Forecasting 2020 US Election: Tough win for Trump, and potential overturn of Biden*

Forecast using Multilevel Regression with Post-stratification

Yi Su

01 November 2020

Abstract

The 2020 US election is one of the most important events in the US for 2020 and especially given that 2020 has been a tough year for US citizens because of numerous human-made and natural events. Among nominates of the US 2020 election, Donald J. Trump and Joe Biden are the two most competent nominates from the two most influential parties, the Republican and the Democratic party. In this report, the technique of multilevel logistic regression with pos-stratification is used to forecast the outcome of the 2020 US election, specifically, predict between Trump or Biden. The regression was fitted on the UCLA Nationscape survey data in a Bayesian setting, and we adapted the American Community Survey dataset as the census data where forecast happens. Using the above approach, we forecast that Trump will lose in terms of the proportion of national votes, but will win the electoral college votes and thus win the election again in the similar way he won the 2016 US election.
keywords: Forecasting; US 2020 Election; Donald J. Trump; Joe Biden; Multilevel Regression with Post-stratification

1 Introduction

President election has been one of the most important political events in the United States of America which happens every 4 years. Most American citizens will be involved in this event and make their decision on which direction the county will go at least in the next 4 years. Meanwhile, the rest of the world will also keep their eye on the election because of the global political position of the US. The election is an invisible war between the parties of the US, among those parties, the Republican and the Democratic are the two oldest and dominant parties.

The Republican party and the Democratic party are the two dominant parties holding a large number of positions in congress. Throughout the history of the US, the competition between the president nominates of these two parties has never stopped. In 2016, Donald J. Trump won the election as the nominate of the Republican party and defeating his opponent from the Democratic party, Hilary Clinton. In 2020, Trump is the Republican nominate again and this time, his main opponent from the Democratic party is Joe Biden, a former vice president of the U.S. during 2009-2017. In general, these two nominates are more likely to win the election than nominates from other parties like the Green party.

In this report, we are interested in forecasting the 2020 U.S. election. Specifically, who is more likely to win among Trump and Biden? To do this, the support of a voter intent survey is essential, and we used the UCLA Nationscape survey dataset (Tausanovitch and. 2020) requested from the URL in reference. The discussion of this survey dataset will be included in *Section 2.1*. Among the many statistical techniques of forecasting election, we used multilevel regression and post-stratification (MRP) to produce estimates of votes. The MRP involves partitioning the data into small cells based on demographic characteristics of our choice, then estimate voter intent (Trump or Biden) in the cell level using a multilevel regression model

*Code and data are available at: https://github.com/YiSu2000/US_Election_Yi_Su

and weight the estimated forecast base on the proportion of the cell. A more detailed explanation of the multilevel part and the post-stratification part are included in *Section 3* and *Section 2.2* respectively. To make the forecasts, we used a census-like dataset, the American Community Survey Dataset (Steven Ruggles and Sobek., n.d.). This is the dataset where partitioning into cells happens and we make predictions from here. The discussion on this dataset is included in *Section 2.2*.

The discussion on the specific procedure of this report and the multilevel regression model is included in *Section 3*, in which we regressed the voting intent based on age, household income, race, and states in the US. In *Section 4*, we present and discuss the resultant model and present estimates of the voting intent between Trump and Biden. Meanwhile, *Section 5* includes discussions of our forecast results as well as some weaknesses and future improvements on our procedure. This report was produced using `R` (R Core Team 2020), as well as some packages which will be mentioned in each section for usage.

The forecast results are divided into two main part, the first part is an overall support rate among the ACS dataset in favor of Trump. The second part is to calculate the mean support rate of Trump in each state and then calculate the number of electoral votes from each state that we forecast Trump to get.

2 Data

2.1 Democracy Fund + UCLA Nationscape Data

The Nationscape (Tausanovitch and. 2020) is a survey conducted from July 2019 to December 2020, collecting demographics of the respondent as well as their voting intent during the 2020 election. The survey samples are provided by Lucid, an online exchange platform focusing on market research. Specifically, the samples were drawn from the online platform based on a set of demographic quotas like age, region, and gender.

The Nationscape aimed at conducting 500,000 interviews in total and roughly 6,250 interviews per week. The survey took the form of an online survey using a survey software controlled by the Nationscape team, however, the respondents were sent to the software directly by the Lucid platform. Since only the respondent will only be directed to the survey software if they match on the Lucid platform, the non-response rate should be reasonably low. Although the quality of the responses is expected to be high from the Lucid platform, the representativeness of the population of interest still needs to be assessed. This is solved by comparing the Nationscape's results to the results of the Pew Research Center's evaluations of online non-probability samples in 2018. The Pew Research Center's 2018 report assessed how various choices impact the quality of the online survey. For more information regarding the UCLA Nationscape dataset, please see the details in the reference section.

After requesting the data on the June-25-2020 phase of the Nationscape survey dataset. We need to modify the original observation levels based on our needs. In this report, we are interested in modeling the voting intent between Trump and Biden by age, household income, race, and states of living. Only 4 variables were chosen because of some hardware limitations which will be discussed in *Section 3* in detail. The original survey data has race and income levels too narrow and thus might cause trouble in our regression.

First, we created a binary variable with 1 meaning vote for Trump and 0 meaning vote for Biden. This will be the response variable of our regression model. We deleted the observations that will vote for nominates other than Trump and Biden, and this caused a reduction in sample size for our model. Further discussion of the potential hazard of this reduction is included in *Section 5.4*.

Second, to make cells wider for a stable sample mean in the MRP process, we redefined some variables related to age, race, and household income. We reduced the household income levels from 24 to 6 wider income groups. For the race, the categories for races are reduced from 15 to 5 wider race groups. Similarly, age was redefined from a discrete variable to a categorical variable with 8 levels, each level representing a decade's group of age. For specific of the group names of each variable, refer to Table 2 which is a frequency table that includes all variable group names except for the states. Since the state is an important factor in the election and we need individual state forecast, there will be no modification on it.

The decision on how many levels or groups there are for each explanatory variable is based on both the

Nationscape dataset and the ACS dataset. For example, since household income is a categorical variable in the Nationscape dataset, we need to categorize the household income numeric variables in the ACS dataset so the two datasets have the exactly same variable categories for forecasting. Another good example is with race, since the ACS dataset groups Chinese and Japanese into one category, we need to group the Chinese and Japanese samples in the Nationscape dataset into one category as well.

Some example observations in the dataset are shown in Table 1). These modifications of the dataset were done with `R` (R Core Team 2020), through R packages `tidyverse` (Wickham et al. 2019), `naniar` (Tierney et al. 2020), `haven` (Wickham and Miller 2020) and `broom` (Robinson, Hayes, and Couch 2020). And Table 1) was created with `knitr` package (Xie 2020) and `kableExtra` package (Dowle and Srinivasan 2019).

Table 1: the first 6 rows of the dataset

Expected Vote in 2020		Income level	Race	Age group	State
Donald Trump	1	$70,000 to $99,999	White	Between 40 to 50	Wisconsin
Donald Trump	1	$175,000 to $249,999	White	Between 40 to 50	Virginia
Donald Trump	1	$35,000 to $69,999	White	Between 70 to 80	Texas
Donald Trump	1	Less than $35,000	White	Between 50 to 60	Washington
Joe Biden	0	$70,000 to $99,999	White	Between 20 to 30	Massachusetts
Joe Biden	0	Less than $35,000	Black/African American/Negro	Between 30 to 40	Texas

Next, we check the frequency of each group in our modified dataset to ensure variation in explanatory variables for the accuracy of the model. Table 2 displays the frequency of each group within each variable instead of plotting the distribution of each categorical variable. These frequencies tell us the distribution of each categorical variable and provide a more compact view of the problems within each variable. It is also a good display of all variable sub-group names. However, since there are 50 states, we only display the first 6 states in alphabetic order.

Although only 6 states are shown, the problem is clear. Some states like Alaska and Arkansas have a significantly fewer number of respondents than large states like California. The same situation happened in all other variables as well, White people have way more number of observations than Chinese or Japanese and Alaskan natives, the number of respondents above 80 is way less than others, and the number of respondents with higher household income decreases as income level increase.

Unfortunately, we can only redefine the age groups since the other variables groups need to line up exactly the same between the two datasets. And some variables are already at the base-line level and could not be further modified. We can choose to specify states into different regions, but that would obey one of our goals of forecasting election base on electoral votes which is highly dependent on winning in each state.

After we joined the age group of above 80 into the age group between 70 to 80. The remaining differences in other variables would be small enough to be compensated by our Bayesian multilevel modeling approach. This approach pools the effect of the minor groups with other more major group cells, which is very beneficial for the states with less than a hundred respondent observations since we cannot redefine state groups.

Table 2: Frequency of each group

Age Group	Count	Income level	Count	Race	Count	State	Count
Under 20	155	Less than $35,000	1648	American Indian or Alaska Native	55	Alabama	71
Between 20 to 30	687	$35,000 to $69,999	1303	Black/African American/Negro	529	Alaska	6
Between 30 to 40	966	$70,000 to $99,999	665	Chinese or Japanese	66	Arizona	122
Between 40 to 50	873	$175,000 to $249,999	228	Other Asian or Pacific Islander	158	Arkansas	31
Between 50 to 60	749	$100,000 to $174,999	797	Some other race	291	California	520
Between 60 to 70	884	More than $250,000	112	White	3654	Colorado	67
Between 70 to 80	387					Connecticut	54
Above 80	52					Delaware	20

Further concerns with this pick of variables are any underlying strong multicollinearity between age and income. We roughly check this by Figure 1, which was created with the `ggplot2` package (Wickham 2016) and `data.table` package (Zhu 2020). We do see an increasing income level trend with aging in general. However, the degree of multicollinearity is not sufficient to violate the assumption of no perfect multicollinearity between explanatory variables since the general proportion of incomes remains steady except for the age group between 20 to 30.

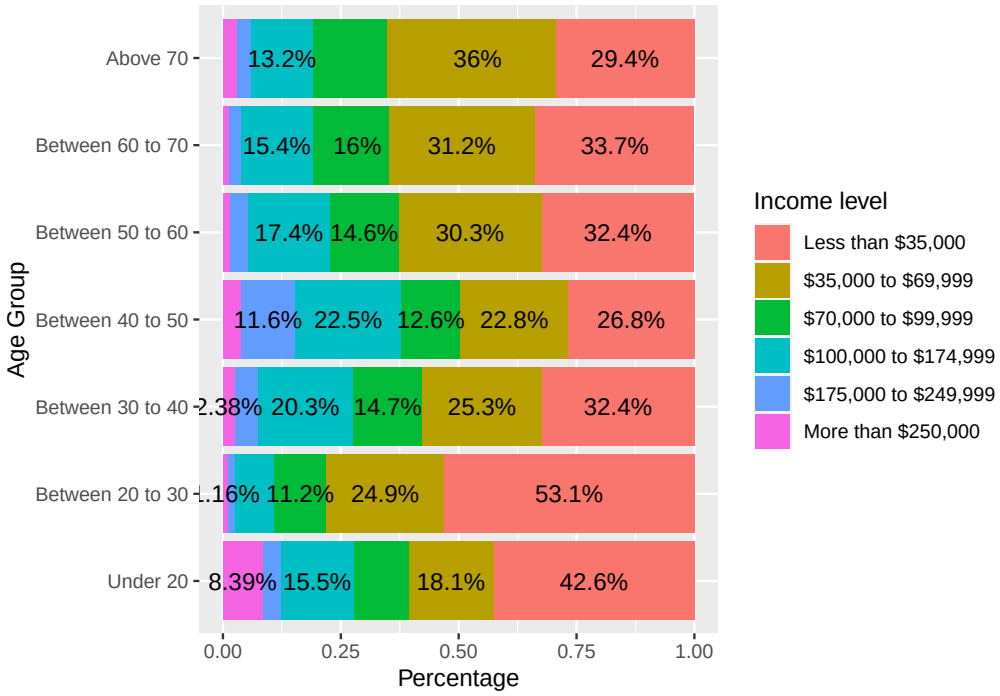

Figure 1: Percentage of income levels for age groups

2.2 American Community Surveys Data

The American Community Survey (ACS) (Steven Ruggles and Sobek., n.d.) are monthly surveys on rolling households that were designed as a substitute for a census. The target population of the ACS is all American households since it is designed to replace a census, and at full implementation, the sample would include about 3 million households across the US. The samples were extracted from the Census Bureau's larger internal data files, and thus shares the same sampling errors. For each sampled household, the ACS records the response from all members of the family. Privacy was protected by restricting geographic variables to the state level and some of the individual variables are Top coded.

The ACS sample design took the form of systematically sampling one household to represent each U.S. county each month. Systematic sampling means sampling using a a scheme among the target sample populations. The monthly sample selected will receive the ACS survey through mail at the beginning of the month. Non-respondents of the mail were contacted via telephone for a computer-assisted telephone interview one month later. If the household is still not responding, one-third of the non-respondents to the mail or telephone survey are contacted in person for a computer-assisted personal interview one month following the last attempt. Notably, the type of house that will be counted as possible receivers of the mail is classified into three types: vacant, occupied mail/CATI, and occupied CAPI. The sampling rate for each of these types are different, for more information on the more detailed sampling design on the house selection and the associated weighting system, please see https://usa.ipums.org/usa/voliii/ACSsamp.shtml.

Since the ACS is sampling on a county level monthly, this makes it close to a census dataset. However, the initial attempt of the survey through mailing may increase the initial non-response rate and decrease the

quality of response. Mailing back takes more effort as the respondent and mailing also have the uncertainty of lost mails during delivery. In the most extreme cases, the survey mail may be miss-classified as junk mail and thus discarded. However, the latter attempts would likely solve this issue but requiring more effort.

The ACS dataset is as good as a census dataset and thus appropriate to our goal of forecasting for the whole nation's vote. For our purpose, we need to filter out all sample observations with age under 18 so the remaining all reach the age to vote, and then assume all observation in the remaining dataset will vote between Trump and Biden. The next step is to create variable groups that are identical to the ones we created in the Nationscape survey dataset. Ensuring the variables in both dataset lines up is how we could implement our model trained from the Nationscape dataset to the ACS dataset and create the estimated forecasts.

First, we notice that the recorded household income in the ACS dataset are numeric variables and we need to first remove the observations with value *9999999* which is a special way of recording missing values. Then we could group the observations into 6 groups by the same logic statements on the numeric income values used in the modified Nationscape dataset. The procedure of grouping the race and age variables is identical to how we grouped the Nationscape dataset. The categorical groups of the race were reduced from 9 groups to 6 groups by merging some smaller groups. The age variable is also grouped into 7-decade groups identical to the modified Nationscape dataset. The above procedure gives a modified ACS dataset that has the identical categorical variables as the modified Nationscape dataset but with a lot more observation. Thus, Table 1 can also be used as example observations of the modified ACS dataset.

Table 3 shows the frequency of each group within the explanatory variables similarly to Table 2 in the previous section. The reason why use a frequency table instead of plotting each variable is the same as for Table 2.

Table 3: Frequency of each group

Age Group	Count	Income level	Count	Race	Count	State	Count
Under 20	60441	Less than $35,000	498152	American Indian or Alaska Native	24013	Alabama	36288
Between 20 to 30	341839	$35,000 to $69,999	620021	Black/African American/Negro	213941	Alaska	4511
Between 30 to 40	372096	$70,000 to $99,999	419774	Chinese or Japanese	41604	Arizona	52176
Between 40 to 50	368195	$175,000 to $249,999	172624	Other Asian or Pacific Islander	95835	Arkansas	22373
Between 50 to 60	437794	$100,000 to $174,999	551253	Some other race	134711	California	284580
Between 60 to 70	432402	More than $250,000	149052	White	1900772	Colorado	42110
Above 70	398109					Connecticut	27132
						Delaware	7178

The distribution of the groups is similar to the Nationscape dataset, and the unequal distribution of frequency in each group still exists as expected. There still exist states that have significantly lower frequency comparing to larger states, the higher house income level groups still have lower frequency among the dataset. Chinese or Japanese and Alaska Native or American Indians still have a significantly lower share of the distribution comparing to the number of White people.

However, this similar pattern is a good sign which indicates that our regression model trained on the Nationscape dataset would be appropriate to use on the ACS dataset. This similarity also indicates the reliability of both datasets in terms of response quality. Since the distribution of variables in the two survey matches, we have more reason to believe that the two surveys captured the true population distributions in general.

The next step is the post-stratification part of the MRP approach. The first step is, to sum all the observations in the dataset sharing the same age group, income group, race, and state. Then we can create new datasets with a calculated proportion of such a combination of variables with respect to the specific variable we are interested in. This essentially gives the distribution of the variable we are interested in and we can re-weight our estimated results to proportions.

Table 4 shows one of such sub-dataset for proportion with respect to age groups. For example, for the first row, the proportion variable means that the proportion of American Indian or Alaska Native living in Alabama with income between $100,000 to $174,999 at the age above 70 is only 0.00025% of all people with

age above 70. This could also be done by calculating the proportion of that cell with respect to the whole ACS dataset samples, which will make the proportion be $n/sum(n)$. The former group-wise proportion is useful when making predictions within each group, while the latter is useful when making a final prediction of the proportion of Trump vote among all the samples in the ACS dataset.

Table 4: Proportion with respect to Age

Age Group	Race	State	Income Group	Count	Proportion
Above 70	American Indian or Alaska Native	Alabama	$100,000 to $174,999	1	2.50e-06
Above 70	American Indian or Alaska Native	Alabama	$35,000 to $69,999	4	1.00e-05
Above 70	American Indian or Alaska Native	Alabama	Less than $35,000	7	1.76e-05
Above 70	American Indian or Alaska Native	Alabama	More than $250,000	1	2.50e-06
Above 70	American Indian or Alaska Native	Alaska	$100,000 to $174,999	9	2.26e-05
Above 70	American Indian or Alaska Native	Alaska	$175,000 to $249,999	8	2.01e-05

3 Model

A multilevel logistic regression model only random effects was fitted on the modified ACS dataset using a Bayesian approach. The regression was done in `R` (R Core Team 2020) using the `brms` package (Bürkner 2017) and `tidybayes` package (Kay 2020). The `brms` package essentially brings the merit of `Stan` using simple `R` syntax and through similar algorithms like Markov chain Monte Carlo (MCMC). MCMC chains allow the model to develop the parameter gradually, and `brms` uses its variant form which is the No-U-Turn Sampler (NUTS) to adapt the samples for posterior distributions, 4 chains and 1000 iteration (not including warm-up) were used to fit the model with the help of `brm` function in `brms`.

As briefly mentioned in *Section 2.1*, this multilevel approach regression would 'pool' information across different groups within a variable so that the group with less information gets that partially shared information from the groups with more information. In this way, the minority group would be closer to the mean outcome of other groups, and the effect of having too little information would be 'diluted' as we have less extreme information going into our model. This ability of getting more informative estimates out of the minority data is the advantage of the MRP approach.

Reflect on our model, we use a model of only random effects, this means we have a different intercept for each subgroup of each explanatory variable. We use a logistic regression since our response variable is a binary variable, with value 1 representing vote for Trump and 0 representing vote for Biden. Our explanatory variables are age groups, race, household income groups, and states which all enter in an additive way in the logit regression, the detailed model is shown in Equation (1).

The choice of age as an explanatory variable is based on the modern history of the US. The US is one of the few countries that have experienced wars in recent decades although they are not domestic wars. This experience of war is likely to influence an individual's opinion about Trump, since the Nationscape dataset does not include veteran status, we use age groups instead and hoping that higher age groups may have higher proportions of veterans whether in WWII or recent ones. The choice of household income and race are also based on some expected difference of opinions about Trump when they have different income level or race. Because of some recent events between races, we do expect a significantly less support rate for Trump among American Africans and the minority races compared to other races. The choice of states is based on its important role in the election, it is almost a tradition that some states have been a supporter of a certain party through out the history, and winning the electoral vote in each state is essential to the 'winner takes all' rule of the election.

However, these choices of explanatory variables are also due to the limitation of the hardware. Unfortunately, the regression needed to be run on local personal devices with relatively poor CPU performance and a limited amount of memory comparing to a desktop device. Four categorical variables with an intercept for each is the maximum performance that the device could get after numerous attempts of more complex regressions. We made attempts including gender, labor force status, education, and citizenship status. Most models are

either successfully built but exceeding our device memory when making forecasts or it breaks when modeling on our device.

$$Pr(y_i = 1) = logit^{-1}(\alpha_{a[i]}^{age\,group} + \alpha_{b[i]}^{income\,group} + \alpha_{c[i]}^{race} + \alpha_{d[i]}^{state}) \tag{1}$$

Equation (1) displays the logistic regression model. We use $\alpha_{a[i]}^{age\,group}$ to represent that each age group a has a unique intercept, this is the same for other variables. Each α represents the age group, race, household income group, and state respectively. And the subscripts $(a, b, c, d)[i]$ represent the specific sub-group that the i^{th} individual belong to. For example, if we have $\alpha_{Below\,20[j]}^{age\,group}$, this means the individual j belongs to the age group of below 20. However, the (a, b, c, d) would be replaced with number 1 to K, for K be the number of the total number of categorical groups in that variable, in the actual regression process.

Each of the α is modeled as normally distributed with mean 0 and a variance that we set prior to. For example, the α for the age group variable would be modeled by:

$$\alpha_{a[i]}^{age\,group} \sim Normal(0, \sigma_{age\,group})\ with\ a = 1, 2, ..., 7$$

Since we are using a Bayesian approach model, we could set prior distribution on the variance (thus standard error) of these αs. In this report, we use the default prior distribution of the `brms` package function `brm` which are student t distributions since there is not much information we could provide to the variances' prior.

The initial model had 20 divergent transitions during the NUTS period. To ensure convergence of the regression, we increased the flexibility of the model by increasing the target average proposal acceptance probability during the adaptation period of NUTS from 0.95 (the default) to 0.98, which fully eliminates all divergent transitions. However, this does not solve the root issue that our model is potentially not a very good fit for the data. Ideally, we could achieve a better fit by adding more control variables in our regression. But as discussed above, the limitation that we had to run the regression locally on our portable devices restricted the degree of layers and explanatory variables we could add to the regression due to the poor performance of hardware.

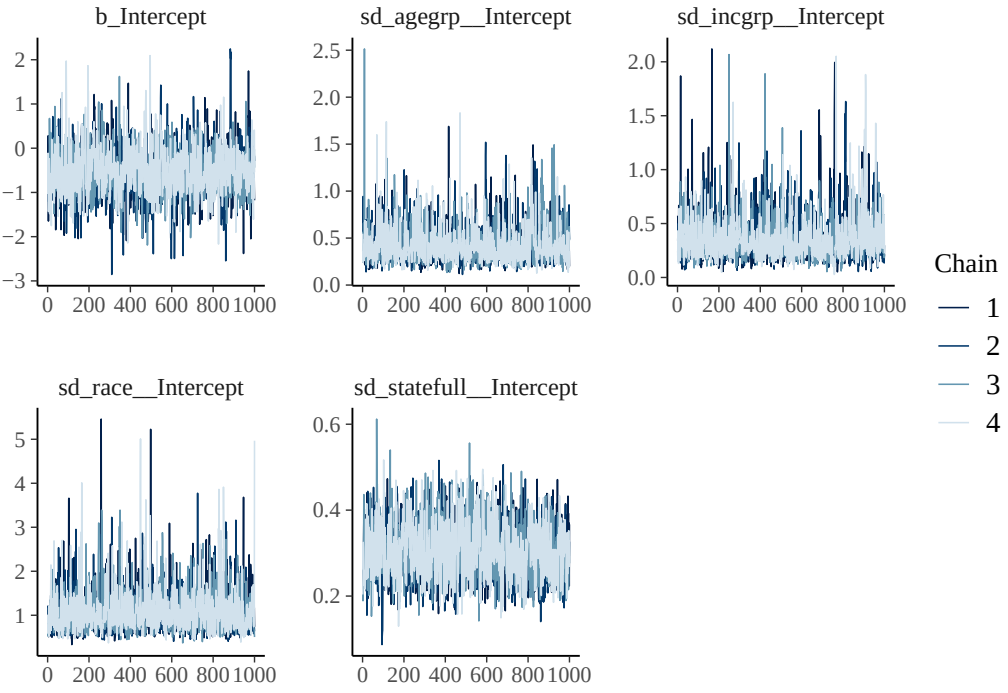

Figure 2: Convergence of Model Parameters

Figure 2 shows the trace plot of parameters for the final regression model we use. We successfully eliminated all divergent transitions during the chain, but the traces are still not quite ideal. An ideal trace would be in a consistent and rapid up-and-down shape with no long term trend. For the standard deviation of the intercepts for age group, income group, and race, there are frequent transitions where the 'up' happened but no 'down'. This could be the aftermath of those previously divergent transitions. However, there is no long-run trend throughout the traces which is good. Even though these traces are not perfect, the model estimation is still convergent with some trade-off between over-fitting and under-fitting in the three intercept groups mentioned above.

4 Results

4.1 Resultant Regression Model

For a multilevel model with only intercepts for the levels of categorical variables, the estimated coefficient of intercepts for each group is replaced by the estimated standard deviations of those intercepts as random effects. Table 5 shows the estimated standard error of the intercepts for each explanatory variable as well as the population level estimated intercept. These estimates are the summary output of the regression model obtained using `brms` package (Bürkner 2017), and the additional root mean square deviation (RMSE) were obtained using `performance` package (Lüdecke et al. 2020) and `clickR` package (Ferrer and Marin 2020) in `R` (R Core Team 2020). The original table produced by `clickR` package (Ferrer and Marin 2020) is not compatible with the actual output of our report's format, thus a substitute is latexed into our report.

Table 5 : *Results*

	Estimate	Std. Error	Number of Levels	CI(95%)	R hat	RMSE
Intercept	-0.521	0.523		(0.217, 1.634)	1.00	
sd(Intercept) Age Group	0.423	0.185	7	(0.194, 0.905)	1.00	
sd(Intercept) Income Group	0.355	0.201	6	(0.124, 0.85)	1.00	
sd(Intercept) Race	1.092	0.474	6	(0.532, 2.32)	1.00	
sd(Intercept) State	0.304	0.062	50	(0.194, 0.436)	1.00	
						0.411

The Intercept is a population-level effect in the regression, in this case, it does not have a meaningful interpretation for its negative values which is outside of our data. However, the R hat of 1.00 means it is convergent in the regression, and a 95% credible interval without 0 means in 95% of the case if you continue to construct the intervals, it will not contain 0 which distinguishes the population level intercept from 0. The estimated standard deviations for the intercepts are all convergent as indicated by the R hat value of 1.00, and we are sure there is decent variation among levels of intercepts with all the 95% credible intervals for estimated standard deviations excluding 0.

Notice that the estimated standard error of the race intercepts is highest among all four group-level effects with the value of 1.092. This is in general a significant amount of deviation given we are using a binary response variable, thus, we believe there will be some extreme differences for the support rate of Trump among some different races. For more direct visualization of the group-level effects, a set of group-wise support rates are included in *Section 4.2* for the age groups, income groups, and races. In *Section 4.3* we present the support rates of each state and forecast the outcome of the election using electoral votes.

Figure 3 shows a receiver operating characteristic curve (ROC) curve of our resultant model, the ROC was produced using the `pROC` package (Robin et al. 2011). The ROC curve plots the False Positive Percentage (FPP) and True Positive Percentage (TPP) when using the model to predict back on the sample dataset for the binary outcome of supporting Trump or supporting Biden. When making predictions, each point on the ROC represents a threshold value of treating the prediction as 1 or 0. The true positive rate is the rate of successfully predicting the response variable, and the false positive rate is the opposite rate. The ROC plots the two rates for every possible threshold of identifying the binary response, and an ideal ROC should be as

close to the upper left corner as possible for the highest TPP and the lowest FPP.

The ROC of our resultant model has an area under the curve (AUC) of 69.5%, the AUC is a general measure of goodness of the ROC and an AUC of 69.5% is usually considered moderate and indicating some decent trade-off between TPP and FPP. The ROC suggests that an optimum choice of threshold provides a TPP of around 80% for an FPP of around 48%, this trade-off between TPP and FPP indicates our model could use some reliability improvement when making predictions.

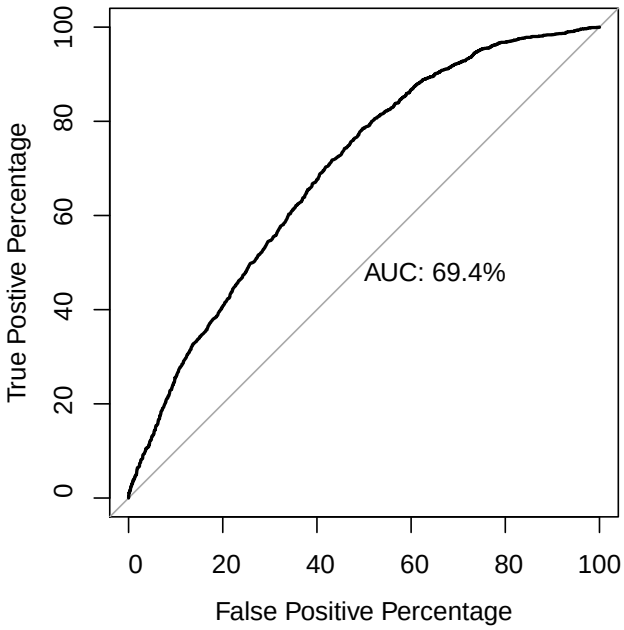

Figure 3: receiver operating characteristic curve (ROC)

4.2 Forecast of overall support rate of Trump

As mentioned in previous sections, we used two different approaches to forecast the outcome of the election. The first approach is to forecast by estimating the support rate of Trump in each cell of the post-stratification dataset (the modified ACS dataset) and times the estimate by the proportion of the cell in the dataset and then sum all cell estimations for a final support rate of Trump. This procedure can be summarize by Equation (2), the n_i represent the number of counts of the i^{th} cell and x_i is the estimated support rate for the i^{th} cell. The summation of the n_is is thus the number of observations in the post-stratification dataset.

$$Support\ rate = \frac{\sum n_i x_i}{\sum n_i} \tag{2}$$

This procedure gives an overall estimating of support rate for Trump of 49.38%, with a margin of error of 22.45%, certainly this is a naive approach to forecasting the election with a higher margin of error. In the real world, the US election uses the winner-takes-all method, and discussion of this method will be discussed in *Section 4.3*. Before going into this alternative approach, some group-level effect of the age groups, income groups, and races are displayed and discussed below for a better understanding of how our model adapt these variables.

Figure 4 shows the estimated proportion of Trump voters (the support rate) in each age group along with a 95% predictive interval. The 95% predictive interval captures 95% of the estimation cases among each age group cells. All graphs in this section are made with the `ggplot2` package (Wickham 2016). In the graph, there is a positive trend as higher age groups tend to have a higher estimated proportion of votes for Trump on average. The support rate peaks at the age between 50 to 60 and gradually decreases as moving on to

older age groups, but are still higher than younger age groups in general. These differences might be due to the veteran effect mentioned before, older age groups have a higher odd of been through a war and they may have a different aspect to politics from that war experience.

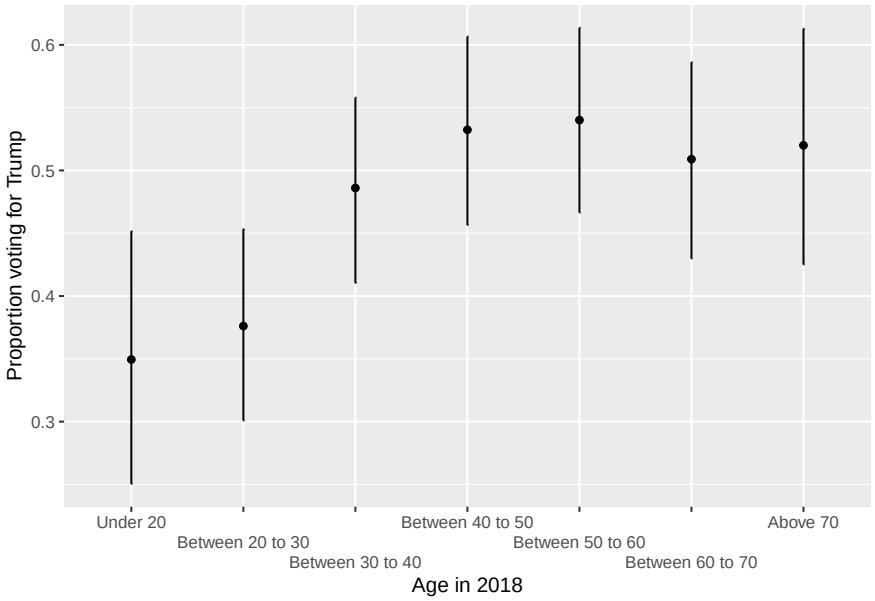

Figure 4: Forcasting Votes for Trump by Age Groups

Figure 5 shows the estimated proportion of Trump voters (the support rate) in each household income group along with a 95% predictive interval. The 95% predictive interval captures 95% of the estimation cases among each income group cells. In general, all income groups with income below $99,999 have no group-wise difference in support rate for Trump. An overall positive relationship between support rate for Trump and income level happened for income from $100,000 to $249,999 and gradually decreases for the group with income more than $250,000. The off-trend effect of the group with income more than $250,000 might be due to the lower amount of samples for individuals with such high household income levels. In general, higher-income people tend to vote for Trump more often on average.

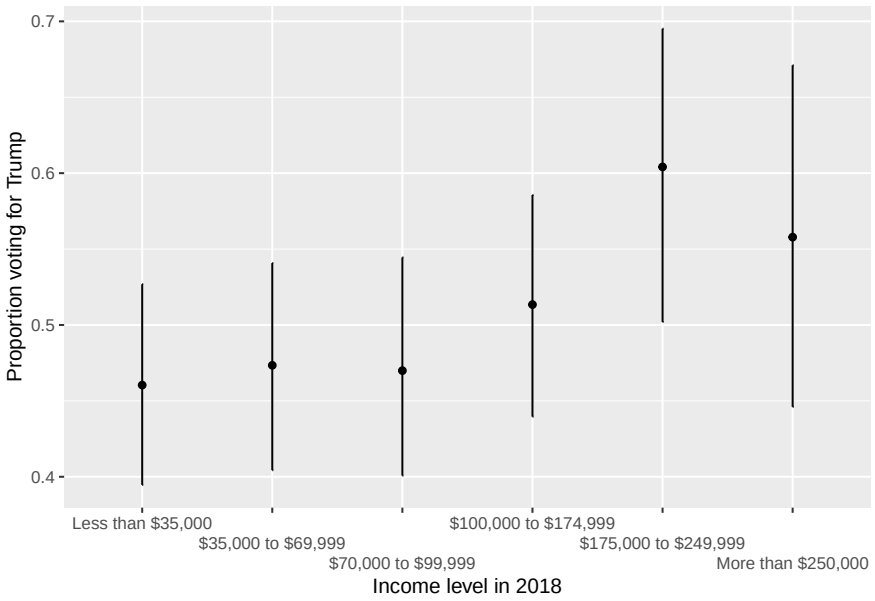

Figure 5: Forcasting Votes for Trump by Household Income Groups

Figure 6 shows the estimated proportion of Trump voters (the support rate) in each race along with a 95% predictive interval. The 95% predictive interval captures 95% of the estimation cases among each race group cells. As expected, African Americans and some other minority race groups have a significantly lower support rate comparing to white people in the US on average. The predictive interval of the African American group does not even exceed a support rate of 20%, and the predictive interval for most other minority races do not exceed 50% support rate. To recap, although the ACS dataset is from 2018, but the UCLA Nationscape dataset which we trained our model is from June 2020 thus the model will predict using 2020 respondent behaviors. These group differences are expected because of some events involving potential racial discrimination that happened in early 2020, as a result, the minorities will have lower trust on average for the president of 2020, Donald Trump.

The American Indian or Alaska Native race group has a higher support rate for Trump on average potentially for two reasons. One is the significantly lower amount of samples in this category and the other is that they do behave similarly to white people in election choices. The latter could happen since The American Indian or Alaska Native groups have lived in North America since the White people come to the continent, throughout the hundreds of years there has been some discrimination against the natives but modern decedents of these first nations have likely merged themselves into the white population to some degree. Thus it would make sense if there is similar political taste among the American Indian or Alaska Native and the white people.

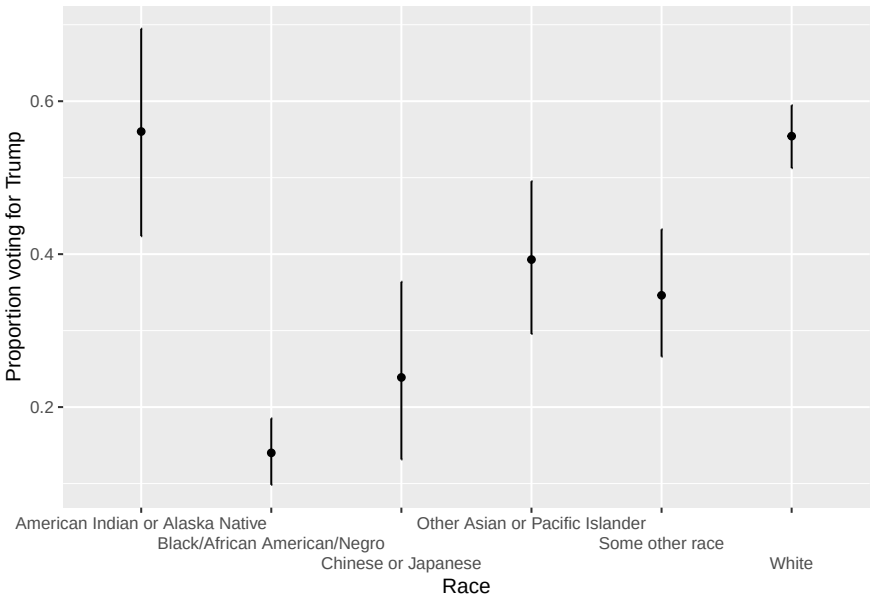

Figure 6: Forcasting Votes for Trump by Race

4.3 Forecast Electoral Vote by Support Rate in Each State

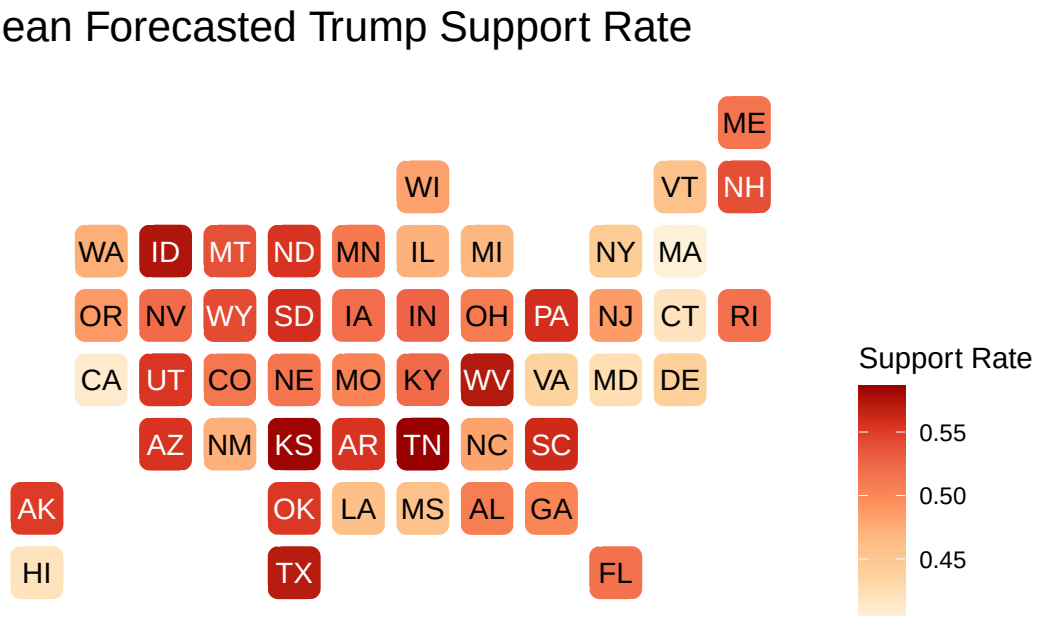

Figure 7: Forcasting Votes for Trump by States

The electoral college in the US is a group of presidential electors selected every four years for electing the president and vice president of the US. There are 538 electors in 2020, and any nominate needs to have the absolute majority of the 538 votes in order to win the election meaning they need at least 270 votes. The 538 electors are separated by states, and the number of electors in each state is based on the number of seats the states have in the US House of representativeness and the US senates. For example, the state of New York has 2 positions in the senate and 27 positions in the house of representativeness, then the state of New York has 29 electoral votes. Among the 50 states of the US, 48 states use the form of 'winner takes all' for the electoral votes, meaning if the nominate wins the most proportion of votes in that state, the nominate will

have the support of all the electoral votes in that state. However, the state of Maine and Nebraska do that in a more complex way involving winning in particular rides of the election. Thus, since our regression model does not predict and level narrower than the state, we estimate the number of electoral votes from the state of Maine and Nebraska as a rounded half of what they actually have. In the real world, the state of Maine and Nebraska have 4 and 5 electoral votes respectively, in our estimate of total predicted electoral votes that Trump gets, we will only count the number of electoral votes as 2 and 3 respectively.

Figure 7 shows the predicted votes as the support rate of Trump in that state. This graph was made with the `statebins` package (Rudis 2020). The darker the color of the state, the more proportion of votes Trump gets in the state. Notice that the color of the state names in Figure 7 are just for readability and do not indicate whether Trump has a support rate of more than 50% in that state. In previous elections, nominates does not necessarily need more than 50% of the vote to win in a state, but since our model only forecast whether they vote for Trump or Biden, they will need more than 50% of the vote in a state for an ensured win.

Table 6 in the Appendix displays the full list of states that Trump has a mean estimated support rate of more than 50% along with their 95% predictive intervals and the number of electoral votes in that state. It shows that in our prediction using the regression model, Trump wins in the state of Maine and Nebraska, as well as the other 29 states on average.

Summing up all the electoral votes including the halved 2 votes for Maine and halved 3 votes for Nebraska, the predicted number of electoral votes that Trump gets is 279. This is larger than the absolute majority of 270 votes among the 538 votes meaning an overall win for Trump. Even though the prediction in *Section 4.2* indicates a slightly overall loss in the overall proportion of votes, Trump wins in both the states that have more electoral votes on average and the smaller populated states and thus wins the overall president election by winning in the electoral college.

5 Discussion

5.1 More on the Group Level Effects

In *Section 4.2*, the figures show that on average, the support rate of Trump increases as age increase and income increase. For the age groups, the positive trend slightly decreases when for the age group with age above 60 but not by much. The mean estimate of support rate gradually increases from approximately 35% to 54%, and the predictive intervals are slightly narrower when age increases. Narrower predictive intervals mean that the group effect is more accurate. To explain why support rate increases as age increases, as mentioned when choosing age group as an explanatory variable for random effects, we are expecting the older population to have a higher proportion of veterans. Especially for the people born from 1940 to 1970, which is the WWII and the post-WWII baby boomers period.

Comparing to the younger generations, this generation was born in the wars and post-war. Throughout their lifetime, they went through the cold war which was almost 2/3 of their life as well as the Korean and Vietnam wars. This generation may have more tolerance to some of the more aggressive policies and decisions of Trump on average, therefore a higher mean support rate for Trump on average. Meanwhile, the younger generation could think that those policies and decisions are less understandable. Since veteran status is not included in the UCLA Nationscape dataset, we are unable to train the regression model with veteran status and thus unable to analyze if our expectations are true.

Income groups behave similarly to age groups, but we have almost no variation for the mean estimated support rates among groups with household income lower than $100,000. A positive relationship only appears for income from $100,000 to $249,999 and gradually decreases for the group with income more than $250,000. While this suggests that we should reconsider the grouping of household incomes for a better fit of the model, there are some explanations for the current effects. The people with household income from $100,000 to $249,999 works in job-fields that are considered "higher-end", such as medical professionals, engineers, and lawyers. The mindset of these people could be different from those who work in fields with lower income in general. Nevertheless, the income groups should need more revision for better prediction on the modified ACS dataset.

The race is less complicated among the group-level effects. Since African Americans are directly affected by the event regarding potential racial discrimination earlier this year, they display an extremely low estimated support rate ranged from approximately 10% to 19%. And most other minority race groups display an average estimated support rate of less than 50% as expected, except for the American Indians and Alaska Natives which was mentioned already. For the estimated support rate of the White people, the 95% predictive interval is narrower than all other race groups, approximately ranging from 56% to 59%. This higher accuracy is likely the merit of having a White people sample that is more than the summation of all other races which is inevitable due to the race structure of the US. The centrality of the interval also tells us the real-world distribution of Trump supporters among the White population is likely to be around that 50%.

Regardless of states, the resultant model will predict the highest support rate for Trump among the White population with age more than 50, household income more than $175,000. Interestingly, these characteristics all follow from the personal characteristics of Trump himself. This indicates that the relationship between the nominate himself and the supporters to some degree might be a novel direction for future exploration of elections.

5.2 Reliability of the forecast

In *Section 4*, the naive way of predicting the population support rate of Trump indicated that Trump will lose to Biden by a hair on average. The naive prediction was naive because it does not reflect on the real winning condition of the US election, winning in states with more electoral votes are much more important than winning more states in general. For example, the state of California has 55 electoral votes for having the most number of positions in the House and representativeness and the Senates , while many other states like the state of Wyoming only have 3 electoral votes. Meanwhile, in a more realistic setting, the prediction by states using electoral votes indicated that Trump will win the election, on average, by 9 votes more than the absolute majority of 270 votes among the 538 electoral votes in total.

These forecasts are all based on average prediction using the post-stratified ACS dataset with lots of uncertainty. In fact, from Table 6 in the Appendix, there are no 95% predictive intervals that exclude 0.50, meaning there are no states that Trump always has more than 50% support rate and more variability in the election. To recap, nominates do not necessarily need more than 50% support rate to win in a state, the threshold is 50% in this report because the regression model is binary. Considering Trump only wins by 9 electoral votes in the mean estimation, Biden still has a decent chance of winning the US election if some states support Biden instead. From the "winner takes all" rule of the electoral college votes, Trump losses if he loses in one of those larger states like the state of Texas, Florida, Georgia, and Ohio which has a lot more electoral votes than 9.

5.3 Weaknesses and next steps

The resultant regression model will need a lot more improvement in terms of prediction correctness and fit, but some improvement can be done using current settings while some are out of our reach. In the Bayesian settings, we are always trying to adapt to the real world, and the foundation of that is having a decent model fitting the survey data on hand.

The MRP approach itself does not stop us from building more complex mix effect models instead of the current random effect model. Thus, the inclusion of other continuous explanatory variables might be helpful, and interaction terms like one between age and income would be useful for controlling variations of the regression. Some potential control variables could also be included like citizenship status, veteran status, and education attainment. However, the inclusion of future predictors needs to be careful with any confounders that potentially exist in the dataset, as well as the degree of multicollinearity with other existing predictors.

Production of predictive draws using complex models will require local devices with stronger performance. As discussed in *Section 3*, the requirement of running the model on local devices restricted the complexity of the regression especially for the regression model using Bayesian settings. These hardware limitations could be solved in non-pandemic periods by renting devices, or through overclocking the local device which is not recommended since it lowers the stability of the device permanently if not done properly. An alternative

approach is to try other frequentist regression approaches like using generalized linear regression models on a maximum likelihood structure.

Another constraint is the datasets itself, even though they are both reliable datasets but they do not capture much of the useful information relating to the election and the requirement of identical variables in the two datasets is another constraint for regression. For example, the former vote in 2016 would be a very informative variable to be included in a regression, but the ACS dataset does not record that while the Nationscape dataset did. Some alternative survey datasets could be the polling datasets, which are likely to have restrictions within each party or organization on the distribution of such datasets. The idea is to have datasets that feature more personal aspects of political tastes, and regression on those aspects should provide more relevant information and more accurate predictions.

In terms of using the polling dataset, the importance of winning the electoral votes in those particularly larger states implies that polling companies building the survey datasets should focus on collecting state-wise differences. But on the other hand, this is expected to be very difficult since the difference in opinions is among the hardest characteristics to measure correctly, and is unlikely to be included in the most census. And if it's not in the census dataset, it cannot be used in the regression for prediction. The other approach is to increase the number of survey responses in states with more electoral votes. The additional variation in survey data would help the regression model to produce a more accurate result with less error.

Appendix

Table 6: List of states that Trump is forecasted to win

State	Mean Estimated Support Rate	2.5th Percentile	97.5th Percentile	Electrol Vote
Alabama	0.5083323	0.3385968	0.6692316	9
Alaska	0.5500960	0.3779428	0.7251219	3
Arizona	0.5552791	0.3793080	0.7182258	11
Arkansas	0.5552081	0.3594947	0.7387934	6
Colorado	0.5125775	0.3283258	0.6939302	9
Florida	0.5163946	0.3581729	0.6738320	29
Georgia	0.5010415	0.3561457	0.6429099	16
Idaho	0.5752379	0.3665448	0.7744405	4
Indiana	0.5265186	0.3363321	0.7099414	11
Iowa	0.5202217	0.3089449	0.7254129	6
Kansas	0.5837116	0.3779112	0.7789664	6
Kentucky	0.5234565	0.3329115	0.7118379	8
Maine	0.5153685	0.2880474	0.7347217	2
Minnesota	0.5113255	0.3107805	0.7026190	10
Missouri	0.5034320	0.3133300	0.6934908	6
Montana	0.5385234	0.3193633	0.7539297	3
Nebraska	0.5144277	0.3016873	0.7279859	3
Nevada	0.5224990	0.3623708	0.6755905	6
New Hampshire	0.5394787	0.3200282	0.7483655	4
North Dakota	0.5545930	0.3295641	0.7656453	3
Ohio	0.5100512	0.3340955	0.6827220	18
Oklahoma	0.5521705	0.3764844	0.7262155	7
Pennsylvania	0.5580115	0.3778855	0.7312585	20
Rhode Island	0.5165172	0.3019726	0.7184983	4
South Carolina	0.5609876	0.3926478	0.7121442	9
South Dakota	0.5584841	0.3362397	0.7702996	3
Tennessee	0.5863008	0.4098392	0.7543660	11
Texas	0.5711524	0.4261633	0.7095639	38
Utah	0.5539963	0.3634798	0.7502949	6
West Virginia	0.5723192	0.3493975	0.7871430	5
Wyoming	0.5419596	0.3094800	0.7680741	3

References

Bürkner, Paul-Christian. 2017. "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (1): 1–28. https://doi.org/10.18637/jss.v080.i01.

Dowle, Matt, and Arun Srinivasan. 2019. *Data.table: Extension of 'Data.frame'*. https://CRAN.R-project.org/package=data.table.

Ferrer, Victoria Fornes, and David Hervas Marin. 2020. *ClickR: Fix Data and Create Report Tables from Different Objects*. https://CRAN.R-project.org/package=clickR.

Kay, Matthew. 2020. *tidybayes: Tidy Data and Geoms for Bayesian Models*. https://doi.org/10.5281/zenodo.1308151.

Lüdecke, Daniel, Dominique Makowski, Philip Waggoner, and Indrajeet Patil. 2020. "Performance: Assessment of Regression Models Performance." *CRAN*. https://doi.org/10.5281/zenodo.3952174.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. "PROC: An Open-Source Package for R and S+ to Analyze and Compare Roc Curves." *BMC Bioinformatics* 12: 77.

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

Rudis, Bob. 2020. *Statebins: Create United States Uniform Cartogram Heatmaps*. https://CRAN.R-project.org/package=statebins.

Steven Ruggles, Ronald Goeken, Sarah Flood, and Matthew Sobek. n.d. "IPUMS Usa: Version 10.0 [Dataset]." Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0.

Tausanovitch, Chris, and.Lynn Vavreck. 2020. "Emocracy Fund + Ucla Nationscape, October 10- 17, 2019 (Version 20200814)." https://www.voterstudygroup.org/publication/nationscape-data-set.

Tierney, Nicholas, Di Cook, Miles McBain, and Colin Fay. 2020. *Naniar: Data Structures, Summaries, and Visualisations for Missing Data*. https://CRAN.R-project.org/package=naniar.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, and Evan Miller. 2020. *Haven: Import and Export 'Spss', 'Stata' and 'Sas' Files*. https://CRAN.R-project.org/package=haven.

Xie, Yihui. 2020. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.

Zhu, Hao. 2020. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.