Turnout in the 2020 United States Presidential Election

Riggs Markham

Christian Brothers University

Mathematics Senior Project

Spring 2021

Table of Contents

Abstract

This study examines the high voter turnout of the 2020 United States Presidential Election, and in particular, examines several factors that might have accounted for different levels of turnout between states. Accounting for the demographic profiles of the states as well as basic partisanship, this study seeks to find the connection between turnout and three primary factors: the margin of victory of the election, the restrictiveness of laws relating to voting, and the local severity of the coronavirus pandemic. My results show that the margin of victory has a large effect on voter turnout, with states that have closer elections experiencing markedly higher turnout that states without competitive presidential elections. Additionally, the restrictiveness of certain voting laws – particularly vote-by-mail regulations and, to a lesser extent, the amount of time available to register for an election – had an effect, with turnout increasing when those legal regimes were less restrictive. My results show that the severity of coronavirus outbreak in a state had very little effect on turnout, which could indicate that the changes in voting laws implemented in response to the pandemic – in particular, making voting-by-mail easier – were effective in blunting the pandemic's expected negative effect on turnout.

Introduction

The 2020 Election

On November 3rd, 2020, the last Americans cast their ballot in the 2020 United States Presidential Election. Held in the midst of the devasting COVID-19 pandemic, the election delivered a victory for the challenger Joe Biden over the incumbent Donald Trump, although the losing candidate would protest this outcome in the coming days.

Surprisingly, considering the presence of a pandemic that made congregations in public spaces extraordinarily dangerous, in this highly contentious election, the United States had the most votes ever recorded in one of its presidential elections. While it is reasonable to expect that record to be periodically broken as the American population continues to rise – that record was broken in every presidential election from 1952 to 1984 [28] – what makes in the 2020 election truly extraordinary is that the United States also had extraordinarily high turnout; according to Michael McDonald of the United States Election Project, preliminary estimates showed that turnout would reach its highest level since 1900, 120 years and 30 presidential elections ago [11].

Furthermore, since in elections in those days excluded massive numbers of Americans - before 1920, women were not afforded the right to vote nationally, and, even after the Civil War and the end of slavery, Black Americans were both *de facto* and *de jure* disenfranchised across the South under Jim Crow from the 1890s until the 1960s under Jim Crow – one can surmise that the 2020 presidential election likely had the highest proportion of American adults voting in American history [28].

Defining Voter Turnout

Voter turnout is the proportion of people in a district who were allowed to vote that actually voted in a specified election. As countries generally limit their franchise based on age – with adults being eligible to vote and children being ineligible – and since that sort of basic demographic data is usually easy to find, it is often a simple task to calculate a country's voting age population (VAP). But since many nations and sub-national units have further, more specific criteria for voter eligibility, one must find the requisite data for those qualifications to calculate the relevant number here, the voting eligible population (VEP) [9].

For example, in the United States, two key groups are included in VAP statistics that are ineligible to vote: non-citizens and some people convicted of crimes. Those pieces of demographic data are often difficult to find, being either politically sensitive issues or complex questions because of American federalism [9].

Therefore, we define the turnout rate in an election to be equal to the total number of legally-cast ballots cast in a specified election divided by the voting eligible population.

History of US Turnout

Voter turnout, specifically in the context of American presidential elections has varied significantly throughout its history. Starting out low at the dawn of the republic, turnout rose until the election of 1840, when it stabilized at around 80% of VAP, remaining there through the election of 1896 [10]. After the election of 1896, turnout rates fell, reaching a low of just under 50% in the 1920 and 1924 election, over the next 90 years, turnout remained between 50% and 65%, with a high period from the 1930s to the 1960s and a low period from the 1970s until the 2000 election. From 2004-2016, turnout hovered at around 60% [10].

One key element of this discussion is the changing franchise over the years. Initially, since many states enforced property requirements for voting, the franchise was essentially made up of only rich, white men. After much of these requirements had been lifted during the period of Jacksonian democracy, turnout rose to its highest point. With partisanship extremely high– note that the Civil War took place right in the middle of this period – and political machines at the height of their power, voter turnout stayed at its height in American history during this 1840-1896 period [28].

Much has been made about the fall of turnout at the end of the 19th century that continued until the 1920s. In the South at least, Reconstruction was done, and the era of Jim Crow began, with many states viciously restricting the rights of Black Americans to vote using methods like poll taxes, literacy tests, and good-old-fashioned white terrorism [28]. But Jim Crow was largely restricted to the South, in the rest of the country, research has attributed this large drop in turnout to voting reforms, specifically the introduction of the “Australian ballot” and individual voter registration [2]. Whereas previously urban political machines had given out ballots containing only one party to voters – along with bribes and the expectation of patronage – the introduction of the secret ballot, also known as the “Australian ballot,” made this practice impossible. Ballots were now printed by the government, putting candidates together under the office for which they were running instead of with the rest of their parties. And since all candidates were listed on a ballot and the votes were cast secretly, voting machines could have no assurances that their bribes were being followed through on [2]. This “Vote Market Hypothesis” – the claim that the increased difficulty of corruption lowered turnout rates – is disputed by Reed who attributed this turnout drop to the ballot design rather than voter secrecy [17]. Along with this was the advent of voter registration, which served as a highly effective means to disenfranchise Black voters during Jim Crow – one of the most important focuses of violence during the Civil Rights Era was campaigns to registered Black southerners to vote. It also had the effect of making it more difficult for people to vote as they had to personally register. The introduction of these measures was typical of the Progressive Era where corrupt practices like patronage and bribery were attacked, which had the effect of lowering the number of people voting [23].

As the franchise increased, with women first being allowed to vote nationally in the 1924 election and with the Civil Rights Era dismantling Jim Crow, the franchise was massively expanded. The voting age was also lowered from 21 to 18 in 1971, giving a group of voters who might vote in low numbers the franchise. These expansions might have lowered the total turnout, expanding the vote to voters who had not consistently voted in the past or might be less interested in voting in general [28].

Since the 2000 election and its highly contentious conclusion, turnout rose consistently above its norm in the 1970-2000 period. With the solidification of “red states” and “blue states” in the Electoral College system, the focus of attention on American presidential elections has been on “swing states,” those where margins of victory are close. The most famed instance of this phenomenon, and likely where it cemented itself in the American psyche, is the 2000 presidential election in Florida. This has often been attributed to increasing partisan polarization and an increased perception of the importance of elections [5].

Turnout Around the World

In general, the United States has significantly lower turnout than other nations of comparable wealth [3]. According to Pew, the U.S. places 30th out of 35 peer nations in the Organization for Economic Cooperation and Development for which there is data [3]. This phenomenon has been explained via many methods, key among which are cultural factors leading to higher voter apathy and the electoral systems in place hurting turnout. Other laws surrounding registration, eligibility, and voting can also have large effects; for instance, Australia and Belgium, polities with consistently high turnout, have compulsory voting laws [3]. Furthermore, when Chile moved off of a compulsory voting system in 2012, its presidential turnout plummeted from 87% in the 2010 election to 42% in the 2013 election [3]. Voting systems can also have a significant effect on turnout. Research by Paskert showed that proportional representation electoral systems have higher turnout, along with countries with compulsory voting [15]. Along with those conclusions on voting systems, high levels of democratic satisfaction were also found to induce higher turnout [15].

Preliminary Analysis of 2020 Turnout

Initial Analysis

Before 2020, media sources expected extremely high turnout due to extremely high polarization among voters. With Democrats detesting the incumbent President Trump and Republicans idolizing him, voters saw a high importance in the 2020 election. In fact, this sense of significance has been on the uptick since at least the 2000 election according to the Brookings Institute [5]. In 2000, on the question "Does it really matter who wins the presidential election?" 50% of voters thought it did really matter versus 44% who believed it didn't [5]. This 4% gap suddenly expanded to 38% in 2004, stayed around there for 2008 and 2012, jumped to 52% in 2016, and finally reached 67% in 2020, with the number of voters responding "Things will be pretty much the same" falling to 16% [5].

Voters in 2020 viewed the differences between electing Donald Trump and Joe Biden as far more significant than, for example, the differences between George W Bush and Al Gore. Pew Research attributed the rise in turnout at least partially to "the bitter fight between incumbent President Donald Trump and challenger Joe Biden" [4].

But as these media predictions were largely deemed irrelevant after it became apparent that the COVID-19 pandemic would have a large effect on the mechanics of voting. This occurred during the Democratic primary election, with many states pushing back their primary election dates until after they thought the pandemic would be over, later in the spring or early in the summer [25]. As we all know, the pandemic did not end that quickly, and as fall approached and the pandemic raged on, many commentators expected the pandemic to hurt turnout in the November election [13].

While these discussions were rife before the election took place, after the election, Santana et al. released a study analyzing elections held during the pandemic worldwide, and found that, while turnout did not generally decline in comparison to elections before the pandemic, political participation is already lower in polities that had worse outbreaks of the disease [18].

Hypotheses

My first hypothesis is that less restrictive voting laws – specifically allowing registration closer to election day, having more days available for early voting, and having more lenient vote-by-mail rules – will result in higher turnout for the states that possess these laws. Many of these laws were passed specifically to make it easier for voters in the midst of a pandemic, so it would indicate the correct judgement of those policy makers if they increased turnout.

My second hypothesis is that states with more severe COVID-19 outbreaks will have lower turnout. The severity of an outbreak will be determined by both the total number of cases and deaths caused by the virus before election day and the new cases and deaths from the week preceding election day. This outcome would be expected to occur due to the vast disruptions to daily life caused by the pandemic in other areas of life, but overwhelming passion from voters or the blunting effects of reforms might limit this effect.

My final hypothesis is that states with tighter margins – smaller differences between the proportions of voters who voted for Biden and Trump – will have higher turnout. With an increasingly partisan electorate, most states are not seen to be competitive in the election. This would indicate that the Electoral College system of the United States and its swing states are pushing down electoral participation in much of America.

Data Collection

For the core, state-level turnout data, I used data for the 2016, 2018, and 2020 elections from the US Election Project run by Dr. Michael P. McDonald of the University of Florida [6, 7, 8]. To compile this data, the US Election Project gathered the ballots counted in each election. They also obtained estimates of the voting age population in each state from the US Census Bureau. From there, to calculate the voting eligible population in each state, they used estimates of the noncitizen proportion of the population and then also used data from criminal justice databases to find the number of people ineligible due to a criminal conviction. Using these numbers for how many people are ineligible, they calculate the VEP and then subsequently the turnout for the state [9].

For my analysis of election laws, I used FiveThirtyEight's "How to Vote in 2020" guide to compile the data [16]. This guide was meant as a tool for voters to inform them of their state's specific election laws, but I used it as a resource to record and classify each states' election law systems [16]. From this site, I recorded the final date available for voter registration, the starting and ending dates for early voting, and the eligibility requirements for vote-by-mail. With voter registration, many states have voter registration available on the day of the election, so for those states, I recorded November 3rd as the final registration date. With this data, I calculated a metric of the number of days before the election that registration ended, which I titled 'days before.' With the start and end dates for early voting, I created a metric of the days between these dates called 'days between.' Now, this metric would be inferior to a metric of the number of days that early voting is available in a state, but since in many states the exact number of days differs by county, I decided to simply choose the widest available early voting data range in each state instead of dealing with all of those exceptions. For the vote-by-mail eligibility, FiveThirtyEight created four classifications of whether one can vote by mail [16]:

- Everyone can vote by mail, and ballots are automatically mailed to voters
- Everyone can vote by mail, and mail-ballot applications are automatically mailed to voters
- Everyone can vote by mail, but nothing is automatically mailed to voters
- You can vote by mail only if you have a valid excuse (the pandemic doesn't count)

I classified these respectively as "auto," "auto app," "all," and "excuse," and then I subsequently encoded them as the integers 1 through 4, and stored them as the metric 'VBM num,' with 1 representing the least restrictive voting laws and 4 representing the most restrictive ones.

For the demographic data, I obtained data from the United States Census Bureau that contained the number of residents in each state, of each sex, of each origin (Hispanic or non-Hispanic), of each race, and of each age [22]. This dataset, titled SC-EST2019-ALLDATA6, contained estimates of each of these tiny cross-sections for each year from 2010 to 2019. For example, this dataset estimates that that in 2012, there were 20,246 66-year-old White Hispanic men residing in the state of Texas [22]. To extract the data that I wanted, I extracted Hispanic from the "origin" column and made it its own race, encoded as the number 7. I also grouped the cross sections by age, creating groups of ages under 18, 18-29, 30-44, 45-59, 60-74, and 75 and up. Using this collected data, I then calculated the proportion of each state's population for each sex, racial group, and age cohort.

For the actual results of the 2020 election, I used data from the MIT Election Lab showing the number of votes cast for each presidential candidate for each state in recent presidential elections [12]. After filtering down to the 2020 election specifically, I grouped all of the votes for candidates of neither the Democratic nor Republican Party into a separate “other” category, and also recorded the margin of votes between Biden and Trump for each state.

Finally, for the COVID-19 data, I used data from the New York Times tracking of the case totals and death totals for each date from early 2020 all the way into the present day [14]. I isolated the totals for election day, November 3rd, 2020, and pushed those into ‘covidcasestotal’ and ‘coviddeathstotal’ variables. I also subtracted the case and death totals on October 27th from those on November 3rd to find the number of new cases and deaths in that week, trying to estimate the severity of the pandemic specifically around election day (e.g. New York had many deaths in Spring 2020, but its pandemic was much calmer by the fall, whereas the Dakotas were having horrible outbreaks at the time of the election) [14]. I stored those in ‘covidcasesweek’ and ‘coviddeathsweek’ variables.

Methods of Analysis

To analyze the effects of voting laws, the COVID-19 pandemic, and the margin of victory on turnout, I performed a multiple linear regression analysis in the programming language R. As my independent variables, I primarily used data from the census on the demographic makeups of states. I used the proportions of each state’s population that is a member of each sex category, racial category, and age cohort as the primary independent variables. Other independent variables, the ones being analyzed, were the voting laws, the coronavirus data, and the data about state margins. The dependent variable here is turnout for the 2020 presidential election. A multiple linear regression is an approach to modelling the relationship between the independent variable and these dependent variables with a linear predictor function [19]. The basic form of a linear predictor function *f(i)* for a data point *i* using *p* independent variables is equation 1, where x_{ik} is the value of the *k*-th independent variable for *i*,

$$f(i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \tag{1}$$

and $\beta_0, \ldots, \beta_p$ are the coefficients indicating the effect of each independent variable on the dependent variable [27]. The linear predictor function can also be written in form of equation 2, where β is the vector composed of the scalars $\beta_0, \ldots, \beta_p$, x_i is the vector composed of the

$$f(i) = \beta \cdot x_i \tag{2}$$

independent variables $x_{i0}, x_{i1}, \ldots, x_{ip}$ for the data point *i* where x_{i0} is an extra pseudo-variable with a fixed value of 1 corresponding to the intercept coefficient β_0, and the operation is the standard dot product between vectors [27]. The relationship between the independent data point y_i for the data point *i* and this linear predictor function is in the form of equation 3, where ε_i is an error term which represents all other factors that influence the dependent variable other than the

$$y_i = \beta \cdot x_i + \varepsilon_i \tag{3}$$

independent variables contained in the model [27].

For the linear regressions in R, I used the lm function, which uses the ordinary least squares algorithm (OLS), which minimizes the sum of squares of the differences between the dependent variable and those predicted by the linear predictor function.

As too many independent variables in a model can cause overfitting, which is a phenomenon where an analysis corresponds too closely to a particular set of data, failing to predict other observations well, in an analysis using a large number of independent variables, it can be useful to get rid of some relatively insignificant independent variables [24]. To combat this, one can use the Akaike information criterion (AIC), which estimates prediction error and the quality of a model. AIC is calculated using formula 4, where *k* is the number of variables in

$$AIC = 2k - 2\ln(\hat{L}) \tag{4}$$

the model plus the intercept and $\hat{L}$ is the maximum likelihood function of the model, which, when using the least squares algorithm, is equivalent to the least squares estimation [24]. A model with a lower AIC value is preferred, containing fewer variables and having the likelihood function of the model be more predictive.

In R, one way to use this AIC metric is via the use of the stepAIC method for model selection [21]. This stepAIC function requires the use of the MASS and CAR packages for R. Using AIC, this method serves to simplify the model without changing its performance significantly. The algorithm finds the AIC value for the current model and then, for each independent variable, it removes them from the model and calculates the AIC value for that. Then, if removing a variable lowered the AIC, then the model with the lowest AIC is used, removing the worst variable. This continues, testing the current model against models that have one of the independent variables either added or removed, until the current model has the lowest AIC value. While this does not necessarily produce the "best" model, it produces a simple model that has a high level of performance [21].

Another obstacle to creating a well-fitting regression model is multicollinearity, which is when one of the independent variables in a model can be linearly predicted from other independent variables to a high degree of accuracy [20]. The way to quantify the severity of an ordinary least squares regression analysis is by using the variance inflation factor (VIF), which is a metric for each independent variable calculated using formula 5 for the factor identified by *i*. In

$$VIF_i = \frac{1}{1 - R_i^2} \tag{5}$$

this formula, R_i is the coefficient of determination for a linear regression model which models the independent factor *i* using all of the other independent variables [20]. A VIF of at least 10 is generally regarded as a high level of multicollinearity, implying that some independent variables should be removed, as they add very little additional information to the model [20]. One instance in which this appears is in the demographic data in this analysis. For example, knowing both the proportion of men and the proportion of women in a state is essentially useless since the proportion of one can be predicted with very high accuracy by the proportion of the other one. This also occurs with age ranges and racial demographic statistics.

Another metric used to measure the fit of the model is the residual: the difference between the observed value and the predicted value for a particular observation [1]. This is the ε_i term from equation 3. Smaller residuals for a term show that they more closely align with the model and imply that the model is more predictive. A standardized form of the residual is the residual divided by an estimate of its standard deviation. calculated by formula 6, where $\varepsilon_i$is the

$$r_i = \frac{\varepsilon_i}{s(\varepsilon_i)} = \frac{\varepsilon_i}{\sqrt{MSE(1 - h_{ii})}} \tag{6}$$

residual of the observation, MSE is the size of the mean square error, and h_{ii} is the leverage of the observation [1]. An observation with a standardized residual with an absolute value over 3 is sometimes deemed an outlier, although, technically, that term should be reserved for a *studentized* residual. The leverage is a measure of how distinct the independent variables of a certain observation are from the other independent variables. For example, Hawaii's unique racial makeup – over 35% Asian American, almost 10% Pacific Islander, and over 20% two or more races; all of which are the highest proportions in any state by a large margin – contributes to its high leverage in this analysis. The leverage score for a particular observation is defined by equation 7, where $[\mathbf{H}]_{ii}$ is the *i*-th diagonal element of the projection matrix **H** [26]. In turn, **H** is

$$h_{ii} = [\boldsymbol{H}]_{ii} \tag{7}$$

defined by equation 8, where **X** is the matrix where each row refers to a specific observation and

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \tag{8}$$

each column refers to one of the independent variables [26]. The value of leverage is bound between 0 and 1 inclusive [26].

Results

As indicated previously, to analyze the effects of voting laws, the COVID-19 pandemic, and the margin of victory on presidential turnout, I created a linear regression model using those variables in question, demographic data, and other results of the 2020 election. To investigate the correlations between these values, I calculated the correlations between the variables involved in the regression, as presented in figure 1.

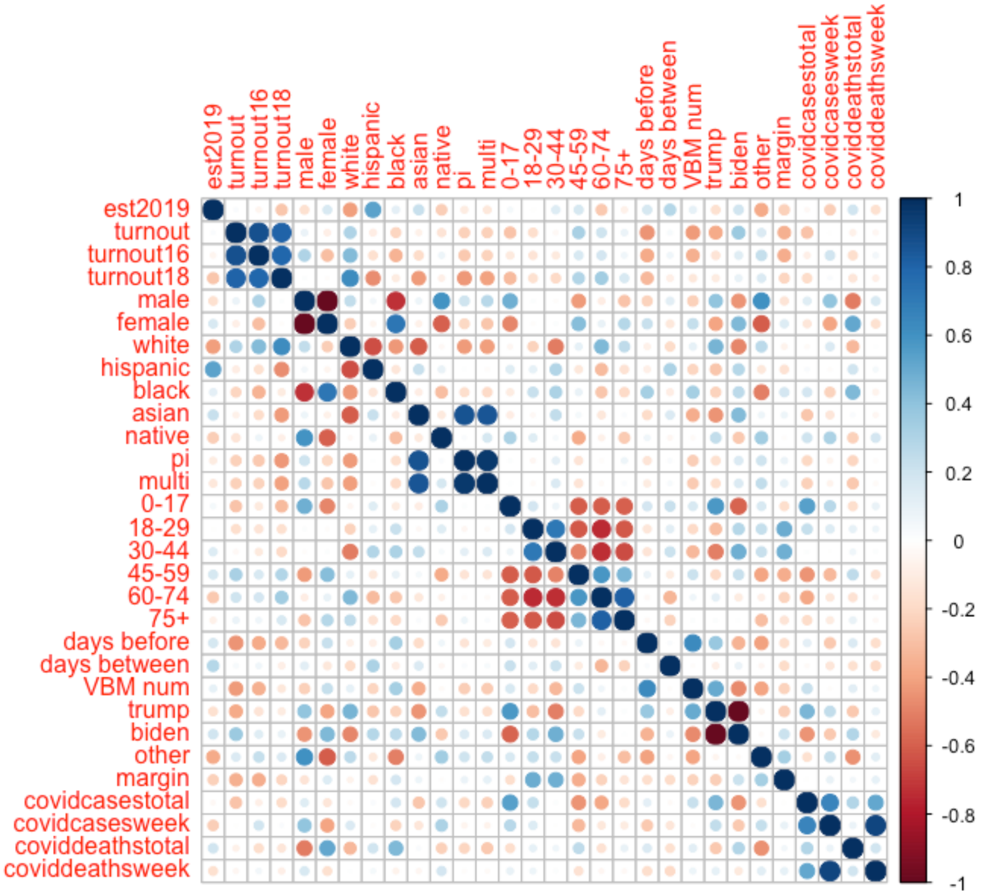

Figure 1: Correlation Plot

One can see clear areas of correlation between specific variables. For example, the most prominent of which are the turnout variables from recent elections, racial demographics, age demographics, and the severity of coronavirus outbreaks. As a consequence of this result, I decided to not use turnout data from previous elections in the linear regression to predict presidential turnout in 2020, since that value would overwhelm all others and would essentially turn the analysis into discovering the changes in turnout from 2016 or 2018 to 2020, which is categorically different from the total analysis of turnout that I would like to achieve here. I also decided to pare down the variables used in the initial regression, specifically variables that could be entirely predicted by other variables. Since many of my independent variables were proportions that added up to 1 – or extremely close to 1 – in each of those cases, I excluded one of the variables from the regression. In particular, I eliminated the ‘female’ (predicted by ‘men’), ‘multi’ (predicted by the other racial variables), ‘0-17’ (predicted by the other age variables), and ‘other’ (predicted by the ‘biden’ and ‘trump’ variables).

These eliminations left me with 22 independent variables, ‘male,’ six racial variables, five age cohort variables, three voting law variables, four coronavirus variables, ‘trump,’ ‘biden,’ and ‘margin.’ With these independent variables, I used the lm function in R to generate a multiple linear regression predicting 2020 turnout based on those variables. A summary of that regression, as given by R’s summary function, is as follows:

```
Residuals:
Min        1Q    Median        3Q       Max
-0.062843 -0.009811 -0.000018  0.012564  0.053099

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -8.949e-01  1.796e+00  -0.498 0.622251
male              2.083e+00  1.781e+00   1.170 0.251840
white             1.140e+00  9.598e-01   1.187 0.245113
hispanic          8.130e-01  9.292e-01   0.875 0.389070
```

```
black             1.069e+00  9.546e-01   1.120 0.272408    
asian             1.216e+00  1.139e+00   1.068 0.294821    
native            1.167e+00  1.191e+00   0.980 0.335512    
pi                5.616e-01  2.557e+00   0.220 0.827784    
`18-29`          -1.949e+00  9.062e-01  -2.151 0.040288 *  
`30-44`           8.964e-01  7.849e-01   1.142 0.263102    
`45-59`          -8.663e-03  8.671e-01  -0.010 0.992099    
`60-74`          -1.307e+00  9.576e-01  -1.365 0.183068    
`75+`             1.223e+00  1.297e+00   0.942 0.354056    
`days before`    -6.890e-04  5.200e-04  -1.325 0.195919    
`days between`   -2.639e-04  4.425e-04  -0.596 0.555794    
`VBM num`        -1.414e-02  8.041e-03  -1.758 0.089620 .  
trump            -3.166e-01  9.841e-01  -0.322 0.750035    
biden             1.664e-02  1.040e+00   0.016 0.987355    
margin           -2.135e-01  5.157e-02  -4.140 0.000288 ***
covidcasestotal  -2.816e-01  9.452e-01  -0.298 0.767965    
coviddeathstotal -1.269e+01  1.831e+01  -0.693 0.493812    
covidcasesweek   -1.725e+00  7.316e+00  -0.236 0.815265    
coviddeathsweek   1.448e+02  2.853e+02   0.507 0.615822    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02798 on 28 degrees of freedom
Multiple R-squared:  0.8716,     Adjusted R-squared:  0.7708 
F-statistic: 8.642 on 22 and 28 DF,  p-value: 1.814e-07
```

With a multiple R^2 value of 0.8716, this model is highly predictive, but with so many variables, only three variables have any statistical significance, ’18-29’, ‘VBM num’, and ‘margin.’ The first shows the depressing reality that young people vote at significantly low number, but those latter two show promise for the hypotheses of the analysis. The model indicates that stricter vote-by-mail laws produces a weak negative effect on turnout and produces an extremely solid indication that wider margins lead to lower turnout – and, inversely, that closer races lead to higher turnout. However, with so many independent variables, many of the variables’ likely exhibit significant collinearity, as indicated in the correlation plot of figure 2. Since they are able to be predicted by other variables, eliminating some of them may reveal more significant effects that are currently disguised by the wealth of variables. To analyze this collinearity, I will use the vif function in R on the model to calculate VIF for the independent variables. The results of this function are as follows:

```
male            14.123801
white           1539.718094
hispanic        591.254345
black           640.025917
asian           237.197503
native          74.477736
pi              73.067342
`18-29`         7.545691
`30-44`         8.767477
`45-59`         7.572895
`60-74`         16.705728
`75+`           9.990757
`days before`   2.675159
`days between`  1.711530
`VBM num`       3.579344
trump           891.898605
biden           994.744030
margin          3.693364
covidcasestotal 8.439575
coviddeathstotal3.463716
covidcasesweek  40.577899
coviddeathsweek 24.831158
```

Since a VIF over 10 is deemed is generally considered too high, it is evident that this model has a huge amount of unnecessary independent variables, particularly among the age groups, the Trump and Biden vote proportions, and the coronavirus data [20].

As indicated previously, I will use the stepAIC function for model selection to choose a simpler and more streamlined model from this initial model. When, executed, the stepAIC calculated an initial AIC of -349.37, and eliminated the ’45-59’ variable first. Subsequently, it eliminated the ‘biden’, ‘pi’, ‘covidcasesweek’, ‘covidcasestotal’, ‘days between’, ‘coviddeathstotal’, ‘75+’, 'coviddeathsweek', and ’60-74’ independent variables, bringing the model to an AIC of -364.28. This smaller model consists of, in order of the next in line to be eliminated if stepAIC were allowed to continue on, ‘male’, ‘days before’, ‘native’, ‘asian’, ‘VBM num’, ’30-44’, ‘hispanic’, ’18-29’, ‘black’, ‘white’, ‘trump’, and ‘margin’. The R generated summary of this model consists of the following:

```
Residuals:
      Min        1Q    Median        3Q       Max 
-0.071284 -0.008286  0.000807  0.009308  0.064665 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -0.7210895  0.5537768  -1.302 0.200710    
male           1.6147663  1.2210553   1.322 0.193928    
white          0.8806834  0.2706556   3.254 0.002393 ** 
hispanic       0.5834995  0.2491229   2.342 0.024510 *  
black          0.7912341  0.2653613   2.982 0.004981 ** 
asian          0.9981169  0.4860569   2.053 0.046953 *  
native         0.8179242  0.4105578   1.992 0.053567 .  
`18-29`       -1.3950539  0.4706681  -2.964 0.005219 ** 
`30-44`        1.2083687  0.5303799   2.278 0.028417 *  
`days before` -0.0008016  0.0004151  -1.931 0.060943 .  
`VBM num`     -0.0132343  0.0061027  -2.169 0.036440 *  
trump         -0.2856721  0.0664686  -4.298 0.000116 ***
margin        -0.2165601  0.0342808  -6.317 2.09e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02525 on 38 degrees of freedom
Multiple R-squared:  0.8581,    Adjusted R-squared:  0.8133 
F-statistic: 19.16 on 12 and 38 DF,  p-value: 1.583e-12
```

Even after having eliminated almost half of the independent variables, the multiple R^2 value is of 0.8581 is just below the larger model’s R^2 of 0.8716. This improved is indicated by the adjusted R^2 value of the respective models, which is a metric that penalizes models for additional independent variables. This smaller model has a higher R^2 value, 0.8133, than that of the larger model, 0.7708. Additionally, nearly every independent variable has a statistically significant effect on turnout.

The fact that stepAIC eliminated every single covid-related variable indicates that the second hypothesis is likely invalid. It also eliminated the 'days between' variable which measured the length of early voting periods. On the other hand, the 'margin', 'VBM num', and 'days before' variables all remained, and all have a p-value less than 0.10, indicating at least some level of statistical significance.

Figure 2 shows the high level of alignment between the model's predicted turnout levels and the actual turnout of the election, with 49/51 states (and DC) having turnout within ±5% of the predicted values. The only exceptions were Florida (with an actual turnout 6.47% higher than predicted) and Nevada (with an actual turnout 7.13% lower than predicted).

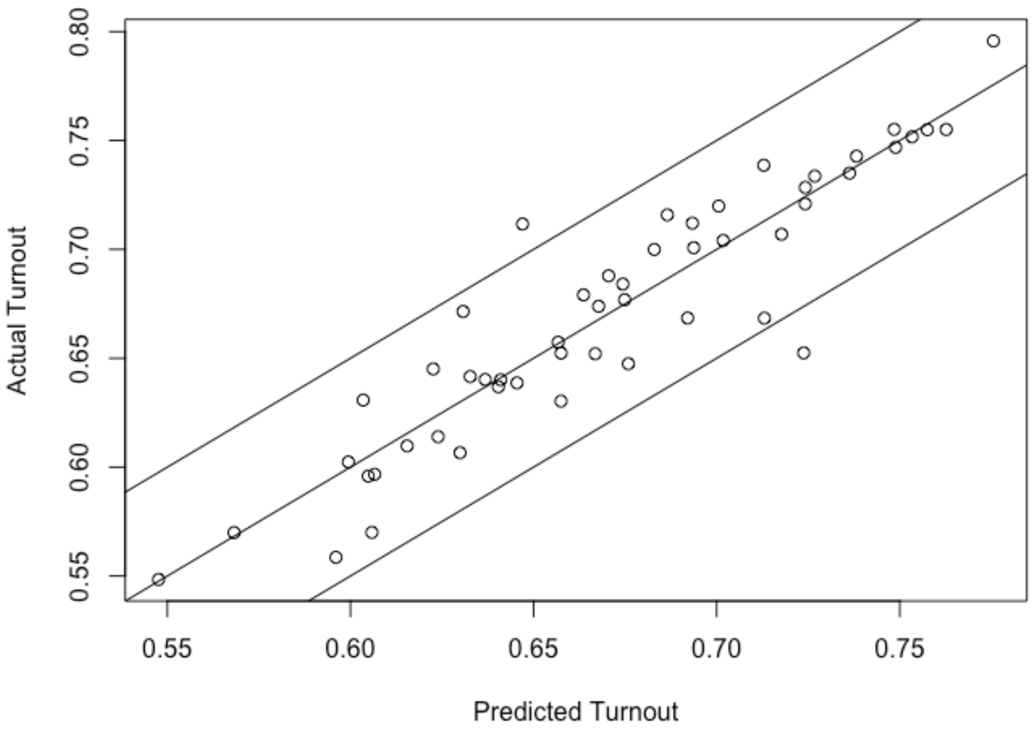

Figure 2: Turnout Predicted by Model vs. Actual Turnout

Figure 3 shows plots of the residuals of the turnout values plotted against the fitted values of the model. Similarly, figure 4 shows a plot of the residuals against the leverages of the expected turnouts for each of the states.

These plots show how the error of the model remained consistent through most of the states, with no specific level of residual, fitted value, or leverage having disproportionate errors. As mentioned earlier, an observation with a standardized residual value with an absolute value above 3 can be considered an “outlier,” and Florida and Nevada are the only two values that even approach that threshold [1]. In terms of leverage, as described earlier, Hawaii has an extraordinarily high leverage, with its anomalous demographic data.

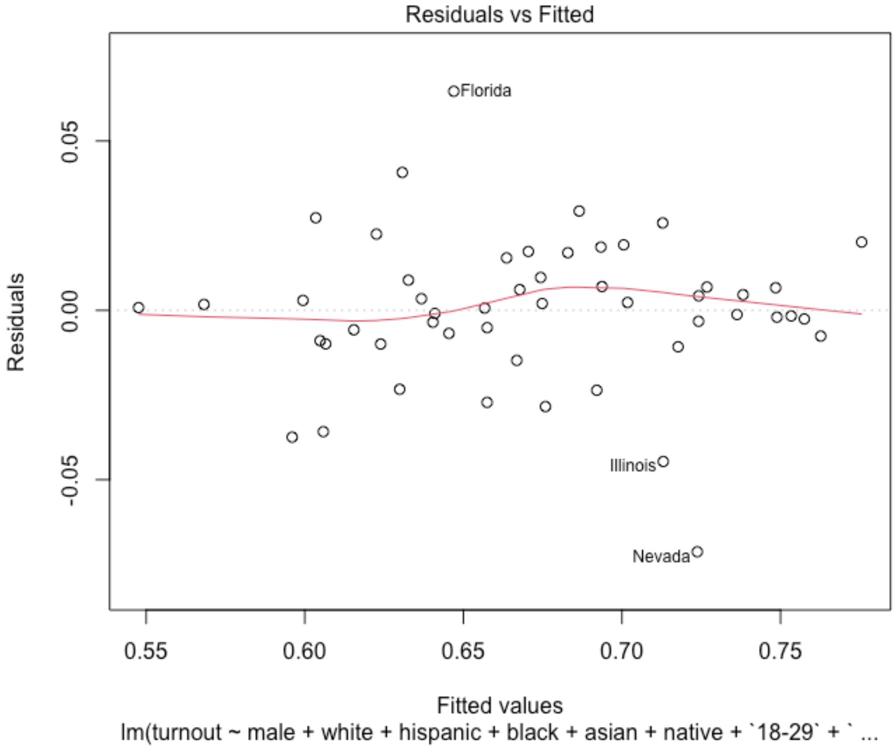

Figure 3: Residuals vs. Fitted Values

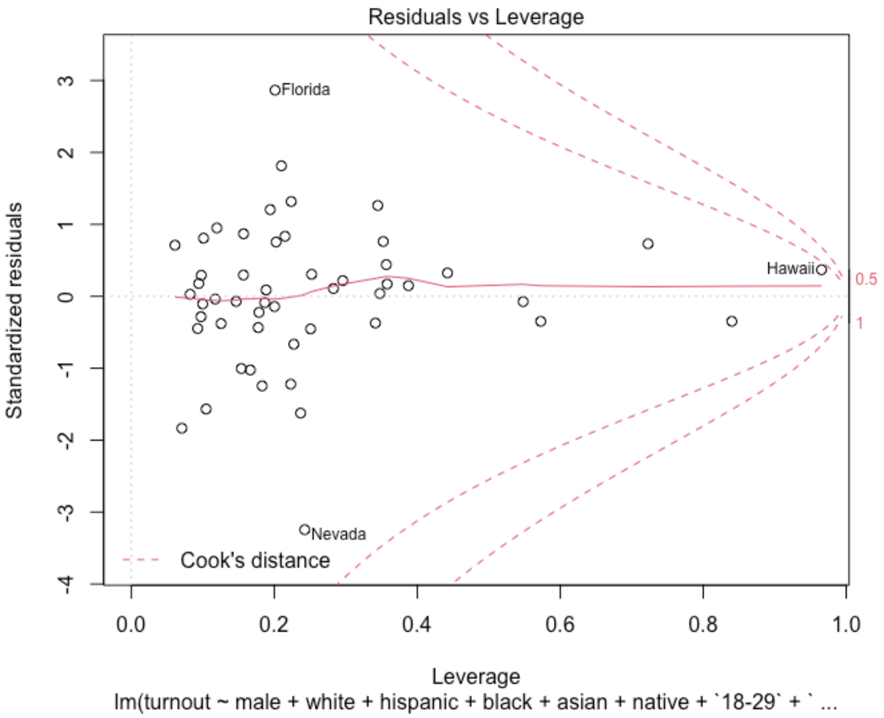

Figure 4: Residuals vs. Leverage

In figure 5 and figure 6, the relationships between turnout and margin are plotted, with each state also revealing the class of its vote-by-mail laws. Figure 5 possesses all 50 states + DC, but since Biden won the District of Columbia by over 85%, it is so far from the rest of the data points, making the plot difficult to read or interpret. Consequently, figure 6 simply zooms in on the part of the plot with the remaining 50 states.

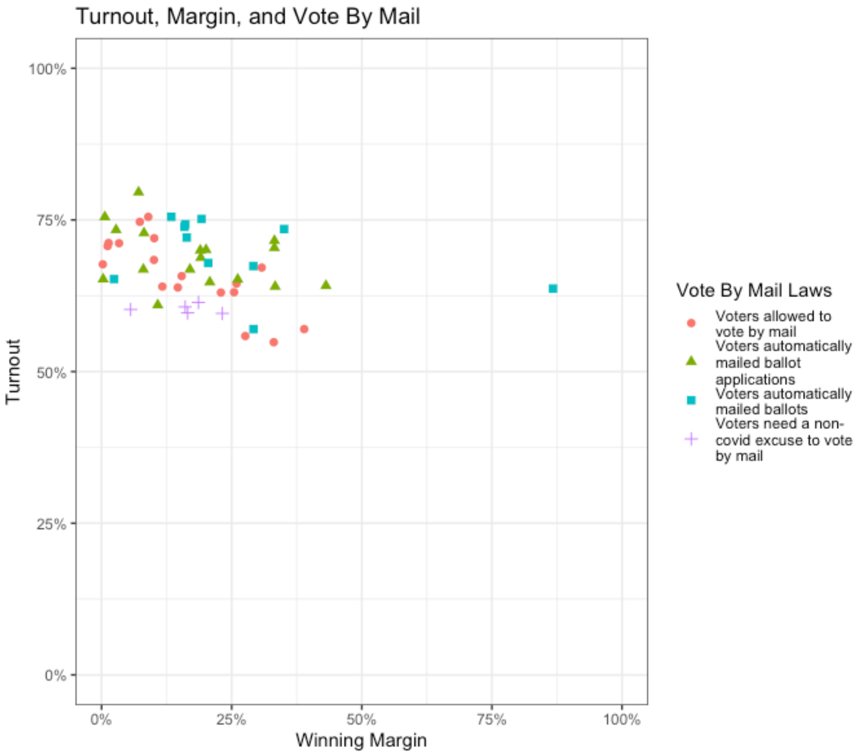

Figure 5: Turnout vs. Margin Plot with Vote By Mail Rules

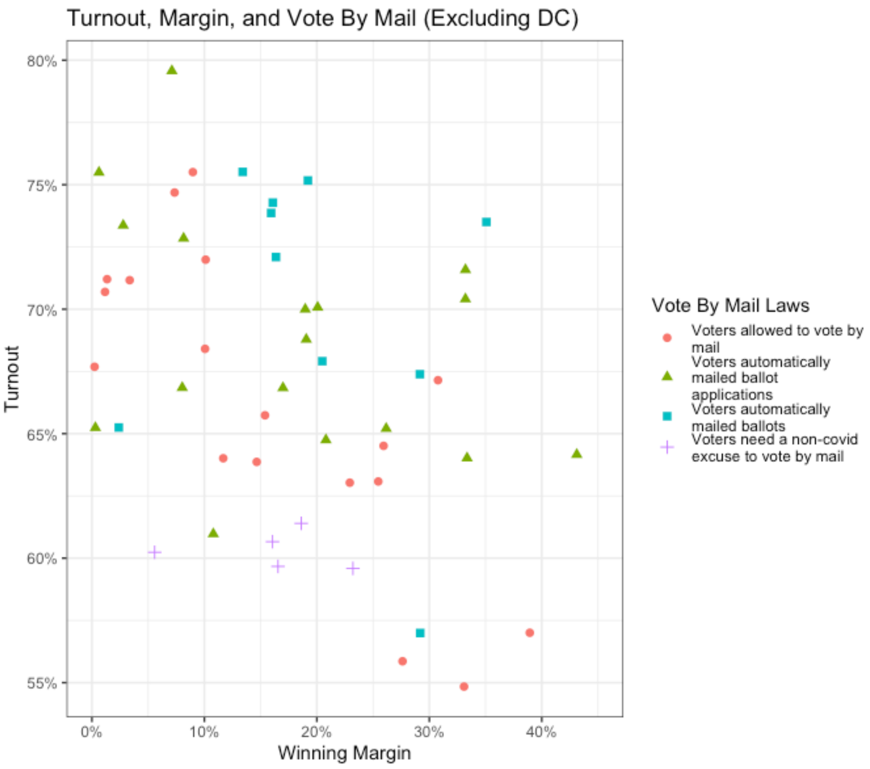

Figure 6: Turnout vs. Margin Plot with Vote-By-Mail Rules (excluding DC)

In the zoomed in plot of figure 6, the states where voters were automatically mailed ballots, represented by squares, are clustered at the top of the graph (with higher turnout), and that those states where one needed a non-coronavirus excuse to be allowed to vote-by-mail were clustered in an area of very low turnout.

The p-values for ‘margin’ and ‘VBM num’ were less than 0.05, showing that the corresponding hypotheses were correct. In particular, the p-value of ‘margin,’ $2.09*10^{-7}$, provides a huge indication that closer margins of victory lead to higher turnouts. In particular, the model predicts that, with 95% confidence, a state having a 1% closer margin will lead to an increase in turnout of 0.15%-0.29%. This would in turn correspond to an election that is 10% closer leading to a turnout increase of ~2%. This is a large increase in a nation where turnout levels only vary by about 20% between the best and worse states. While the model was less confident in ‘VBM num,’ giving it a p-value of only 0.036, it also indicated a significant effect on turnout. This value indicates with 95% that switching from the most restrictive (requiring a non-COVID excuse) to the least restrictive (mailing ballots to all voters) will increase turnout by between 0.3% and 7.6%. The model was even less confident in the ‘days before’ variable, having a p-value of 0.061, which is too high to fit the traditionally accepted 0.05 p-value threshold. Further research is needed, but this could possibly indicate a relationship between higher turnout and registration deadlines closer to election day. With 95% confidence, the model indicates that allowing registration for an additional week will result in an effect between 1.14% higher turnout and 0.02% lower turnout. All the other variables in question had very little observed effect on turnout, being eliminated by the stepAIC algorithm, indicating that, overall, their effects on turnout are insignificant.

Conclusion

The linear regression model constructed and exhibited above supports the hypotheses that having tighter presidential races and more permissive vote-by-mail laws raised turnout in states in the 2020 presidential election. However, it did not support the hypotheses regarding the severity of coronavirus outbreaks lowering turnout or other liberalized voting rules raising turnout.

This result vindicates the work of activists and politicians campaigning to make vote-by-mail easier, and should discourage the rhetoric of those who say that this sort of voting liberalization only displaces in person voters instead of adding new ones. This result also places the focus on the Electoral College and the phenomenon of swing states that it creates. Since the turnout differentials between solidly blue or red states and swing states are so large, this system decreases political participation in most areas of the country while increasing it in a select few. Surprisingly, COVID did not hurt turnout as expected. This could indicate that the efforts taken by governments to prepare for the election and give people more options were effective and solved the problem, but it could also indicate that partisanship is so overwhelming that even life or death situations will not dissuade voters from voting for their beloved heroes and against their hated enemies.

Regardless, if one is interested in increasing electoral participation, studying the successes of the vote-by-mail system while reckoning with the peculiarity of the Electoral College is a valuable exercise to see how Americans actually participate in politics.

Works Cited

[1] "9.3 - Identifying Outliers (Unusual Y Values) | STAT 462." *Eberly College of Science*, The Pennsylvania State University, 2018. https://online.stat.psu.edu/stat462/node/172.

[2] D'Angelo, James and Ranalli, Brent. "How the Secret Ballot Ended the Gilded Age." *The Congressional Research Institute*, 9 Aug. 2020. https://congressionalresearch.org/SecretBallot.html.

[3] Desilver, Drew. "In past elections, U.S. trailed most developed countries in voter turnout." *Pew Research Center*, 3 Nov. 2020. https://pewrsr.ch/2LjNokk.

[4] Desilver, Drew. "Turnout soared in 2020 as nearly two-thirds of eligible U.S. voters cast ballots for president." *Pew Research Center*, 28 Jan. 2021. https://pewrsr.ch/3oAN3MB.

[5] Galston, William A. "Election 2020: A Once-in-a-Century, Massive Turnout?" *Brookings*, 14 Aug. 2020, www.brookings.edu/blog/fixgov/2020/08/14/election-2020-a-once-in-a-century-massive-turnout.

[6] McDonald, Michael P. "2016 November General Election Turnout Rates." *United States Elections Project*. http://www.electproject.org/2016g. Accessed 26 Apr. 2021.

[7] McDonald, Michael P. "2018 November General Election Turnout Rates." *United States Elections Project*. http://www.electproject.org/2018g. Accessed 26 Apr. 2021.

[8] McDonald, Michael P. "2020 November General Election Turnout Rates." *United States Elections Project*. http://www.electproject.org/2020g. Accessed 26 Apr. 2021.

[9] McDonald, Michael P. "Frequently Asked Questions." *United States Elections Project*. http://www.electproject.org/home/voter-turnout/faq.

[10] McDonald, Michael P. 2021. “National General Election VEP Turnout Rates, 1789-Present.” *United States Elections Project*. http://www.electproject.org/national-1789-present. Accessed 26 Apr. 2021.

[11] McDonald, Michael P. [@ElectProject]. “I posted PRELIMINARY estimates of the 2020 state and national turnout and voting-eligible population turnout rates 160 million people voted …” Twitter, 4 Nov. 2020. https://twitter.com/ElectProject/status/1323897443398942726.

[12] MIT Election Data and Science Lab. “U.S. President 1976-2020.” *Harvard Dataverse*, 2017. https://doi.org/10.7910/DVN/42MVDX. Accessed 29 Apr. 2021.

[13] Montellaro, Zach. “Pandemic Threatens Monster Turnout in November.” *POLITICO*, 31 Mar. 2020, www.politico.com/news/2020/03/31/states-struggle-voting-pandemic-155700.

[14] The New York Times. “Coronavirus (Covid-19) Data in the United States.” *GitHub*, 2021. https://github.com/nytimes/covid-19-data. Accessed 29 Apr. 2021.

[15] Paskert, Michael, "Effects of Voting Behavior and Voter Turnout" (2014). *Senior Honors Projects*, 44, John Caroll University https://collected.jcu.edu/honorspapers/44.

[16] Rakich, Nathaniel, et al. “How To Vote In The 2020 Election.” *FiveThirtyEight*, ABC News Internet Ventures, 2 Nov. 2020. https://projects.fivethirtyeight.com/how-to-vote-2020/. Accessed 8 Apr. 2021.

[17] Reed, Daniel C. “Reevaluating the Vote Market Hypothesis: Effects of Australian Ballot Reform on Voter Turnout.” *Social Science History*, vol. 38, no. 3-4, 2014, pp. 277–290. JSTOR, www.jstor.org/stable/90017036. Accessed 28 Apr. 2021.

[18] Santana, Andrés, et al. “The Coronavirus Pandemic and Voter Turnout: Addressing the Impact of Covid-19 on Electoral Participation.” *SocArXiv*, 18 Nov. 2020. https://osf.io/preprints/socarxiv/3d4ny/. Accessed 9 May 2021.

[19] Stuart W Grant, Graeme L Hickey, Stuart J Head, Statistical primer: multivariable regression considerations and pitfalls, *European Journal of Cardio-Thoracic Surgery*, Volume 55, Issue 2, February 2019, Pages 179–185, https://doi.org/10.1093/ejcts/ezy403.

[20] Tripathi, Ashutosh. “What Is Multicollinearity?” *Data Science Duniya*, 15 June 2019, https://ashutoshtripathi.com/2019/06/13/what-is-multicollinearity.

[21] Tripathi, Ashutosh. “What Is StepAIC in R?” *Data Science Duniya*, 10 Apr. 2021, https://ashutoshtripathi.com/2019/06/10/what-is-stepaic-in-r.

[22] United States Census Bureau. “State Population by Characteristics: 2010-2019.” *Census.gov*, Jun. 2020. https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-detail.html. Accessed 28 Apr. 2021.

[23] Waxman, Olivia B. “The History Behind 2020’s Record Voter Turnout Numbers.” *Time*, 5 Nov. 2020. https://time.com/5907062/record-turnout-history/.

[24] Wikipedia contributors. "Akaike information criterion." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 17 May. 2021. https://en.wikipedia.org/w/index.php?title=Akaike_information_criterion&oldid=1023710107.

[25] Wikipedia contributors. "2020 Democratic Party presidential primaries." *Wikipedia, The Free Encyclopedia*, 7 May. 2021. https://en.wikipedia.org/w/index.php?title=2020_Democratic_Party_presidential_primaries&oldid=1021974155.

[26] Wikipedia contributors. "Leverage (statistics)." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 28 Mar. 2021. https://en.wikipedia.org/w/index.php?title=Leverage_(statistics)&oldid=1014652063.

[27] Wikipedia contributors. "Linear predictor function." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 3 Dec. 2020. https://en.wikipedia.org/w/index.php?title=Linear_predictor_function&oldid=992098633

[28] Wikipedia contributors. "Voter turnout in United States presidential elections." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 19 May. 2021. https://en.wikipedia.org/w/index.php?title=Voter_turnout_in_United_States_presidential_elections&oldid=1024023389.