PREDICTING THE POPULAR VOTE FOR UNITED STATES PRESIDENTIAL ELECTIONS: FUNDAMENTAL VARIABLES VERSUS POLITICAL VARIABLES

Valeria Werner

June 2021

A thesis submitted to the department of

Mathematical Methods in the Social Sciences (MMSS)

Advised by Professor John Bullock

Northwestern University

Table of Contents

Abstract

Presidential Elections can be forecast using a variation of methods. Statisticians, politicians, and many others attempt to predict elections using a variety of methods and variables. Specifically, there are two prevailing methods of predicting US presidential elections which are explored in this paper: fundamentals based models and political variables based models. The paper highlights the advantages and disadvantages of both types of variables by building two separate models by using the same data set. The model consists of two stages: a bivariate regression stage to weed out variables with less predicting power and a multilinear regression model to put together the remaining variables. Ultimately, the model demonstrates that around 40% and 75% of voting behaviour can be predicted by fundamental variables and political variables respectively.

Acknowledgements

I would first like to thank my parents, for reminding me that “the best thesis is a finished thesis;” I did it. Although my work cannot come close to the caliber of their work, I would not have been able to do any of this without their love, support, and endless guidance. To my friends and classmates, thank you for all your support and help throughout my undergraduate career, for the long nights in the library, the endless hours on zoom, and everything in between. And, for offering to show up to my final presentation even when it was virtual. Lastly, thank you to my advisor, Professor John Bullock — thank you for taking the time to read everything I wrote, give feedback, and answer any question I had, all despite having never met me in person.

1. Introduction

Election prediction is an extremely relevant and rapidly evolving topic. Its importance in policy, law, economics, and many more subjects makes its study extremely popular. There exist two main categories of variables — political and fundamental — which shape most of the existing models. There have also been plenty of studies outlining the flaws and shortcomings of different models. In the current literature, a major limitation is the lack of comparison between the two different types of variables and how their respective models perform when compared to one another. My thesis will build on existing models and studies that already outline the best practices for fundamentals-based models and politically rooted models by building a regression that compares which type of variables has more predictive power when it comes to the popular vote of US presidential elections. Ultimately, I will aim to show which variables within the categories have the most predictive power and which of the two categories of variables is most important in predicting elections.

The US electoral system is particularly interesting given the existence of the electoral college which creates a gap between the popular vote and the actual election results. This discrepancy led to controversial election results, like the 2000 and 2016 elections, which inspired many professionals to tackle the issue of predicting the outcomes of future elections. Especially in 2016, when poor interpretations of data, errors in sampling, and a general underestimation of Donald Trump's campaign led many to wrongly predict election results by a wide margin. Furthermore, we live in an environment where we consume news and information 24/7 — our phones travel with us wherever we go and constantly update us with news headings, political ads, polling results and others' opinions. The 24/7 news cycle affects campaign strategies, public perception, and most importantly (for the purpose of this paper) people's voting behaviours.

The problem when it comes to predicting elections is that there are a plethora of variables all of which could influence an election. Some researchers focus on national polls and infer electoral results; others claim that more accurate results are yielded when state level data is interpreted and then applied to the overall electoral result. Although there is a lot of research on the subject that focuses on fundamental polling variables or political polling variables, there is not much research comparing the two methods in one study. The goal of this research paper is to compare methods of predicting elections and determine which specific variables are actually most important in determining the outcome of US presidential elections: fundamental polling variables, or political ones.

2. Literature Review:

2.1. Fundamentals Based Election Prediction

A fundamentals-based model consists of variables that cannot be rationalized. Prediction models that only use fundamental variables essentially allude to the idea that elections are predetermined and the specific candidates, or the campaigns they run have little to do with the ultimate outcome of the election. The use of fundamental variables has been well studied and documented; the effect of these variables on polls and election results was documented by Gelman and King (1993) where they concluded that although the 'horse-race' gets a lot of media attention people ultimately make their decisions based on 'enlightened preferences'. Ultimately, these preferences are the public's beliefs about fundamental variables after hearing the campaigns and perspectives of either candidate (Gelman & King, 1993).

Building on Gelman and King, the 'Bread and Peace' Model constructed by Douglas A. Hibbs, Jr (2000) took into account 24 different variables including unemployment and inflation, but ultimately determined that the weighted-average growth of per capita real personal disposable income over the term and US military fatalities due to unprovoked, hostile deployments of American armed forces were the only two variables that were necessary to predict election results. Similarly, when evaluating the 2000 US presidential election, Bartels and Zaller (2000) concluded that the most significant variable in predicting the results was RDI, but the large standard errors make conclusions about the magnitude of the effect of RDI on actual election results less capable of predicting elections. Later, in 2008, Hibbs used his model to try to predict the presidential election — he predicted Obama would win the popular vote by 7.5

percentage points; he won by 7.2 percentage points (Hibbs, 2008). These studies highlight a gap and a flaw in current literature: if the campaigns help voters form their 'enlightened preferences' then there are political variables at play that influence whether a candidate can manipulate these preferences in their favor. Additionally, the results from fundamentals-based models are not always reliable or accurate — Nate Silver performed a study evaluating 58 different models and concluded that across all the models including only fundamental variables the average difference between the actual outcome and the predicted outcome was 9.8 percentage points, in models that included 'horse-race' variables the difference was only 6.9 percentage points (Silver, 2012). Furthermore, the US is not always involved in unprovoked conflict so there is a gap in exploring if other forms of civil unrest lead to the same election predictions.

2.2. Election Prediction Using Political Variables

Political variables and models include a lot more variation and have some more subjective variables. Researchers have long tried to predict election results using variables such as national polls, state polls, debate results, and many more; a ranking assumption model will be used by Fair (2020) to try and predict this year's election. This model takes the vote in each state for the Democratic candidate for president; then it ranks the states by the probability that the candidate wins the state, if he wins state *i*, he wins every state ranked above state *i*. Then the probability that the Democratic candidate wins the election is the probability that they win state *i,* where state *i* is the pivot state that secures 270 electoral votes. Similarly to this model, Silver (2011) talks about Allan Lichtman's prediction in his 1984 book "The Keys to the White House". Lichtman's model consists of 13 key factors and if at least eight of them are scored in the incumbent's favor, then he will win the election. Lastly, Lauderdale and Linzer (2015) argue that due to discrepancies between the electoral college and the popular vote, fundamentals based models are inconsistent and overstated. Instead, they used a Bayesian forecasting model at the state level to try to predict electoral results, but they concluded that it is not possible to make very accurate predictions about election results.

All three studies have strengths and weaknesses. They all focus on either political variables or a mixture of fundamental and political variables, but none of them compare the separate effects of both categories. Additionally, the 13 key factors include very subjective factors that affect the results of the model (Silver, 2011) and there are different regression techniques, such as random effects, lasso regressions, and other Bayesian statistic tools referenced by Lauderdale and Linzer that can be performed to attempt to get better results. Finally, the focus on state-level data reveals a few factors about specific swing states, but when considering a national election it is important to take into account the interstate correlations.

2.3. Accuracy and Quality of Existing Models

Finally, a large portion of the literature surrounding the subject of predicting US presidential elections is about evaluating and critiquing existing models and the predictive results they yield. In evaluating their own forecasts FiveThirtyEight determined that some of them are pretty inaccurate and do not add much value; nevertheless, their political forecasts have a higher accuracy than their fundamentals based models because they have a better sense of what affects the outcomes (Boyce & Wezerek, 2019). This evaluation method is not that great given that they compare their models predicting Major League Baseball (MLB) game outcomes to those predicting the outcomes of the U.S. House elections; and these two methods are not comparable. In order to do this they present the idea of ‘unskilled’ models which simply take averages of historic outcomes. But these ‘unskilled’ models do not exist for US House elections, as the candidates are not always the same and the article only presents the results of the ‘unskilled’ models for MLB games. In critiques of FiveThirtyEight models, studies highlight how incentives can shape a model’s forecast, and how they change the way results and uncertainty are presented (Gelman, Hullman, Morris, & Wlezien, 2020). They explain that presenting data in an overconfident, or underconfident way can affect the way we perceive election polls and predictions. The way certain questions can be phrased in polls, the samples of the population that are taken, and normative statements can be misleading and misinforming to the public. Other critiques state that underestimating the correlation of uncertainties between states can lead to extremely weird and seemingly impossible outcomes (Gelman, 2020). In FiveThirtyEight’s 2021 Election Forecast, there existed a scenario where Biden won in Idaho, Wyoming, and Alabama; but not New Jersey. Andrew Gleman explains that this forecast is possible if the correlation of uncertainties between states is too low and the forecasting relies too heavily on state-level polls (2020). Several studies analyze different forecasts and critique their models, as well as attempt to explain why certain models failed. Nonetheless, there is a lack of analyses in which differing models are critiqued side by side to see which gives more weight to what type of variable.

3. Data

All of the data was collected in January 2021. The data was collected for the incumbent candidate, as I am attempting to predict the share of the popular vote obtained by the candidate of the incumbent party. The main intention behind data collection was to go as far back as possible while still being able to collect data on all variables of interest for every election cycle. All data concerning the economic and financial state of the country was obtained through the Federal Reserve Bank of St. Louis. The data concerning mortality and death rates was obtained from the United Nations World Population Prospects, the presidential approval ratings were obtained from The American Presidency Project at the University of California, Santa Barbara, and the polling data was drawn from a few different sources — Gallup Poll, FiveThirtyEight, and RealClearPolitics. To make sure all data was available for every election, the data set only goes as far back as the election of 1960, the last election made available by all sources.

All together the data set spans 60 years and 16 election cycles. It consists of four fundamental variables: inflation, all-cause mortality, Real Disposable Income, and unemployment. Each of these variables is represented in several different ways throughout the data set, including as a descriptor of change, a percentage, or even different methods of weighting different quarters depending on their proximity to an election. Specifically, the data set includes annual and third-quarter statistics for all economic fundamental variables, as well as four different methods of weighting all four quarters in the year preceding an election for the independent variable of Real Disposable Income.

The data set also contains two core political variables: polling numbers, and approval ratings of the current President for the incumbent party. Once again, both variables are segmented in various ways to represent the weights of polls and approval rating closer to the date of the election. Specifically, the polls are represented as a year-long average and an average for only October and November polls of the election year, and approval ratings are documented for the entire year preceding the election, as well as for the quarter preceding the election.

3.1. Variables

3.1.1. Fundamental Variables

As stated previously, the four categories of variables available for the fundamentals based model are: all-cause mortality, inflation, unemployment, and Real Disposable Income (RDI). All-cause mortality is represented in two different ways; the first is simply the number of deaths in the year of the election per 10,000 people due to any cause — this includes US military fatalities owing to hostile deployments of American armed forces in foreign conflicts, deaths due to heart attacks, cancer, and anyone who dies in the US population regardless of cause. The second representation is the percent change in annual all-cause mortality per 10,000 people from one election cycle to another — documented on a scale from -100 to 100.

The second category of fundamental variables is inflation and it is only represented in one way. Inflation in the data is measured by the consumer price index and reflects the annual percentage change in the cost to the average consumer of acquiring a basket of goods and services yearly. The variable is not seasonally adjusted and was calculated by the World Bank.

The unemployment rate in the data set is represented in two separate ways. Both ways represent the number of unemployed people as a percentage of the labor force, which is restricted to people over 16 years of age. The measure of the labor force does not include people who reside in institutions (e.g., prison, mental institutions, or assisted living facilities). The first representation of unemployment in the dataset is the annual unemployment rate for the year of the election in question. The second measure of unemployment is the stand-alone third-quarter unemployment rate as measured on the year of the election from June 1st to September 30th of the same year.

The fourth, and last, category of fundamental variable is RDI represented 12 different ways, two different representations for six different calendar weights. The first representation of RDI is in billions of chained 2012 dollars at a seasonally adjusted annual rate. The second representation is the percent change from a year ago of billions of chained 2012 dollars at a seasonally adjusted annual rate. The first calendar weight for the RDI measures —represented both in billions and as a percent change, is an annual measure, which equally weights all 4 quarters of the election year. The second variation of RDI measures, is the third-quarter RDI which simply takes the same two statistics for the quarter preceding the election. This is similar to the third-quarter unemployment rates explained above. The other four calendar weights are defined as weighted RDIs (or WRDIs) and they take all four quarters of RDI and place different weights on the quarter before the election versus the three quarters preceding it. The four measures of weighted RDIs are calculated such that the weight on the third-quarter increases for everyone — meaning that the second WRDI weighs the third-quarter RDI more heavily than the first WRDI, and the third WRDI weighs the third-quarter RDI more heavily than the second WRDI, and so on. The specific weights of the four different measures of weighted RDIs are outlined below:

$$\text{WRDI\_1} = 0.8 \times \left(\frac{1}{3}\left(\sum_{i=1}^{3} x_i\right)\right) + 0.2 \times x_4$$

(1)

$$\text{WRDI\_2} = 0.6 \times \left(\frac{1}{3}\left(\sum_{i=1}^{3} x_i\right)\right) + 0.4 \times x_4$$

(2)

$$\text{WRDI\_3} = 0.4 \times \left(\frac{1}{3}\left(\sum_{i=1}^{3} x_i\right)\right) + 0.6 \times x_4$$

(3)

$$\text{WRDI\_4} = 0.2 \times \left(\frac{1}{3}\left(\sum_{i=1}^{3} x_i\right)\right) + 0.8 \times x_4$$

(4)

Where $x_i$represents the four quarters leading up to an election, and $x_4$is the quarter directly preceding it (i.e. the third-quarter of the calendar year of the election, or third-quarter RDI).

3.1.2. Political Variables

The two categories of political variables in the dataset are approval ratings for the sitting president, and polling numbers. The approval ratings were simply what percentage of the sampled population at the time of the poll approves of the current president. The polls were documented quarterly as a percentage by the American Presidency Project at the University of California, Santa Barbara. They are represented in two different ways in the dataset, the first is as an annual approval rating, which is simply the average of the approval ratings for the four quarters. The second way it is represented is simply as the third-quarter approval rating which is the percentage reported for the quarter preceding the election. The polling numbers are also represented in two ways. The first, is an annual average of all the polls for the year preceding the election taken by Gallup Poll, FiveThirtyEight, or RealClearPolitics, depending on the year. The second is the average of only polls taken in the months of October and November, right before the election.

3.2. Data Collection Methods

As described previously, the data set was collected by piecing together a variety of different sources. For the data obtained through the Federal Reserve Bank of St. Louis, The American Presidency Project at the University of California, Santa Barbara, and the United Nations World Population Prospects it was publicly available and datasets were downloaded from their websites. For all other sources, a web scraper was built using Python and the data was collected directly from their websites.

3.3. Advantages of the Dataset and Results

The data set is extremely expansive for fundamental variables; it covers a wide range of economic variables that could affect public perception of an incumbent's ability to lead the nation. Additionally, it includes several ways for all variables to be accounted for and provides ample opportunities to demonstrate which specific characteristic of fundamental variables affects election results (i.e. change year to year, raw numbers, or change from the previous presidency). Additionally, the dataset continues to build on previous research by including the statistics for the 2020 presidential election; therefore creating a more robust and expansive dataset. Especially given that more recent elections document polling, predictions, and general statistics more carefully than before and they provide a plethora of information to work with.

3.4. Disadvantages of the Dataset and Results

The data set also has several drawbacks. Although it spans over 50 years, there are only 16 concrete data entries given that those were the only years when primary presidential elections in the United States happened. Furthermore, the data focuses heavily on fundamental variables, instead of evenly on both types of desired variables. This is mainly due to the fact that historically more data is collected on the economy in any given year than about politics. Additionally, daily polling trackers and detailed documentation on approval rating are relatively new phenomena and there is a large discrepancy between the political data available for the 1960s versus now. Nevertheless, in order to truly be able to compare their respective predictive powers, it would be ideal to have a similar number of variables for each category. This flaw comes with one more pitfall, which is that experts have disagreed on the efficiency of using fundamental variables to predict election results (Silver, 2012). Nate Silver warns of studies that use the large availability of economic data to manipulate their models to fit the data by using illogical combinations of variables. He also claims that some fundamentals based models implement variables of future economic statistics, or values that compromise and undermine the point of the study. These factors can invalidate studies of models that claim to be able to predict 90% of voting behaviour. create issues of reverse causality by. He additionally states that many times fundamental models overstate the power of the economy, and make sweeping claims about the economy’s effect on election results.

4. Methods

The following section will describe the methods employed throughout the research process and will dive into the reasoning behind the methodology used. Specifically, it will describe the process of weeding out variables and selecting the most important independent variables — with many representations of the same variables, the process involved a lot of repetition and ultimately only a couple representations of each variable were significant and will be described in detail.

Originally my intention was to perform a series of ridge and lasso regressions on the data. With the wealth of related independent variables, these regressions would have helped pare down the number of variables while still describing their effect on the popular vote, and they would reduce the effect of the independent variables correlation to each other, by shrinking the coefficients that are close to zero. They are also both methods that allow for easily interpretable results. However, due to the limited data set and the low number of elections available, this was not possible. Instead, this paper utilizes Ordinary Least Squares (OLS), which still very clearly shows the effect of the variables on the popular vote, but does not account for multicollinearity in the same way, and does not automatically help weed out less significant variables.

4.1. Hypotheses and Predictions

Initially, I believe the fundamental variables would have a significant and differentiable impact on voting behaviours and patterns. Real Disposable Income (RDI), inflation, and unemployment are all factors that affect people's daily lives and are influenced by the acting government, even if not directly by the president. Specifically, following the ‘Bread and Peace’ Model constructed by Douglas A. Hibbs, Jr (2000), I believed that RDI would have the most impact on the popular vote. Otherwise, I expected the third-quarter political variables — for both polling and incumbent approval ratings — to be powerful indicators of the election results, because they are direct measures of the public’s feelings at times close to the election.

4.2. Models

The two different types of models used were simple bivariate regression and multiple linear regression. For both models separate regressions were run for the two categories of variables, in order to best be able to interpret their predicting abilities. Furthermore, the bivariate model was used to weed out insignificant variables, or variables that were represented in multiple ways in the data set. This allowed for the selection of the most predictive variables in the second stage of regressions, the Multiple Linear Regression (MLR) stage; where two regressions with all the most important variables in both categories (political and fundamental) were run.

4.2.1. Bivariate Regression

As mentioned above, the bivariate regressions were used as weed-out tools where every variable in the data set was run individually. The models followed the following structure:

$$y_i=\beta_0+\beta_1 x_i+\varepsilon_i$$

where *y* is representative of the dependent variable and *x* is representative of the independent variables. In this study every the dependent variable was run in two bivariate regressions; one against the percentage of the popular vote obtained by the incumbent party's candidate (i.e. "VOTE") and one against the two-party vote (i.e. "2VOTE") — which is the percentage of the vote obtained by the incumbent party's candidate out of votes cast only for the Republican or Democartic Parties.

Once again, at this stage for the fundamental variables, the independent variables were: all-cause mortality per 10,000 people (i.e. "DEATH"), the change in all-cause-mortality from the previous election cycle (i.e. "%DEATH"), annual inflation (i.e. "INFL"), annual Real Disposable Income (i.e. "RDI"), the year to year change in RDI expressed as a percentage (i.e. "%RDI"), the third-quarter RDI for that election year (i.e. "3RDI"), the year to year change in third-quarter RDI (i.e. "%3RDI"), the annual unemployment rate (i.e. "AUNEM"), the third-quarter unemployment rate (i.e. "3UNEM"), a weighted RDI statistic (i.e. "WRDI_1"), the year to year percent change in WRDI_1 (i.e. "%WRDI_1"), a second weighted RDI statistic (i.e. "WRDI_2"), the year to year percent change in WRDI_2 (i.e. "%WRDI_2"), a third weighted RDI statistic (i.e. "WRDI_3"), the year to year percent change in WRDI_3 (i.e. "%WRDI_3"), and finally a fourth weighted RDI statistic (i.e. "WRDI_4"), and the year to year percent change in WRDI_4 (i.e. "%WRDI_4").

For the fundamental variables, the independent variables were: the average annual approval rating of the sitting president (i.e. "APPROVAL"), the third-quarter approval rating of the sitting president (i.e. "3APPROVAL), the average polling numbers for the incumbent candidate in the year leading up to the election (i.e. "AVGPOLL"), and the average polling numbers for the incumbent candidate in October and November (i.e. "OCTNOVPOLL").

These bivariate regressions led to various results which are explained in detail in the next section, however, for the next step they led to the important conclusion that for the purposes of this research the two-party vote led to more clear and significant results for the majority of the variables. This is due to the fact that for the majority of the variable categories the two-party vote led to larger coefficients and more statistical significance; and later, when testing the MLRs, all of the regressions with the two-party vote had larger adjusted r-squared values. Therefore, for the next step, MLR, the two-party vote was taken as the default dependent variable.

4.2.2. Multiple Linear Regression (MLR)

To test the hypothesis and how these variables work in tandem, I selected the most statistically significant form of each variable and ran various multiple linear regressions (MLRs) for both categories of variables. The models took on the following structure:

$$y_i=\beta_0+\beta_1x_2+\beta_2x_2....+\beta_nx_n$$

where, once again, y is the dependent variable, or the two-party vote, and the x_i's are the independent variables, or a series of fundamental or political variables. For this second stage every possible combination of the most significant variables for each category was run in separate regressions to obtain the most accurate predicting model for both categories of variables. So, for the fundamental regressions one RDI variable was selected, one death rate variable was selected, and so on. Then every possible combination of those four categories of variables was run. Lastly, the same process was repeated with the political variables.

5. Results and Analysis

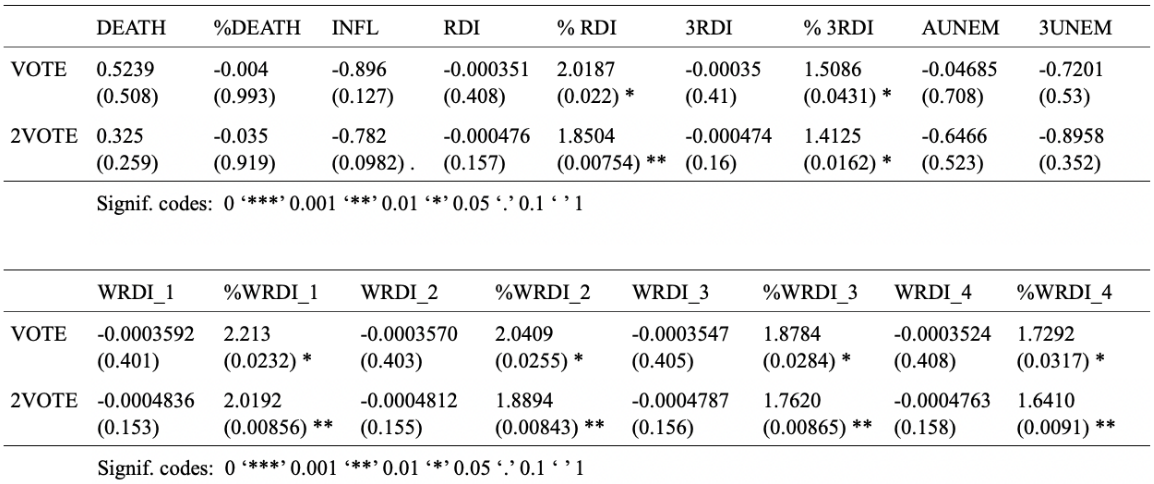
Next we will discuss the findings from both stages of regression, including the weed out of different variables and the final results of the separate MLR regressions. In total there were 54 regressions, consisting of 42 bivariate regressions and 12 multiple linear regressions; or 45 regressions utilizing purely fundamental variables and 9 regressions using purely political variables.

5.1. Fundamental Variables

5.1.1. First Stage - Bivariate Regression

Below are the results of the bivariate regressions for the fundamental variables for both the popular vote, and the two-party vote:

Table 1: First Stage Bivariate Regressions of Fundamental Variables



	DEATH	%DEATH	INFL	RDI	% RDI	3RDI	% 3RDI	AUNEM	3UNEM
VOTE	0.5239 (0.508)	-0.004 (0.993)	-0.896 (0.127)	-0.000351 (0.408)	2.0187 (0.022) *	-0.00035 (0.41)	1.5086 (0.0431) *	-0.04685 (0.708)	-0.7201 (0.53)
2VOTE	0.325 (0.259)	-0.035 (0.919)	-0.782 (0.0982) .	-0.000476 (0.157)	1.8504 (0.00754) **	-0.000474 (0.16)	1.4125 (0.0162) *	-0.6466 (0.523)	-0.8958 (0.352)

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

	WRDI_1	%WRDI_1	WRDI_2	%WRDI_2	WRDI_3	%WRDI_3	WRDI_4	%WRDI_4
VOTE	-0.0003592 (0.401)	2.213 (0.0232) *	-0.0003570 (0.403)	2.0409 (0.0255) *	-0.0003547 (0.405)	1.8784 (0.0284) *	-0.0003524 (0.408)	1.7292 (0.0317) *
2VOTE	-0.0004836 (0.153)	2.0192 (0.00856) **	-0.0004812 (0.155)	1.8894 (0.00843) **	-0.0004787 (0.156)	1.7620 (0.00865) **	-0.0004763 (0.158)	1.6410 (0.0091) **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

After the first stage of bivariate regressions for the fundamental variables, several discoveries led to the selection of the key variables for the second round of regressions. The first key observation was that for all-cause mortality the statistic measuring the change from the previous election cycle in deaths per 10,000 was more accurate, but less statistically significant. The coefficient for the variable representing the change in all-cause mortality (i.e. "%DEATH") was $ß_1$= -0.035 and the p-value was 0.919 (see Table 1). Whereas, the coefficient for deaths per 10,000 people (i.e. "DEATH") was $ß_1$= 0.325 and the p-value was 0.259. Despite being less statistically significant, "%DEATH" was selected for the MLR regressions given that intuitively speaking the coefficient for the all-cause mortality should be negative — when there is a notable increase in the number of people dying in a year that should reflect poorly in the current leadership and therefore cause the incumbent party to lose votes. The second thing worth noting was that the only measure for inflation (i.e. "INFL") had a negative coefficient (as expected) and had a relatively high degree of statistical significance (see Table 1). Furthermore, the variable for inflation has a relatively large coefficient given that these regressions imply that a one percentage point increase in inflation would lead to 0.78 percentage point fall in the candidates two-party vote. Therefore the measure for inflation was used in the next round of multiple linear regressions.

For the statistics regarding unemployment all of the coefficients are negative, as expected. And for both the popular vote and the two-party vote, the third-quarter unemployment rate (i.e. "3UNEM") had larger coefficients and lower p-values. Therefore, the third-quarter unemployment rate was selected for the MLR stage, given that the coefficient was both larger and more statistically significant than for the annual unemployment. The coefficient for the variable was $ß_1$= -0.8958 and the p-value was 0.352; implying that a percentage increase in the third-quarter unemployment would lead to a 0.90 drop in the incumbent party's candidate two-party vote.

Another discovery was that for the measures of RDI, a more rounded-out statistic that encompasses the RDI throughout the year of the election has more effect on the popular vote, rather than a RDI measure that is weighted heavily on the third-quarter (right before the election). This is apparent because WRDI_1 has the largest coefficient of any of the weighted RDIs, the annual measure for RDI, and the third-quarter RDI. It is also noteworthy that all of the two-party vote and RDI bivariate regressions have relatively low p-values. Additionally, due to the scale of the RDI variable the year to year change was a more impactful and appropriate predictor of the two-party vote, rather than the RDI reported in billions of chained 2012 dollars at a seasonally adjusted annual rate. These two key observations led to the selection of the variable "%WRDI_1", or year to year change in annual RDI, for the MLR stage of the study. The coefficient for this variable was β_1= 2.0192 and the p-value is 0.00854 (see Table 1). Meaning that the variable is highly statistically significant and a percentage point increase from the RDI the year before the election will roughly translate into a 2% increase in the incumbent party's candidate two-party popular vote.

From the first stage of bivariate regressions, four out of the 17 fundamental variables were selected for the multiple linear regression stage of the study, in each of the four categories (death rate statistics, inflation, unemployment, and RDI). The four variables selected were: the all-cause mortality statistic measuring the change from the previous election cycle in deaths per 10,000 people, yearly inflation presented as a percentage, the third-quarter unemployment rate, and the yearly percent change in "WRDI_1" (which is described in detail in equation 1).

5.1.2. Second Stage - MLR

For the second stage of regressions all possible combinations of the four selected variables were performed and all their results were reported.

Table 2: Multiple Linear Regressions for Fundamental Variables

	%DEATH	INFL	%WRDI_1	3UNEM	Adjusted R-squared
Regression 1	-0.04458 (0.891)	-0.78304 (0.111) .	NA	NA	0.05883
Regression 2	-0.4025 (0.16760)	NA	2.4029 (0.00396) **	NA	0.4054
Regression 3	0.04795 (0.894)	NA	NA	-0.89232 (0.368)	-0.08059
Regression 4	NA	-0.4401 (0.2901)	1.7580 (0.0256) *	NA	0.3669
Regression 5	NA	-0.7191 (0.179)	NA	-0.2750 (0.778)	0.0634
Regression 6	NA	NA	1.9665 (0.0112) *	-0.6830 (0.3594)	0.3527
Regression 7	-0.3518 (0.25973)	NA	2.3228 (0.00736) **	-0.4125 (0.58951)	0.3719
Regression 8	-0.3681 (0.2112)	-0.3732 (0.3592)	2.1486 (0.0132) *.	NA	0.4012
Regression 9	-0.01982 (0.955)	-0.72350 (0.199)	NA	-0.25806 (0.808)	-0.01437
Regression 10	NA	-0.3377 (0.4721)	1.7860 (0.0289) *	-0.4245 (0.6098)	0.3295
Regression 11	-0.3525 (0.2690)	-0.3393 (0.4650)	2.1421 (0.0181) *	-0.1523 (0.8581)	0.3488

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

As seen in Table 1, all combinations of the four key fundamental variables were run. Ultimately, the table shows a large variance between the success of the different models in predicting the two-party popular vote. All regressions including the variable “%WRDI_1” were able to account for approximately 30% of the variance in the data, as demonstrated by the adjusted r-squared. Furthermore, Regression 2 (which closely follows the ‘Bread and Peace Model’ that was previously introduced) had an r-squared of 40.54%; implying that a simple two variable model is able to explain around 41% of the variance of election outcomes’ and was the most explicative model (see Table 2). Interestingly enough, this is almost exactly the percentage (40%) Nate Silver had previously reported when explaining why fundamentals based election predictions were not as indicative as some might believe.

As expected "%DEATH" had a negative coefficient of $ß_1$= -0.4025, and a relatively low p-value; indicating a certain degree of statistical significance (see Table 2). These results imply that for a one percentage increase in all-cause mortality during the year before an election, the candidate for the incumbent party will lose one 0.40% of the two-party popular vote. Now onto the year to year change in "WRDI_1", as expected the coefficient for this variable is positive and as seen throughout the paper it is one of the variables with the lowest p-value and most statistical significance. The coefficient for "%WRDI_1" was $ß_2$= 2.4029 and the p-value was 0.00396 (see Table 2); meaning that a relative percent increase in the RDI from the year before could lead to two times as many percentage points more for the incumbent party's candidate. The increase in statistical significance relative to the measure of all-cause mortality is reasonable, because unlike all-cause mortality (which is measured per 10,000 people), RDI is extremely tangible to all citizens. It directly affects how much people have at their disposal at any given moment in time and indicates how wealthy people feel. This could directly impact someone's perception of the success and capabilities of the current government and president. It is important to note, however, that the same sentiment about tangibility could be applied to unemployment and it is therefore not the only possible explanation for why RDI seems to have such a significant impact on voter behaviour. Overall, the model is not completely accurate and fundamental variables cannot completely predict election outcomes. However, if they explain approximately 40% of behaviour these two variables can still be significant and account for a large percentage of the population's decision making.

Another explicative model that resulted from the four key fundamental variables that were previously selected was Regression 8, which included the year to year change in all-cause mortality, inflation, and the year to year change in "WRDI_1”. This model had a r-squared value of 0.4012 (as seen in Table 2), and therefore, similarly to Regression 2, explains about 40% of voter behaviour. As expected both “%DEATH” and “INFL” had negative coefficients of $ß_1$= -0.3681 and $ß_2$= -0.3732 (see Table 2). What these results imply is that for a one percentage increase in all-cause mortality or inflation during the year before an election, the candidate for the incumbent party will lose one 0.37% and 0.37% of the two-party popular vote for each respective variable. The coefficient for “%WRDI_1” was $ß_3$= 2.1486 and a p-value of 0.0132. Although slightly less accurate at predicting the two-party vote, it is still interesting to note the relative effects of each variable and their predictive capabilities. After all, both variables have extremely similar adjusted r-squared values.

5.1.3. Predictions

The predictions created by Regression 2 (see Table 2) are listed below, along with a row indicating whether the incumbent candidate did win the election. If the predicted value was above 50% and the candidate won the election then, for the purposes of this study, the prediction was counted as correct. See predicted values below:

Table 3: Fundamental Model Predictions

	1960	1964	1968	1972	1976	1980	1984	1988
Actual	49.95	61.93	49.59	61.81	48.93	44.71	59.15	53.94
Win	0	1	0	1	0	0	1	1
Predicted	50.02	57.96	53.13	53.64	52.58	45.04	59.04	52.95

	1992	1996	2000	2004	2008	2012	2016	2020
Actual	46.52	54.73	48.4	51.21	45.7	51.06	48.2	46.9
Win	0	1	0	1	0	1	0	0
Predicted								50.13

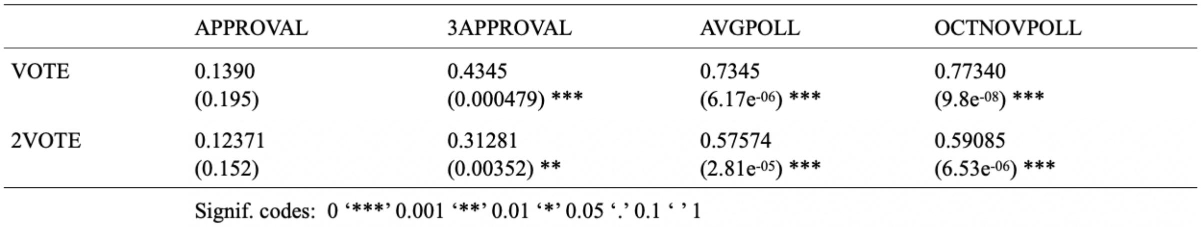
Overall, the fundamental models are pretty accurate when it comes to predicting the winner of the majority of the two-party vote. The most predictive model predicted the correct candidate for 9 out of the past 16 elections, over the last 60 years (see Table 3) — that is a 56.25% success rate. Interestingly enough, the model is more accurate when the incumbent candidate won the election, which could imply an implicit bias towards incumbent parties, or sitting presidents during election years.

5.2. Political Variables

5.2.1. First Stage - Bivariate Regression

Below are the results of the bivariate regressions for the political variables for both the popular vote, and the two-party vote:

Table 4: First Stage Bivariate Regressions of Political Variables



	APPROVAL	3APPROVAL	AVGPOLL	OCTNOVPOLL
VOTE	0.1390 (0.195)	0.4345 (0.000479) ***	0.7345 ($6.17e^{-06}$) ***	0.77340 ($9.8e^{-08}$) ***
2VOTE	0.12371 (0.152)	0.31281 (0.00352) **	0.57574 ($2.81e^{-05}$) ***	0.59085 ($6.53e^{-06}$) ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

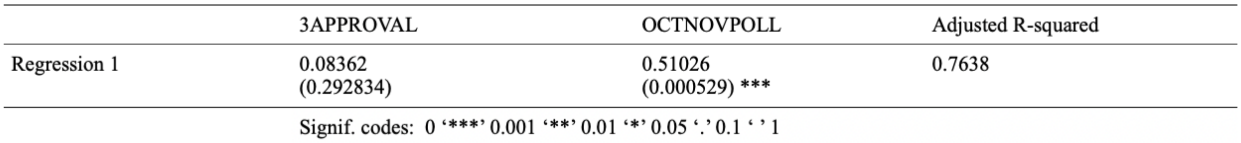
When considering the political variables the weed out process was less straightforward given that they were all pretty statistically significant (see Table 4) and that, like elections, polls and approval ratings are a direct measure of the country’s feelings towards the current government. If one approves of the current president, it is likely that you would vote for them or their party in the next election. This is not always the case, but it is increasingly common given the polarization of the current political climate.

In order to ensure continuity throughout the experiment the two-party vote (i.e. “2VOTE”) was chosen for the MLR stage of political variables as well as fundamental variables, despite the fact that the popular vote lead to larger coefficients and, for most variables, higher statistical significance. This makes sense given that when it comes to the public’s opinion third-party candidate can be considered, and would be factored in. Furthermore, the third-quarter approval ratings (i.e. “3APPROVAL”) and the average of October and November polling numbers for the incumbent party’s candidate (i.e. “OCTNOVPOLLS”) were chosen. This was due to their higher statistical significance and the fact that they had larger coefficients, which were closer to one (see Table 4). Since, these variables are percentages that measure approval a coefficient closer to one would indicate a more impactful and direct relationship to the two-party popular vote. The coefficients for these two variables were $ß_1$= 0.31281 and $ß_1$= 0.59085, respectively, and their p-values were both significantly below 0.05. Although these coefficients are both large enough to have an impact on the two-party vote, it is odd that direct measures of the public’s opinion are not closer to one. And furthermore, they are not much larger than the fundamental variables, which are indirect measures of the success of the sitting president.

5.2.2. Second Stage - MLR

For the second stage of regressions a multiple linear regression was run including both the political variables — the October and November polling results and the third-quarter approval rating of the sitting president. The results can be seen below:

Table 5: Multiple Linear Regressions for Political Variables



	3APPROVAL	OCTNOVPOLL	Adjusted R-squared
Regression 1	0.08362 (0.292834)	0.51026 (0.000529) ***	0.7638
	Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1		

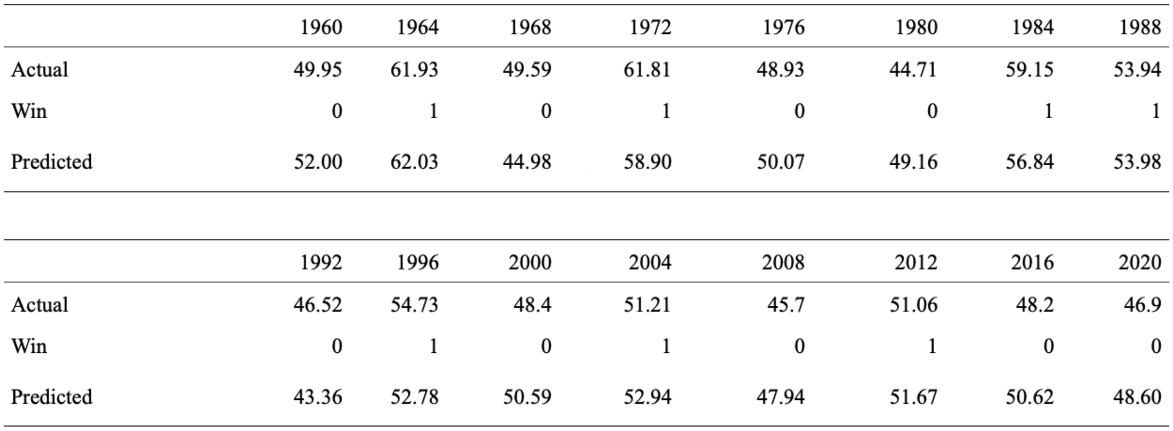
Given that there were only two key variables for the category of political variables there was only one regression to run for this category. In this model, where only the two political variables are included: the polling variable has a coefficient of *ß*$_2$= 0.51026 and a p-value of 0.000529, and the approval rating variable has a coefficient of *ß*$_1$= 0.08362 and a p-value of 0.29. This coefficient can have two distinct and opposing interpretations. The first is that it is an extremely significant effect given that a rise of one percentage point in the polls would lead to a 0.5% increase in the candidates' two-party popular vote. The second interpretation is that since polls are supposed to represent a predictor of who people will vote for, this coefficient is extremely low and should be a lot closer to one. The model has an adjusted r-squared of approximately 0.7638 meaning that 76% of the variance can be explained by the model. Therefore, despite having smaller coefficients than the fundamentals based model, it is overall more accurate.

Lastly, it is important to note that the coefficient and p-value for the polling numbers in Regression 1 have actually gotten worse with the MLR. The variable was more accurate and impactful in bivariate regressions, despite the MLR model being more accurate and having a higher adjusted r-squared (see Table 5). This is probably due to issues of multicollinearity between polling numbers and approval ratings — as stated previously, people who approve of the current president are more likely to vote for the incumbent party than not. The multicollinearity could also explain why the coefficient and p-value for the approval rating variable have also gotten worse in the MLR regression; they are less explanatory and less significant.

5.2.3. Predictions

Overall, the political variables model is pretty accurate when it comes to predicting the winner of the majority of the two-party vote. As before, if the predicted value was above 50% and the candidate won the election then, for the purposes of this study, the prediction was counted as correct. The predicted two-party vote by the political model can be seen below:

Table 6: Political Model Predictions



	1960	1964	1968	1972	1976	1980	1984	1988	1992	1996	2000	2004	2008	2012	2016	2020
Actual	49.95	61.93	49.59	61.81	48.93	44.71	59.15	53.94	46.52	54.73	48.4	51.21	45.7	51.06	48.2	46.9
Win	0	1	0	1	0	0	1	1	0	1	0	1	0	1	0	0
Predicted	52.00	62.03	44.98	58.90	50.07	49.16	56.84	53.98	43.36	52.78	50.59	52.94	47.94	51.67	50.62	48.60

The model predicted the correct candidate for 12 out of the past 16 elections, over the last 60 years (see Table 6) — that is an accuracy rate of 75%. Although when framed as a percentage the difference between the political variables model and the fundamentals based model seems significant, in reality there is only a difference of 3 elections in 60 years. Additionally, the political variables are more robust and have more data points than the fundamentals based models; and the data is taken right before the election. It is impressive that the gap between the two predictions is not larger. This relative closeness could be due to factors such as how a question is phrased, what exact populations are sampled, human error, or even people who change their mind. These factors can affect polls in a way that the fundamentals based models are not susceptible to. Overall, the political model is a more accurate predictor of the two-party vote than the fundamentals based model.

6. Conclusion

This study shows that there is a difference between fundamentals based models and political variable models. More specifically it demonstrates that approximately 40% of voting behaviour can be explained through two or three key fundamental variables, whereas approximately 75% of voting behaviour can be explained through two key political variables. And that each set of variables comes with its own flaws and drawbacks. Although, overall, the political variables were more accurate in predicting the final two-party popular vote; they also had more recent and robust data. Furthermore, the study demonstrates that the year to year change in all-cause mortality, inflation, and the year to year change in a weighted measure of Real Disposable Income (RDI) have a significant effect on voting behaviour and can affect the outcomes of elections. Specifically, that the percent change in RDI is extremely significant; with a coefficient of β_3= 2.1486 a change of one percentage point in the RDI in the year of an election can lead to approximately two percentage points gained in the two-party popular vote.

As discussed previously there are some limitations to the data set and the analysis. First and foremost, although there are several observations that can be made from the popular vote and the two-party popular vote, United States presidential elections are not decided by the popular vote, they are decided by the electoral college. There have been multiple presidents who won the election, but lost the popular vote — most notably George W. Bush (2000) and Donald Trump (2016). Second of all, the data set for both fundamental and political variables is limited and does not span as far back as would be necessary to run different types of regressions and do more in depth analysis of the variables and their predictive capabilities. Furthermore, it would be interesting to investigate global voting patterns and explore whether the trends and affecting factors in United States elections also affect behaviour worldwide. It could be that, with a wider data set, including more countries, a study could be able to predict the popular vote in any election and it could reveal what citizens everywhere care about most when deciding who they would vote for. Finally, although the purpose of this study was to compare the two categories of variables, there is something to be said for combining them and seeing exactly how to use their predictive capabilities in conjunction to get the most accurate and predictive results possible.

7. Bibliography

Abramowitz, A. I. (2016). Will Time for Change Mean Time for Trump? PS: Political Science & Politics, 49(4), 659–660. https://doi.org/10.1017/S1049096516001268

Bartels, L. M., & Zaller, J. (2001). Presidential Vote Models: A Recount. PS: Political Science & Politics, 34(1), 9–20. https://doi.org/10.1017/S1049096501000026

Cuzán, A. G. (2012). Forecasting the 2012 Presidential Election with the Fiscal Model. PS: Political Science and Politics, 45(4), 648–650. https://www.jstor.org/stable/41691404

Fair, R. (n.d.). Ranking Assumption for 2020 Presidential election and 2020 Senate elections. Retrieved November 2, 2020, from https://fairmodel.econ.yale.edu/vote2020/inrank.htm

Gelman, A. (2020, August 31). Problem of the between-state correlations in the FiveThirtyEight election forecast « Statistical Modeling, Causal Inference, and Social Science. https://statmodeling.stat.columbia.edu/2020/08/31/problem-of-the-between-state-correlations-in-the-fivethirtyeight-election-forecast/

Gelman, A., Hullman, J., Wlezien, C., & Morris, G. E. (2020). Information, incentives, and goals in election forecasts. Judgment & Decision Making, 15(5), 863–880. http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=146375735&site=ehost-live

Gelman, A., & King, G. (1993). Why Are American Presidential Election Campaign Polls so Variable When Votes Are so Predictable? British Journal of Political Science, 23(4), 409–451. https://www.jstor.org/stable/194212

Hibbs, D. A. (2000). Bread and Peace Voting in U.S. Presidential Elections. Public Choice, 104(1), 149–180. https://doi.org/10.1023/A:1005292312412

Hibbs, D. A. (2008). Implications of the "Bread and Peace" Model for the 2008 US Presidential Election. Public Choice, 137(1/2), 1–10. https://www.jstor.org/stable/40270847

Historical polling for United States presidential elections. (2021). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Historical_polling_for_United_States_presidential_elections&oldid=1018185798

Lauderdale, B. E., & Linzer, D. (2015). Under-performing, over-performing, or just performing? The limitations of fundamentals-based presidential election forecasting. International Journal of Forecasting, 31(3), 965–979. https://doi.org/10.1016/j.ijforecast.2015.03.002

Mehta, A. B., Ritchie King and Dhrumil. (2018, June 28). National President: general election Polls. FiveThirtyEight. https://projects.fivethirtyeight.com/polls/

RealClearPolitics - Election 2020 - General Election: Trump vs. Biden. (n.d.). Retrieved May 13, 2021, from https://www.realclearpolitics.com/epolls/2020/president/us/general_election_trump_vs_biden-6247.html#polls

Silver, N. (2011, August 31). Despite Keys, Obama Is No Lock. FiveThirtyEight. https://fivethirtyeight.blogs.nytimes.com/2011/08/31/despite-keys-obama-is-no-lock/

Silver, N. (2012, March 26). Models Based on ‘Fundamentals’ Have Failed at Predicting Presidential Elections. FiveThirtyEight. https://fivethirtyeight.com/features/models-based-on-fundamentals-have-failed-at-predicting-presidential-elections/

Wezerek, J. B., Gus. (2019, April 4). How Good Are FiveThirtyEight Forecasts? FiveThirtyEight. https://projects.fivethirtyeight.com/checking-our-work/