Using Statistics to Predict the Outcome of the 2016 Presidential Election

Nicole Barbour, Elisabeth Mersino, Carlos Rivera, and Gorka Bravo Martinez

California State University Monterey Bay, Seaside, California

Abstract

Predicting the outcome of the presidential elections in the United States has become more accurate over the years, with past studies using factors like the state of the economy and how ready citizens are for a change in the White House to formulate a statistical model. In our study, we sought to predict the outcome of the 2016 presidential election using survey data from individuals from each state from the Understanding America study. Our overall goal was to 1) predict the winning presidential candidate in each state and District of Columbia and 2) predict the final percentage of the national popular vote for each candidate. We fit a logistic regression model, with the probability of Clinton being president as a 1 and probability of Trump as a 0 as a response variable and with race, income level, age, education level, sex, and house size variables from each individual in the survey as predictors. By fitting our model and running a prediction function with R programming language, we were able to predict the outcome of the 2016 election with Trump as the winner (57.6%) and Clinton the loser (42.4%) for the national popular vote, as well as predict the winner in each state. Results of this study will serve to inspire future studies that seek to predict the outcome of a presidential election with individual survey data.

Introduction

Past studies have aimed to predict the outcome of a presidential election with statistical robustness before the actual election, but predictions vary in their dependent variables. Each election year presents a new mixture of these variables with factors such as the state of the economy, whether or not the country is at war, and how the long the current president has been in office having an impact on individual voting decisions (Miller et al 1986; Albanowitz 1988).

Assessing multiple factors such as these is essential to gaining the most accurate predictions possible. Predicting the candidate election for each state based on partisan affiliation and preference alone has been demonstrated to not be the most effective, with past elections demonstrating policy preferences having more of an influence in an individual's vote than party attachments (Page & Jones, 1979). Instead, individual voting preference has been shown to be dictated more by the schema of the candidate, or, mental representation of how the candidate is perceived, than by which party or issues the individual sides with, with candidates being thought of in a categorized, bin-like manner (Miller et al., 1986). These factors influencing an individual's voting choice can be incorporated into a model predicting the outcome of an election.

Historically, the outcomes of presidential elections were predicted using a "time for change" model. Lewis-Beck and Rice (1984), predicted presidential election results on a national level statistically before the election using post World War II data in the form of an aggregated time series. They formed a multivariate forecasting model with economic performance and presidential popularity as predictors of the popular vote for president; economic conditions were the change in real Gross National Product (GNP) per capita and popularity the president’s approval rating in May.

Recent studies have also used the popularity and open-access of social media to predict the popularity of specific parties in an election. In the case of Tumasjan et al. (2010), the outcome of the national parliament federal election in Germany for 2009 was predicted using Twitter posts using a text analysis software to show how the number of Tweets with the name of a party reflects the outcome of that party in the election.

Important in the outcome of the election is the electoral college. The electoral college was first introduced at the 1787 Constitutional Convention to discuss a balance between the federal and state government, with various smaller states have higher leverage in the election process, preventing them from being lost in the voting weight of larger, more populated states, and preventing the election from being manipulated politically (Neale, 2012). The electoral college uses chosen candidates to cast votes in order to represent American voters. These individuals are nominated in various ways based on each of the states, with any U.S. citizen being eligible as an electoral candidate. Since there are various ways to nominate candidates, it is left to American voters to vote for the candidates themselves before the primary election. Once an electoral college candidate is selected their vote pledges on the presidential candidates are given based on the American voters (Hoffman, 1996). The electoral college works primarily with the major two parties; the smaller third parties are not adequately represented during the presidential election. Despite this, the electoral college is a small part of the election process.

In order for a model to have accuracy in forecasting the outcome of an election, it must show a strong relationship between the popular vote share and the electoral vote share, as the majority electoral vote in most states determines whether a candidate is set for presidency (Lewis-Beck & Rice, 1984). The upcoming presidential debate of 2016 has two main candidates, Hillary Clinton representing the democratic party and Donald Trump representing the republican party. Predicting the outcome of this election, in each state and nationally, will involve formulating a model that incorporates variables that influence the individual vote. In order to predict the outcome of the 2016 election, a generalized linear model will be formulated with a response variable (probability of Trump or Clinton) and six explanatory variables using data provided by the Understanding America study: race, age, education, income, sex, and house size.

Methods and Materials

Data Collection

Data was obtained from the 2016 University of Southern California’s Dornsife Center for Economic and Social Research and the LA Times Presidential Election Poll as part of the Understanding America study. This is composed of poll data collected from around 3000 respondents, all members of the Understanding America Study election panel, based on three questions: 1) whether they will vote 2) which candidate they will vote for and 3) which candidate they think will win. These respondents were selected randomly from United States households and data was weighted to match race and gender attributes from the U.S. Census Current Population Survey and the results of the 2012 presidential election.

Variable Creation

The response variable of interest in our model was the probability of either Trump or Clinton winning. This probability was found by asking respondents daily on how they felt about the election and which way they would vote. Within R, each probability was transformed into either a 0 (Trump) or Clinton (1), based on the probability. For the Clinton probability variable, values less than 0.50 were classified as 0 and values greater than 0.50 were classified as 1. Our explanatory variables were found based on the individuals in the study, with race, age, sex, state of residency, and house size being recorded for each respondent. Race, education level, income level, state of residency, and sex were categorical variables; age, education level, income level, and house size were continuous variables.

Analytic Methods

Our model used logistic regression with Clinton and Trump as a binary response variable (Clinton 1, Trump 0). All data was manipulated in R, using the packages “haven”, “gdata”, ”Hmisc”, “car”, “ggplot2”, “boot”, and “caret”. Separate models were run for each state in the United States; a global model was run to predict the final percentage of the national popular vote for each major candidate. Variable selection was performed using significance testing and AIC values of the global model. Predictions for each model were made using the “predict” function in R. Predicted percentages greater than 50 percent were categorized as in favor of Clinton, whereas those less than 50 percent were categorized as in favor of Trump. Percentages exactly equal to 50 percent were split between Trump and Clinton. Skewed state data (aka. only voting for Trump, all females, etc) were not included in the analysis. Model fit was tested with an ROC curve and AUC values.

Results

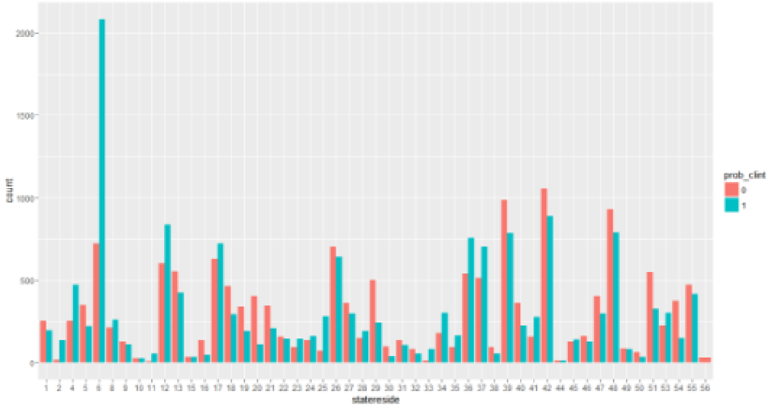


Figure 1. Bar graph showing count data for each state (see Figure 8 for FIPS state codes), with 0 representing Trump and 1 representing Clinton.

As seen in Figure 1, the counts of individuals voting for either Clinton or Trump varied by state (see Figure 8 for state FIPS codes). State 6, California, has a distinct preference for Clinton, whereas state 56, Wyoming, only has data present for Trump. States with higher count data for Clinton are as follows: state 4 (Arizona), state 6 (California), state 8 (Colorado), state 11 (DC), state 12 (Florida), state 17 (Illinois), state 23 (Maine), state 24 (Maryland), state 25 (Massachusetts), state 33 (New Hampshire), state 34 (New Jersey), state 35 (New Mexico), state 36 (New York), state 37 (North Carolina), state 41 (Oregon), state 45 (South Carolina), and state 53. (Washington). State 10 (Delaware), 15 (Hawaii), and 44 (Rhode Island) had equal counts for Trump and Clinton. The rest of the states had count preferences for Trump.

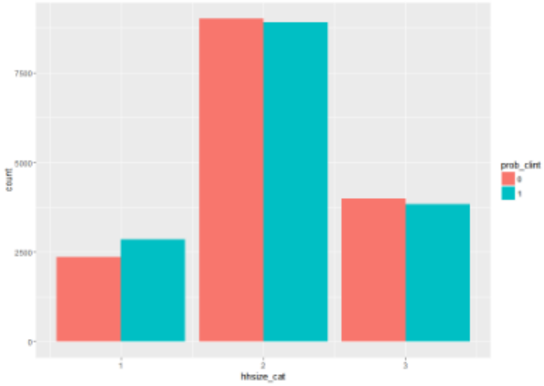


Figure 2. Bar graph showing count data for Trump (0) and Clinton (1), with smaller households on the x-axis being category 1, medium category 2, and larger households category 3.

In Figure 2, a smaller household size (category 1) has a slightly higher preference for Clinton, whereas larger household sizes (categories 2 and 3) have a slightly higher preference for Trump.

In Figure 3, a lower and higher income (category 1, <$35,000 ;category 2, >=$75,000) is associated with a preference for Clinton, whereas a middle income (category 2, $35-75,000) with a preference for Trump.

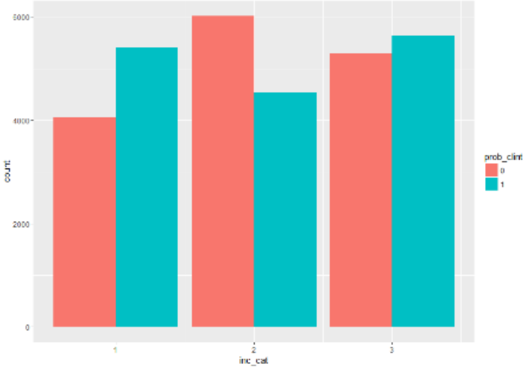


Figure 3. Bar graph showing count data for Trump (0) and Clinton (1), with category 1 on the x-axis being lower income levels (<$35,000), category 2 being medium income levels ($35-75,000), and category 3 being higher income levels (>=$75,000).

In Figure 4, a trend is seen in the distribution of data for Clinton and Trump. For Trump, a normal distribution is seen, with median values around age 40. For Clinton, there appears to be a bi-modal distribution, with a peak around age 20 and age 40.

In Figure 5, non-Hispanic whites (category 1) not only have a distinct preference for Trump, but also have much higher count data than all other race categories. All other categories, which include Hispanic, black, and non-Hispanic races, demonstrate a preference for Clinton.

Figure 6 shows the difference in preference based on sex, with males (category 1) having higher count data for Trump and women (category 0) having higher count data for Clinton.

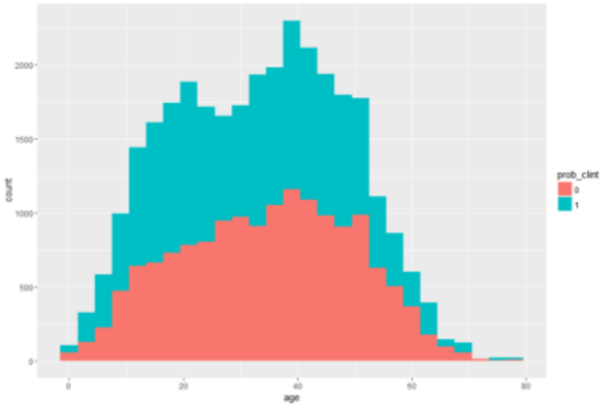


Figure 4. Histogram showing count data for each category, with age on the x-axis and counts being split between Trump (0) and Clinton (1).

Figure 7 illustrates highest education level achieved for those voting for Trump or Clinton. The graph shows higher count data for Clinton in higher education categories (Bachelor’s degree up to Doctoral degree). Trump, however, has higher count data for lower education categories (less than 1st grade up to Associates degree in college).

All variables were found significant (p<0.05, AIC 35429) in the global model (race, gender, age, education, household size, income size, and state residency), with Clinton as a response variable (see global model output in Appendix). However, the ROC curve had a relatively poor relationship between specificity and sensitivity, with an AUC value of 0.694 (see Appendix for figure), illustrating that our model did a poor job of distinguishing between success (Clinton winning) and failure (Trump winning).

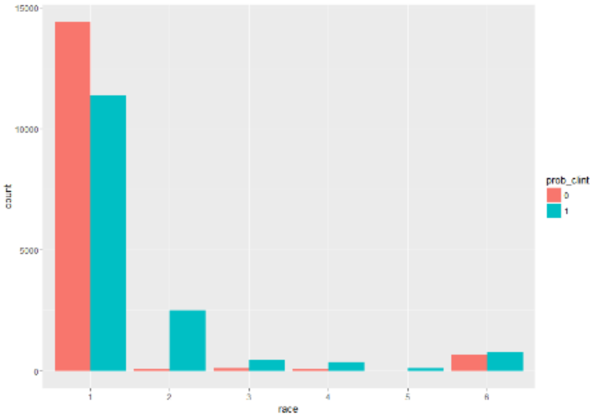


Figure 5. Bar graph showing count data for race categories on the x-axis, with category 1 being non-Hispanic white and all other categories being non-white and counts being split between Clinton (1) and Trump (0).

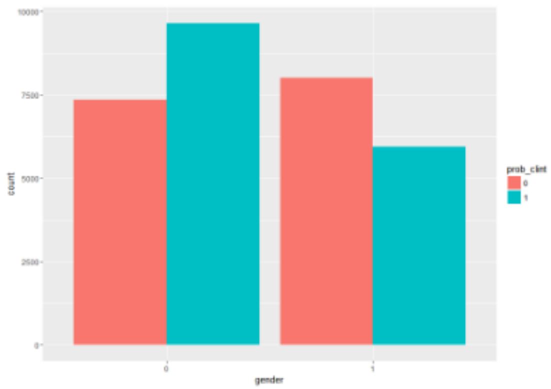


Figure 6. Bar graph showing count data for men (1) and females (0) on the x-axis, being split for Trump (0) and Clinton (1).

Discussion

For each state model, as well as the global model, final percentages for each candidate are given (see Table 1). The results of the global model gives the national popular vote winner as Trump (57.6% vs. 42.4%). For the states, Clinton won in Alaska, Arizona, California, Colorado, District of Columbia, Florida, Maine, Massachusetts, Mississippi, New Hampshire, New Jersey, New Mexico, New York, North Carolina, Oregon, Utah, and Washington. The other 33 states were won by Trump.

For our explanatory variables, we found that the odds of Clinton winning were multiplied by .993 for every 1 unit increase in age when all other variables were held fixed. The odds of Clinton winning with a male voter are 0.589 times the odds of Clinton winning with a female voter. For race, the odds of Clinton winning with for a non-white voter are around 61 times the odds of Clinton winning with a white voter. For education, the odds of Clinton winning were multiplied by 1.26 for every 1 unit increase in education level, when all other variables were held fixed. Similarly, for household size, the odds of Clinton winning were multiplied by .856 for every 1 unit increase in household size. Finally, for income level, the odds of Clinton winning were multiplied by .834 for every 1 unit increase in income level. Overall, these explanatory variables demonstrate a higher likelihood for voting for Clinton with a younger age, a female voter, a non-white voter, higher education level, smaller household size and smaller income level.

The main purpose of this research paper was to set out and predict the outcome of the 2016 election between the two presidential candidates, Trump and Clinton. The overall goal of the model fit was to predict the winning presidential candidate in each state (and the District of Columbia) and to predict the final percentage of the national popular vote for each major candidate. With the logistic regression model, we were able to predict the winning candidate nationally as well as at the state level, using variables for each individual in the survey.

However, our model fit was only fair; despite having all variables significant and seeing trends in each variable for candidate preference, the data might not have represented the outcomes for each state accurately. Our study also looked at individual preference from surveys, which could be biased. For example, some of the samples from the states were skewed (eg. only females took the survey). Our data, provided by the University of Southern California’s Dornsife Center for Economic and Social Research, was only given to 3000 respondents, all of which were members of the Understanding America Study election panel. Additionally, our study used variable selection only for the overall global model; this choice was made to smooth the overall results and apply consistency.

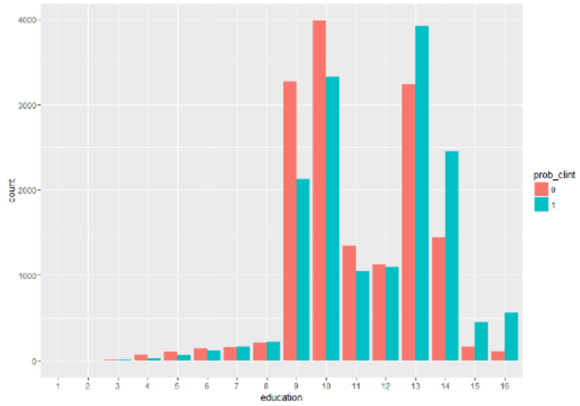


Figure 7. Bar graph showing count data for highest education levels on the x-axis and split for Trump (0) and Clinton (1); education levels increase from left to right, with category 2 being up to 4th grade, category 8 being up to 12th grade, and category 16 being a doctorates degree.

Performing variable selection for each state might have been better able to illustrate the differences between each state. Finally, our model did not take into account the individual characteristics of each state; for instance, California and Texas are both huge states and so had a larger number of participants in the survey whereas small states had far fewer. This means that higher populated states had more weight in the model and possibly could have skewed the results toward a particular candidate.

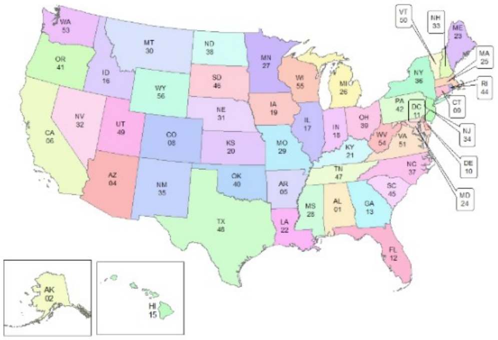
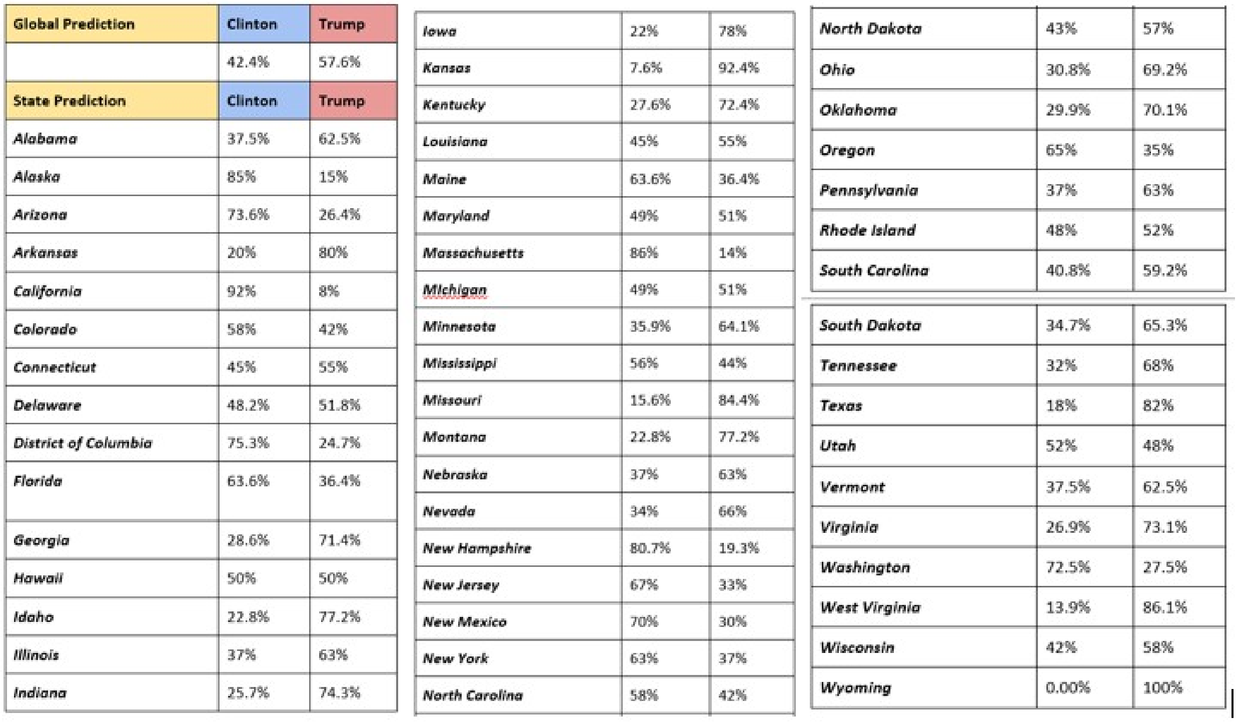


Figure 8. State FIPS codes (courtesy of http://www.icip.iastate.edu/maps/refmaps/STFIPS).

Future studies in predicting the outcome of the presidential election with survey data might use a larger sample size overall, fairly predict for each state based on their characteristics (eg. size), incorporate variables like the state of the economy (similar to the “Time for Change” model), and possibly include past election outcomes (trends for each state). we were able to analyze our data using a logistic regression. Overall, however, were able to predict percentages for each state and for the national popular vote on which candidate is most likely to win. Our results pointed towards Trump (ahead of Clinton by around 15%) and additionally, were able to show how variables like race, age, sex, state of residency, house size, and education level impacted an individual’s candidate choice.

Table 1. Global and State predictions for each candidate.



Global Prediction	Clinton	Trump
	42.4%	57.6%
State Prediction	Clinton	Trump
Alabama	37.5%	62.5%
Alaska	85%	15%
Arizona	73.6%	26.4%
Arkansas	20%	80%
California	92%	8%
Colorado	58%	42%
Connecticut	45%	55%
Delaware	48.2%	51.8%
District of Columbia	75.3%	24.7%
Florida	63.6%	36.4%
Georgia	28.6%	71.4%
Hawaii	50%	50%
Idaho	22.8%	77.2%
Illinois	37%	63%
Indiana	25.7%	74.3%
Iowa	22%	78%
Kansas	7.6%	92.4%
Kentucky	27.6%	72.4%
Louisiana	45%	55%
Maine	63.6%	36.4%
Maryland	49%	51%
Massachusetts	86%	14%
Michigan	49%	51%
Minnesota	35.9%	64.1%
Mississippi	56%	44%
Missouri	15.6%	84.4%
Montana	22.8%	77.2%
Nebraska	37%	63%
Nevada	34%	66%
New Hampshire	80.7%	19.3%
New Jersey	67%	33%
New Mexico	70%	30%
New York	63%	37%
North Carolina	58%	42%
North Dakota	43%	57%
Ohio	30.8%	69.2%
Oklahoma	29.9%	70.1%
Oregon	65%	35%
Pennsylvania	37%	63%
Rhode Island	48%	52%
South Carolina	40.8%	59.2%
South Dakota	34.7%	65.3%
Tennessee	32%	68%
Texas	18%	82%
Utah	52%	48%
Vermont	37.5%	62.5%
Virginia	26.9%	73.1%
Washington	72.5%	27.5%
West Virginia	13.9%	86.1%
Wisconsin	42%	58%
Wyoming	0.00%	100%

References

Abramowitz, A. I. (1988). An improved model for predicting presidential election outcomes. *PS: Political Science and Politics, 21*(4), 843. doi:10.2307/420023

Colomer, J. , & Johnston, R. (2005). Handbook of electoral system choice. *British Journal of Politics and International Relations*, *7*(2), 281-291.

Hoffman, M. M. (1996). The illegitimate president: Minority vote dilution and the electoral college. *The Yale Law Journal*, *105*(4), 935–1021. doi:10.2307/797244

Lewis-Beck, M. S., & Rice, T. W. (1984). Forecasting presidential elections: A comparison of naive models. *Political Behavior, 6*(1), 9-21. doi:10.1007/bf00988226

Miller, A. H., Wattenberg, M. P., & Malanchuk, O. (1986). Schematic assessments of presidential candidates. *Am Polit Sci Rev American Political Science Review, 80*(02), 521-540. doi:10.2307/1958272

Neale, T. (2012). The electoral college: How it works in contemporary presidential elections. *Congressional Research Service*

Page, B. I., & Jones, C. C. (1979). Reciprocal effects of policy preferences, party loyalties and the vote. *Am Polit Sci Rev American Political Science Review, 73*(04), 1071-1089. doi:10.2307/1953990

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Election forecasts with twitter: How 140 characters reflect the political landscape. *Social Science Computer Review, 29*(4), 402-418. doi:10.1177/0894439310386557

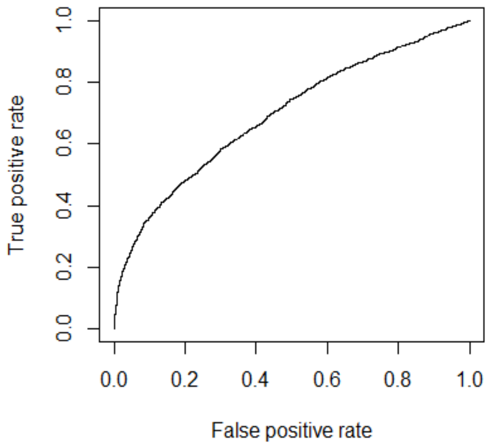
Warf, B. (2009). The u.s. electoral college and spatial biases in voter power. *Annals of the Association of American Geographers*, *99*(1), 184-204.

Appendix

Global Model Output:

```
> global.model<-
glm(prob_clint~age+gender+race+education+hhsize_cat+inc_cat+statereside,family=bino
mial,data=modeldf)

> summary(global.model) #all variables significant

Call:
glm(formula = prob_clint ~ age + gender + race + education + 
    hhsize_cat + inc_cat + statereside, family = binomial, data = modeldf)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.3318  -0.9781   0.0981   1.0280   2.6561  

Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)   -2.934e+00  1.794e-01 -16.352  < 2e-16 ***
age           -6.349e-03  9.155e-04  -6.934 4.08e-12 ***
gender1       -5.291e-01  2.673e-02 -19.795  < 2e-16 ***
race2          4.122e+00  1.223e-01  33.711  < 2e-16 ***
race3          1.745e+00  1.129e-01  15.452  < 2e-16 ***
race4          1.139e+00  1.300e-01   8.767  < 2e-16 ***
race5          1.847e+00  2.670e-01   6.919 4.54e-12 ***
race6          4.588e-01  6.028e-02   7.612 2.70e-14 ***
education      2.344e-01  6.992e-03  33.527  < 2e-16 ***
hhsize_cat    -1.551e-01  2.213e-02  -7.008 2.42e-12 ***
inc_cat       -1.804e-01  1.883e-02  -9.580  < 2e-16 ***
statereside2   3.053e+00  3.075e-01   9.926  < 2e-16 ***
statereside4   1.712e+00  1.694e-01  10.107  < 2e-16 ***
statereside5   7.349e-01  1.756e-01   4.184 2.86e-05 ***
statereside6   2.277e+00  1.547e-01  14.723  < 2e-16 ***
statereside8   1.446e+00  1.766e-01   8.191 2.60e-16 ***
statereside9   1.042e+00  2.118e-01   4.921 8.61e-07 ***
statereside10  1.610e+00  3.180e-01   5.063 4.12e-07 ***
statereside11  2.782e+00  4.296e-01   6.476 9.41e-11 ***
statereside12  1.503e+00  1.588e-01   9.468  < 2e-16 ***
statereside13  2.801e-01  1.689e-01   1.658 0.097221 .  
statereside15  8.254e-01  3.039e-01   2.716 0.006612 ** 
statereside16  5.180e-01  2.339e-01   2.215 0.026754 *  
statereside17  1.478e+00  1.589e-01   9.298  < 2e-16 ***
statereside18  1.030e+00  1.679e-01   6.134 8.57e-10 ***
statereside19  9.205e-01  1.748e-01   5.267 1.38e-07 ***
statereside20  1.150e-01  1.861e-01   0.618 0.536568    
```

```
statereside21  8.325e-01  1.754e-01   4.745 2.08e-06 ***
statereside22  6.930e-01  2.025e-01   3.423 0.000620 ***
statereside23  2.072e+00  2.028e-01  10.220  < 2e-16 ***
statereside24  1.121e+00  2.053e-01   5.461 4.74e-08 ***
statereside25  2.576e+00  2.044e-01  12.600  < 2e-16 ***
statereside26  1.273e+00  1.587e-01   8.023 1.03e-15 ***
statereside27  1.308e+00  1.685e-01   7.766 8.09e-15 ***
statereside28  6.184e-01  2.073e-01   2.983 0.002850 **
statereside29  6.906e-01  1.693e-01   4.079 4.52e-05 ***
statereside30  4.819e-01  2.506e-01   1.923 0.054475 .
statereside31  1.191e+00  2.009e-01   5.931 3.01e-09 ***
statereside32  6.471e-01  2.430e-01   2.663 0.007749 **
statereside33  3.036e+00  3.428e-01   8.856  < 2e-16 ***
statereside34  1.648e+00  1.798e-01   9.168  < 2e-16 ***
statereside35  1.792e+00  2.000e-01   8.961  < 2e-16 ***
statereside36  1.776e+00  1.592e-01  11.157  < 2e-16 ***
statereside37  1.407e+00  1.610e-01   8.737  < 2e-16 ***
statereside38  6.753e-01  2.367e-01   2.853 0.004335 **
statereside39  1.124e+00  1.561e-01   7.198 6.10e-13 ***
statereside40  5.463e-01  1.756e-01   3.111 0.001865 **
statereside41  1.945e+00  1.811e-01  10.744  < 2e-16 ***
statereside42  1.338e+00  1.551e-01   8.624  < 2e-16 ***
statereside44  1.744e+00  4.582e-01   3.806 0.000141 ***
statereside45  9.811e-01  2.047e-01   4.793 1.64e-06 ***
statereside46  1.429e+00  1.934e-01   7.390 1.46e-13 ***
statereside47  8.724e-01  1.706e-01   5.115 3.14e-07 ***
statereside48  1.100e+00  1.567e-01   7.016 2.28e-12 ***
statereside49  1.471e+00  2.223e-01   6.617 3.67e-11 ***
statereside50  8.311e-01  2.707e-01   3.070 0.002140 **
statereside51  5.118e-01  1.684e-01   3.039 0.002371 **
statereside53  1.577e+00  1.742e-01   9.054  < 2e-16 ***
statereside54  4.643e-01  1.804e-01   2.574 0.010046 *
statereside55  1.460e+00  1.631e-01   8.952  < 2e-16 ***
statereside56 -1.232e+01  9.576e+01  -0.129 0.897660
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 42904  on 30949  degrees of freedom
Residual deviance: 35307  on 30889  degrees of freedom
  (5936 observations deleted due to missingness)
AIC: 35429

Number of Fisher Scoring iterations: 12
```



Roc curve for global model. AUC value=0.694.

R Code:

-main.R

```
#####################################
options(warn=-1)
#Presidential Election Data Analysis
#Code to read in the data files to R
#Unfried 10/4/16
#####################################
rm(list=ls())
#Haven package has function to read .dta files into R
#You will need to install it before you load it
library(haven)
#Hmisc package has "describe" function used below
#You will need to install it before you load it
library(Hmisc)
library(gdata)
library(car)
library(boot)
library(caret)
library(ggplot2)
#read_dta function is from Haven package; choose.file() does not work
polldata <- read_dta("polldata.dta")
fulldata <- read_dta("fulldata_10112016.dta")
#otherwise polldata is a "tibble" instead of a data frame
pd <- as.data.frame(polldata)
fd <- as.data.frame(fulldata)
source("functions.R")
#Variables for global model:
#qual <- c("gender", "race", "bornus")
#quant <- c("education", "age", "hhincome", "statereside")
globalV <- c("gender", "race", "bornus", "education", "age", "hhincome", "statereside") #<- rbind(qual, quant)
#Put as factor each necessary variable:
#fd$statereside = as.factor(fd$statereside)
fd$gender = as.factor(fd$gender)
fd$race = as.factor(fd$race)
fd$bornus = as.factor(fd$bornus)
#Get the data cleaned for fd dataset.
data = getNewData(fd, globalV)
data = as.data.frame(data)
#Loop for getting the model for each state:
states <- states_list(data) #get a list with the data of each state in each position of the list.
modelS = c()
states.prediction <- as.data.frame(matrix(0,1:56, 1:6))
for(i in 1:56) {
 id <- i
 if(i != 3 && i != 7 && i != 14 && i != 43 && i != 52 ) { #there are not state 3, 7...
   newdata = fixData(states[[i]]) # fix the data for not having varibles that not change their values in all of the row.
   model <- glm(formula = clint ~ . , family = binomial(), data = newdata)
   #cross <- cv.glm(newdata, model, K = 10)
   #keeping the prediction of each state.
   p <- predict(model, data = subset(newdata, select = c(1:length(newdata)-1)),
type = 'response')
   p.states <- ifelse(p > 0.5, 1, 0)
   clinton <- sum(p.states == 1) / length(p)
   trump <- sum(p.states == 0) / length(p)
   prob <- mean(p)
   states.prediction[i, 1] <- id #states id number
   states.prediction[i, 2] <- NA #Addig Na so that we could add name of each state
latter in this place.
   states.prediction[i, 3] <- clinton #prob of clinton to winnig depends on the model
   states.prediction[i, 4] <- trump #prob of trump for winning depends on the model
   states.prediction[i, 5] <- prob #this is a probability of the mean of the odds for
clinton winning
   a<-na.omit(newdata$clint)
   states.prediction[i, 6] <- sum(a==1) / length(a) #probability of clinto winning
depending on the dataset, counting how many people says they will vote clinton.
   #Print every number i of state and summary of the model of the state[i].
##  print(i)
##  print(summary(model)$coef)
   #print(vif(model))
 }
}


states.prediction[1:2, 2] <- state.name[1:2]
states.prediction[4:6, 2] <- state.name[3:5]
states.prediction[8:10, 2] <- state.name[6:8]
states.prediction[11, 2] <- "District of Columbia"
states.prediction[12:13, 2] <- state.name[9:10]
states.prediction[15:42, 2] <- state.name[11:38]
states.prediction[44:51, 2] <- state.name[39:46]
states.prediction[53:56, 2] <- state.name[47:50]
states.prediction[states.prediction$names == "NA"] <- NA

names(states.prediction) <- c("id", "names", "clinton", "trump",
"clinton_mean_prob"

                              , "no_model_mean")

#------------------------------------------------------------------------------------------------

#Logistic regression model for clinton prediction if she loses or wins.

 modelG <- glm(formula = clint ~ ., family = binomial(), data = data)

 #cVG <- cv.glm(data, modelG)

#start prediction procees for the actual data:

global.pred <- predict(modelG, data = subset(data, select = c(1:length(data)-1)),
type = 'response')

fitted.pred <- ifelse(global.pred > 0.5, 1, 0)

global.predictions <- data.frame(matrix(0, 1:56, 1:5))

#each vector means the same as sin the states.prediction, but this is for the global
prediction of the model.

global.predictions[,1] <- 1

global.predictions[,2] <- sum(fitted.pred == 1) / length(fitted.pred)

global.predictions[,3] <- sum(fitted.pred == 0) / length(fitted.pred)

global.predictions[,4] <- mean(global.pred)

a<-na.omit(data$clint)

global.predictions[,5] <- sum(a==1) / length(a)

names(global.predictions) <- c("id", "clinton", "trump", "clinton_mean_prob",
"no_model_mean")

##ROC curve analysis with cross validation test:

library(ROCR)

test <- as.data.frame(data[200000:length(data),])

p <- predict(modelG, newdata=subset(test,select=c(1:length(test)-1)),
type="response")

pr <- prediction(p, test$clint)

prf <- performance(pr, measure = "tpr", x.measure = "fpr")

plot(prf)

auc <- performance(pr, measure = "auc")

auc <- auc@y.values[[1]]
```

```
##Graphics for every variable:

raceGraph <- ggplot(na.omit(data), aes(race))+geom_bar(aes(fill=clint),
position="dodge")

ageGraph <- ggplot(na.omit(data), aes(age))+geom_bar(aes(fill=clint),
position="dodge")

genderGraph <- ggplot(na.omit(data), aes(gender))+geom_bar(aes(fill=clint),
position="dodge")

educationGraph <- ggplot(na.omit(data), aes(education))+geom_bar(aes(fill=clint),
position="dodge")

bornusGraph <- ggplot(na.omit(data), aes(bornus))+geom_bar(aes(fill=clint),
position="dodge")

hhincomeGraph <- ggplot(na.omit(data), aes(hhincome))+geom_bar(aes(fill=clint),
position="dodge")

stateGraph <- ggplot(na.omit(data), aes(statereside))+geom_bar(aes(fill=clint),
position="dodge")
```

-Functions.R

```
#changing the probability of vote for clinton and trump, getting to 1 clinton
# and 0 voting for turmp:
clint_01 <- function(clint, trump) {
 count <- 0
 for(i in 1:length(clint)) {
   if(clint[i] > trump[i]) {
     clint[i] = 1
     #fd$prob_trump[i] = 0
   }else if(clint[i] < trump[i]) {
     clint[i] = 0
     #fd$prob_trump[i] = 1
   } else if(clint[i] == trump[i]) {
     if(count %% 2 == 0) {
       clint[i] = 1
       count = count +1
     } else {
       clint[i] = 0
```

```
      count = count +1
     }
   }
 }
 clint = as.factor(clint)
 return(clint)
}
#Get states subsets:
states_list <- function(fd) {
 states = list()
 for(j in 1:56) {
   if(j != 3 || j != 52 || j != 14 || j != 43) {
     aux <- subset(fd, statereside == j, drop = FALSE)
     states[[j]] <- drop.levels(aux, reorder = FALSE)
   }
 }
 return(states)
}
#Return the data with the variables that you pas in the var vector from the fd
dataset:
getNewData <- function(fd, var) {
   data = subset(fd, select = var)
   data = drop.levels(data, reorder = FALSE)
   data$clint <- clint_01(fd$prob_clint, fd$prob_trump)
   data$clint <- as.factor(data$clint)
   data[data == "NaN"] = NA
   return(data)
}
#fix the data of the states, cause there are some states that one or more
variable(s) can have the same value for all of the rows
fixData <- function(data) {
 aux = data
```

```
  count = 1
  a <- apply(aux, 2, function(x) length(unique(x)) == 1)
  for(i in a) {
    if(i == TRUE) {
      aux[[count]] <- NULL
      count = count -1
    }
    count = count +1
  }
  return(aux)
}
```