An Analysis of US Counties from the 2016 General Election Swing States

Brianna Vetter and Ari Fleischer

Summary:

The 2016 US General Election had notoriously unpredictable results. Most major news outlets had incorrectly predicted the victor. These incorrect election predictions were based on prediction polls for several key states. These key states are generally referred to as "Swing States."

This project has two main goals. The first goal is to analyze demographic data for US counties in the swing states to isolate factors that may have impacted local election results. The second goal is to create two prediction models using these factors that can predict the percentage of voters in a country that will vote for each major party candidate. The hope is that an accurate county level prediction model might make it easier to spot polling biases for future elections.

It was decided that two models would be needed because there are numerous candidates that run in every presidential election. In the 2016 general election, there were 4 political parties that had a measurable amount of the vote in each county. These parties were the Democratic Party, the Republican Party, the Green Party and the Libertarian Party [2]. For the sake of our analysis only the two major parties, Democratic Party and the Republican Party, will be used. However, this does not negate the effect of the other parties on the data, so it cannot simply be assumed that the models will be inverses of each other. Two models must be made for accurate analysis.

Three datasets were chosen for this project. All three datasets were collected from Kaggle. The first is called US Household Income Statistics. The data it contains was collected by Golden Oak Research Group for use in real estate and business investment research [1]. However, the research group states that much of the data comes from US census data. The results of the primaries and the general election will be from a set called US Election [2]. This dataset also contains a lot of demographic information such as racial and education level by county. The 2016 Election Polls dataset was originally collected by Five Thirty Eight [3]. This set only gives state level results.

The 2016 Election Polls dataset were used to identify the swing states for analysis. The US Household Income Statistics dataset and US Election datasets were combined and put into tidy format. This combined dataset was filtered to only included the swing states, and the parameters and model were based on this information. The dataset was split into train/validation/test for further analysis. The models were built with stepwise model selection. Once the two models were built, ANOVA testing was used to determine if all model parameters were needed to explain the relationships in the data. The results of this project include two

final models and a list of their parameters, as well as the RMSE for each on the test set of the data.

Methods:

The first step of the project was to load our datasets into RStudio using read_csv() for the income and state polls datasets [1], [3] and read_delim() for the election dataset [2]. For the US Household Income Statistics dataset, missing values were originally zeros. These had to be converted to N/A. As the information in this dataset was by city, it had to be grouped_by() the county. The US Election dataset originally had a single column for state and county. This was fixed using separate(). With both datasets in tidy format, they were combined using inner_join(). The function na.omit() was used on the combined dataset so that only complete cases remained. This removed only 34 rows from the combined dataset.

Next, a separate data frame was made for the US Election dataset. This set was modified was modified in several way to collect the statewide winner. The summation of the 2016 Democrat and Republican votes were grouped by state, and an extra column was added to the data frame using mutate() to record the statewide winner. The 2016 Election Polls dataset was first mutated to include an extra column called Predicted Winner that was determined by doing an ifelse() with the adjust poll numbers for Trump and Clinton. Two mutated data sets were then inner_joined(). The new combined set was filtered to only include rows with a grade B+ or better and that had a date within one month of the general election. Also, they conditionally had to be in a state where the predicted winner of the state did not match the statewide winner. After visualizing the number of wrong polls per state, it was concluded that Florida, Michigan, North Carolina, Pennsylvania, and Wisconsin were the swing states for the 2016 election (Figure 1). Every single one of these states had predicted that the Democratic party would win, whereas the Republican party won on election day.

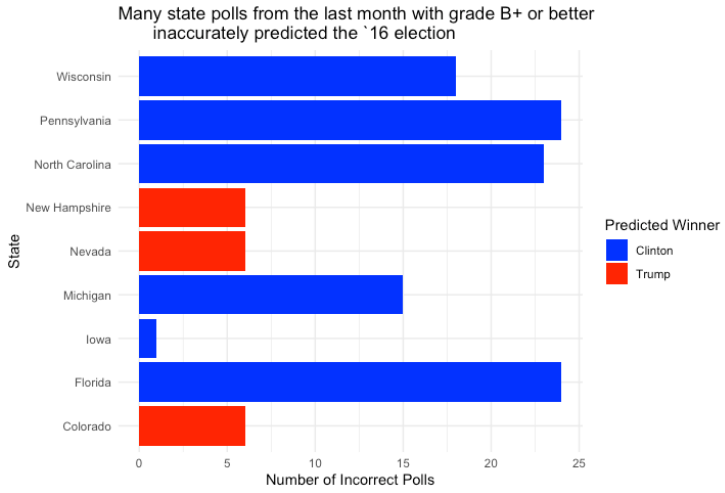

Figure 1 – Swing State Selection Figure using Datasets [2] and [3]

The combined election data set was then filtered to only include the 5 swing states. 230 rows remained in this dataset with most states having between 40-50 counties present in the dataset. The breakdown by state can be seen in figure 2. Based on this information, it was concluded that the swing state data would need to be kept in one data set. A 50/25/25 train/validation/test split was chosen so that a reasonable amount of data would be in the validation and test sets.

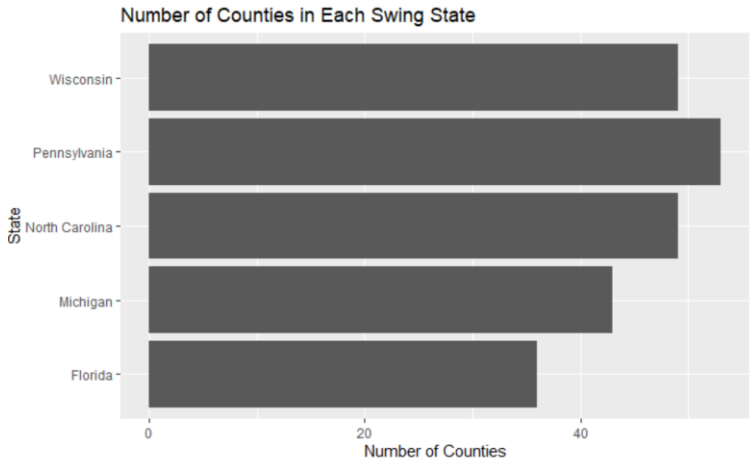

Figure 2 – Combined Election Dataset Number of Counties in Each Swing State

Next, the combined election dataset was used to visualize numerous different possible predictors against the response variables. The response variables were the percent of the vote in each county for the Republican party and the Democratic party. All possible predictors were visualized against both response variables using ggplot2 scatterplots and any necessary log transformations on the predictors was done. Seventeen of these possible predictors were selected to move on to the stepwise model selection stage. The predictors can be grouped into five categories: education, money, race, economic status, and miscellaneous (Table A1).

Stepwise model selection was performed on the 17 parameters twice, once for each response variable (Figures 5, 6). ANOVA testing was done to determine whether any predictor(s) could be removed from the models and have the models still sufficiently explain the relationships in the data. The null hypothesis, H_0, is that the reduced model adequately explains the relationships in the data. The alternative hypothesis, Ha, is that the full model better explains the relationships in the data. An alpha cutoff of 0.05 was used for the testing.

Results:

Stepwise model selection on the 17 possible predictors was used to obtain the models for Democrats `16 Vote Percentage and Republicans `16 Vote Percentage as response variables. The best predictor for both response variables was White (Not Latino) Population Percentage (Figures 3, 4). The stepwise model selection gave the same predictors in the same order for both models, but with different RMSE values. The predictors from the stepwise model are, in

order: White (Not Latino) Population Percentage, At Least High School Diploma Percentage, Asian American Population Percentage, log Poverty Rate, Other Race or Races Percentage, and log Unemployment Rate (Figures 5, 6).

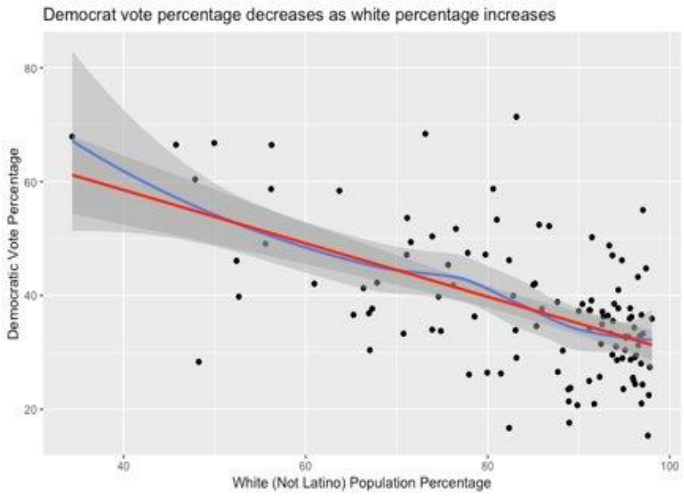

Figure 3 - Democrats `16 Vote Percentage Vs. White (Not Latino) Population Percentage

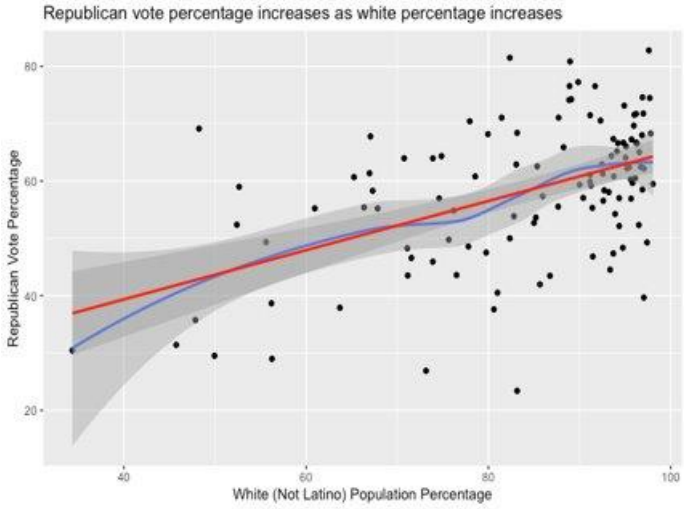

Figure 4 - Republicans`16 Vote Percentage Vs. White (Not Latino) Population Percentage on the bottom

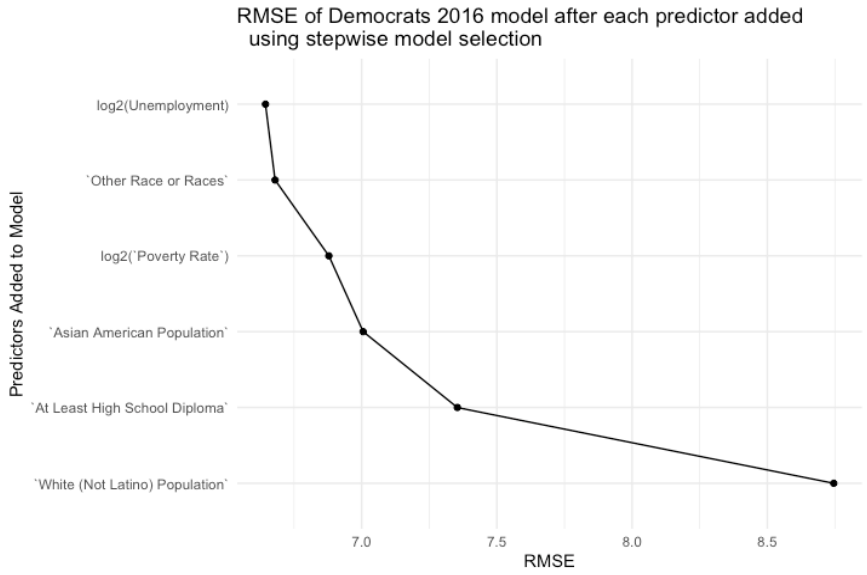

Figure 5 - Stepwise model selection results for Democrats '16 Vote Percentage. Predictors go from bottom to top in the order added to the model

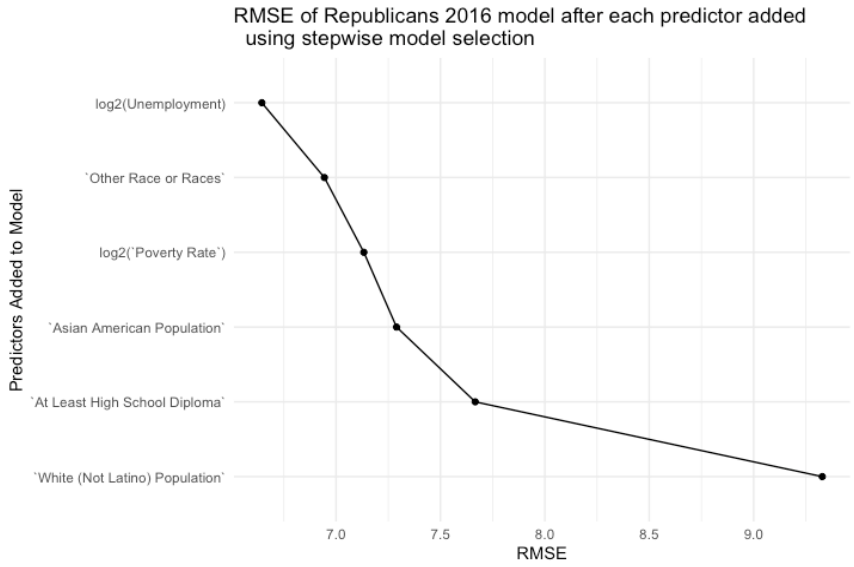

Figure 6 - Stepwise model selection results for Republicans `16 Vote Percentage. Predictors go from bottom to top in the order added to the model

When viewing the summary of the models obtained from stepwise model selection, not all of the added predictors significantly influenced the vote percentage. The predictors from log Poverty Rate and on in both models all have p-values above 0.05, which is the alpha cutoff (Tables 1, 2). Therefore, we cannot reject the null hypothesis, that that predictor has no effect on the vote percentage, with the other predictors present in the model. As a result, ANOVA testing was used. The full models are the response variables, including all six predictors. The reduced models are the response variables with only the first three predictors. The resulting p-value from the ANOVA testing for both models was greater than our alpha value of 0.05 (Table 3). Therefore, the null hypothesis failed to be rejected. Therefore, the reduced model (only

White (Not Latino) Population Percentage, At Least High School Diploma Percentage, and Asian American Population Percentage as predictors) adequately explains the relationships in the data for both response variables.

Response Variable	Predictor	Coefficient	P-Value	Significance Level
Democrats 2016 Vote Percentage	Intercept	-53.24500	0.00953	**
	White (Not Latino) Population Percentage	-0.65381	2.42e-16	***
	At Least High School Diploma Percentage	1.42746	2.69e-11	***
	Asian American Population Percentage	1.45482	0.03036	*
	Log Poverty Rate	3.46912	0.07586	.
	Other Race or Races Percentage	-2.64792	0.09874	.
	Log Unemployment Rate	-3.57762	0.16488	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 1 – Coefficients and P-Values of the Predictors in the Democrats `16 Vote Percentage Model Obtained by Stepwise Model Selection

Response Variable	Predictor	Coefficient	P-Value	Significance Level
Republicans 2016 Vote Percentage	Intercept	165.70826	5.11e-12	***
	White (Not Latino) Population Percentage	0.62509	3.06e-14	***
	At Least High School Diploma Percentage	-1.54452	1.14e-11	***
	Asian American Population Percentage	-1.50362	0.0343	*
	Log Poverty Rate	-3.76547	0.0686	.
	Other Race or Races Percentage	2.35937	0.1635	
	Log Unemployment Rate	4.24026	0.1200	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 2 – Coefficients and P-Values of the Predictors in the Republicans`16 Vote Percentage Model Obtained by Stepwise Model Selection

Response Variable	Full Model	Reduced Model	P-Value
Democrats 2016 Vote Percentage	`Democrats 2016` ~ `White (Not Latino) Population` + `At Least High School Diploma` + `Asian American Population` + log2(`Poverty Rate`) + `Other Race or Races` + log2(Unemployment)	`Democrats 2016` ~ `White (Not Latino) Population` + `At Least High School Diploma` + `Asian American Population`	0.07043
Republicans 2016 Vote Percentage	`Republicans 2016` ~ `White (Not Latino) Population` + `At Least High School Diploma` + `Asian American Population` + log2(`Poverty Rate`) + `Other Race or Races` + log2(Unemployment)	`Republicans 2016` ~ `White (Not Latino) Population` + `At Least High School Diploma` + `Asian American Population`	0.07671

Table 3 - Results of ANOVA Testing Using Alpha Cutoff of 0.05

After ANOVA testing, the final models for both Democratic `16 Vote Percentage and Republican `16 Vote Percentage response variables include the same three predictors: White (Not Latino) Population Percentage, At Least High School Diploma Percentage, and Asian American Population Percentage. The RMSE of the Democrats model on the test set was 7.841428, while the RMSE of the Republicans model on the test set was 8.157785 (Table 6). While the predictors for both models were the same, the absolute value of the coefficients are different (Table 4, 5). This means that the absolute value of the effect a one unit increase in the predictor has on the response variables is not the same for both models. Also, the signs of the coefficients for each predictor are flipped between the two models (Table 4, 5). This makes sense because the response variables are inversely related. For example, White (Not Latino) Population Percentage has a negative relationship with Democrats `16 Vote Percentage, but a positive relationship with Republicans `16 Vote Percentage. Additionally, At Least High School Diploma Percentage and Asian American Population Percentage have a positive relationship with Democrats `16 Vote Percentage, but a negative relationship with Republicans `16 Vote Percentage.

Response Variable	Predictor	Coefficient	P-Value	Significance Level
Democrats 2016 Vote Percentage	Intercept	-18.84095	0.0866	.
	White (Not Latino) Population Percentage	-0.62381	< 2e-16	***
	At Least High School Diploma Percentage	1.25463	1.47e-12	***
	Asian American Population Percentage	1.22105	0.0485	*
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 4 – Coefficients and P-Values of the Predictors in the Final Democrats `16 Vote Percentage Model

Response Variable	Predictor	Coefficient	P-Value	Significance Level
Republicans 2016 Vote Percentage	Intercept	127.73403	< 2e-16	***
	White (Not Latino) Population Percentage	0.60020	5.61e-15	***
	At Least High School Diploma Percentage	-1.37902	2.82e-13	***
	Asian American Population Percentage	-1.32041	0.0436	*
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 5 – Coefficients and P-Values of the Predictors in the Final Republicans`16 Vote Percentage Model

Response Variable	Final Model (Using lm)	RMSE on the Test Set
Democrats 2016 Vote Percentage	`Democrats 2016` ~ `White (Not Latino) Population` + `At Least High School Diploma` + `Asian American Population`	7.841428
Republicans 2016 Vote Percentage	`Republicans 2016` ~ `White (Not Latino) Population` + `At Least High School Diploma` + `Asian American Population`	8.157785

Table 6 - RMSE of the Final Models on the Test Set

Discussion:

The two models wound up having the same predictors. This was expected because Democrat vote percentage and Republican vote percentage are highly correlated. The models were created using certified election results and full demographic data. Therefore, the data was not subject to the same sampling biases present in polling data. The models can be used as a baseline to try and figure out where the 2016 election polls were biased, resulting in incorrect predictions. The models can also be used to predict how a county will vote in future elections based on the county's demographic breakdown for the predictors in the models. If the data proves consistent from year to year, it is possible that models will have a stronger role in election predictions in the future due to their avoidance of sampling bias. In addition, the models can be used to improve representation in future polling by providing a baseline to compare polling data to. This baseline can make sure that the polls have an accurate sampling of the predictors present in the models.

None of the predictors in the final model came from the income dataset [2]. Knowing this, it would be interesting to replicate the analysis on the swing state data. This is because the election dataset contained more complete information for every county in each of the five swing states. It contained 389 counties in total, while the income dataset had more incomplete rows. The join with the income dataset is what caused the number of complete counties to be reduced by about 40% to only 230 counties. These additional 159 rows might result in different models. Exploratory data analysis would need to be redone to determine the transformations for the predictors, and stepwise model selection would need to be redone with the new transformed 15 predictors, excluding the two from the income dataset. The new models could then be compared to the original results. Since the complete data for every county in each of the five swing states is available, both the new and the original models can be tested to see if the models give more accurate election results than the polling data did.

An early plan was to create models for each swing stare. Unfortunately, as no swing state has over 100 counties it will not be feasible to make separate models for each of them. This is because the data still needs to be split into at least a training and testing set to avoid overfitting, and the population is too small.

Another idea for expanding the project is to build models for the response variables from all the non-swing states. This could be done in several ways. One would be to create a separate model for each response variable form states that always have vote Democratic or states that always vote Republican. Another way would be to create one big model for each response variable using all the non-swing states. The models of the non-swing states can then be compared to the swing states model to see if how the best predictors vary from the swing states to the non-swing states. It would be interesting to test the accuracy of our models on the results from the 2020 presidential election. It would also be interesting to rebuild the same models for the 2020 election data using the same swing states and predictors but updated demographic information. This could be used to determine whether vote percentage is best modeled by the same predictors from one election cycle to the next.

Statement of contributions:

Ari Fleischer: Did the coding to figure out which were the swing states we should analyze by looking at the state polls. In addition, helped with accounting for N/As in the income dataset and transformed the income dataset from by city to by county. Also, made any transformations that were needed for the predictors and then did the stepwise model selection and ANOVA testing for both models. Additionally, wrote the results and discussion section of the report.

Brianna Vetter: Loaded in initial datasets and transformed into tidy form. Joined datasets and mutated some data to create additional parameters. Visualized many possible parameters against both response variables and isolated likely options for stepwise model selection.

References:

[1] Golden Oak Research Group, LLC, "U.S. Income Database Kaggle". Publication: 05-August-2017. Available: https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations?select=kaggle_income.csv. [Accessed, 27, October 2020].

[2] B. Tunguz, "US Elections Dataset," *Kaggle*, 13-Oct-2020. [Online]. Available: https://www.kaggle.com/tunguz/us-elections-dataset?select=usa-2016-presidential-election-by-county.csv. [Accessed: 29-Oct-2020].

[3] FiveThirtyEight 2016 Election Forecast, "2016 Election Polls," *Kaggle*, 03-Nov-2016. [Online]. Available: https://www.kaggle.com/fivethirtyeight/2016-election-polls. [Accessed: 29-Oct-2020].

Appendix:

Type	Predictor
Response	Democratic Vote Percentage, Republican Vote Percentage
Education	Percentage of population with Graduate Degrees, Percentage of population with at least a Bachelor's degree, Percentage of population with at least a High School Diploma
Money	Log median income, Mean income
Race	White (Not Latino) Population, African American Population, Asian American Population, Other Race or Races, Log Latino Population, Log Native American Population
Economic Status	Log Children Under 6 Living in Poverty, Log Adults 65 and Older Living in Poverty, Log Unemployment, Log Poverty Rate
Miscellaneous	Percent of the population that voted, Log median age

Table A1 – Predictors Chosen for Possible Addition to the Models

GitHub Link For the Project Code:

https://github.com/AriFleischer13/DS5110FinalProject/blob/main/AFBVDS5110FinalProject.Rmd