Date:	31 October 2016
To:	Christian Lucero
From:	The Old Dominions
Subject:	Prediction and Modelling of Virginia for the 2016 Presidential Election

1 Executive Summary

When attempting to make a prediction for the upcoming presidential election in the state of Virginia, we decided to build a model that focused on polling data as the primary data source. After using linear regression to build the model, our group decided to test the accuracy of the model using data from previous elections[4]. Once satisfied with the accuracy of our model, we used our model to predict the results of the 2016 presidential election, with a final prediction of the Democratic Candidate Hillary Clinton receiving 54.47 percent of the vote, and the Republican candidate Donald Trump receiving 45.53 percent of the vote.

2 Introduction and Problem Setting

When faced with the problem of predicting the outcome of an election, the Old Dominions had several key problems that would have to be faced. Though there is a sizable chunk of past elections to use for guidance, changes in demographics and societal trends made us hesitant to look too far back for fear of using data that might be inapplicable to todays election. Additionally, we made the decision to ignore third party influence in this election as we assumed that due to the nature of the candidates the third party is primarily pulling from individuals who are dissatisfied with both parties, and should not be pulling primarily from one party over another.

After extensive thought and research, the Old Dominions decided that there were too general ways of going about solving the problem. In both cases, linear regression would be used to formulate a model that based on the data that was input, could predict a value for the outcome of the election. The difference between the two main approaches therefore was not a difference in modelling methods but a distinction in the type of data that should be input.

The first option we very seriously considered was building a model based on population demographics in different counties in Virginia. The idea behind this model is: using past datasets we can make a prediction based on how different demographics (for example, different age groups, genders, race, etc.) are likely to vote. Once we have a general model built with past election datasets, we can feed data from 2016 in and feel comfortable with our prediction.

The other option that the Old Dominions thought was viable was a model built using polling data. The general idea behind this model is as follows: using the fact that the election is such a popular topic and there is a sizable amount of polls that exist in relation to it, we can build a model analyzing the trends over time in how individuals feel about voting for a specific candidate, and use this model to predict what will happen on election day.

In the end, the Old Dominions decided to use the second model, using polling data instead of demographic data. The decision was not easy and required debate, however in the end we recognized that although the demographic model has many strengths, it is strongest in predicting the winner of a typical election. However, the upcoming election is very far from typical due to various factors regarding the candidates, and we feared that a model built on demographics would not address the specifics in regards to the candidates. A polling model was therefore more appealing as specific opinions based on the candidates are able to be accurately reflected.

3 Data Description

After some cleaning and adjustments to the data, we were able to condense our primary datasets down to the essentials. The csv files used were taken from a site that gave overall polling data for all states in the years 2008[1], 2012[3], and 2016[2]. These datasets included categorical variables such as the pollster used, the state the pollster was located, and the dates of the recorded pollsters. The quantitative variables in the datasets were the percentage of votes for the Democratic and Republican Party candidate (some third party but only in 2012 and 2016), the days in the year leading up to the election, and the number of electoral votes for the given state.

The cleaning of the datasets involved sub setting each election year so that the only quantitative variables that were examined came from the state of Virginia. Another cleaning procedure that was essential was sub setting the 2008 and 2012 datasets to number of days that corresponded to a timeframe from the first day of polling in January to the day that represented the October 19th polling day. This procedure was necessary because the 2016 dataset that we used for our prediction stopped at October 19th so in order to compare our model to previous years, it was important to make the previous election year datasets stop polling percentages at October 19th. The last step that was necessary for cleaning the datasets was implementing an algorithm that made the percentage of democratic and republican votes add up to one hundred percent. This cleaning method was necessary because the 2008 dataset did not include any third party voting percentages so in order to be consistent, we assumed the third party vote was negligible in all years.

4 Methodology

Once we had obtained the data we intended to use for the purposes of our model, we were able to use simple linear regression to build an initial model. One problem we ran into, however, was that our expected predictions were summing to less than 100%, and as we had made the assumption that third party influence was negligible this was problematic. We decided to make a correction to the data, adjusting each candidates predicted percentage to make the combined value 100%. We accomplished this goal by dividing the predicted percentage of each candidate by the total predicted percentage.

The initial model:

$$Y_i = \beta_0 + \beta_1 X_i \tag{1}$$

The fitted model for the Democratic Candidate:

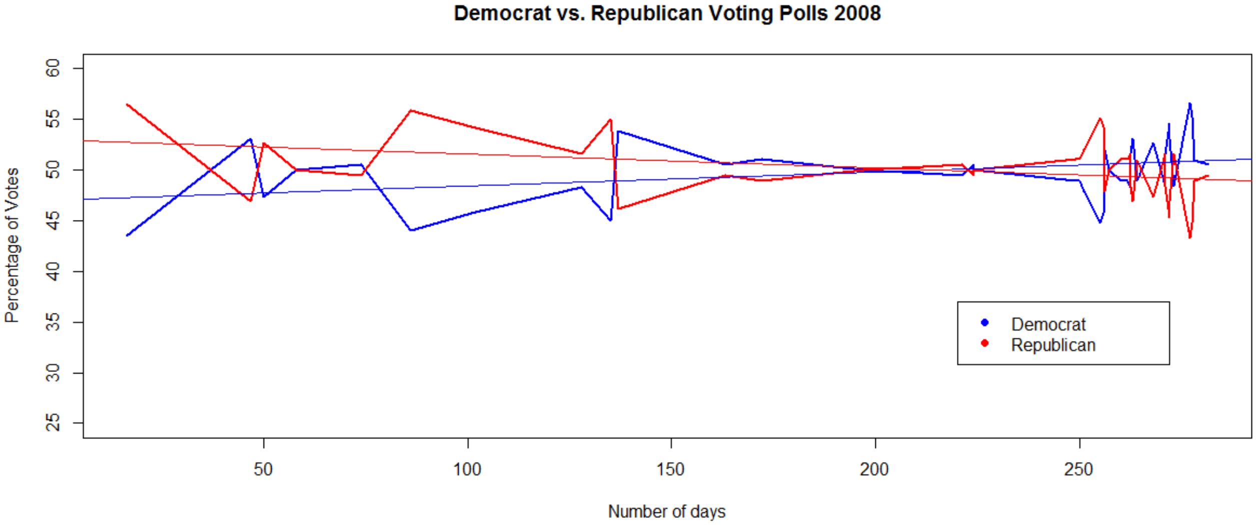
$$\hat{Y}_d = 0.00322 X_d + 53.467 \tag{2}$$

The fitted model for the Republican candidate:

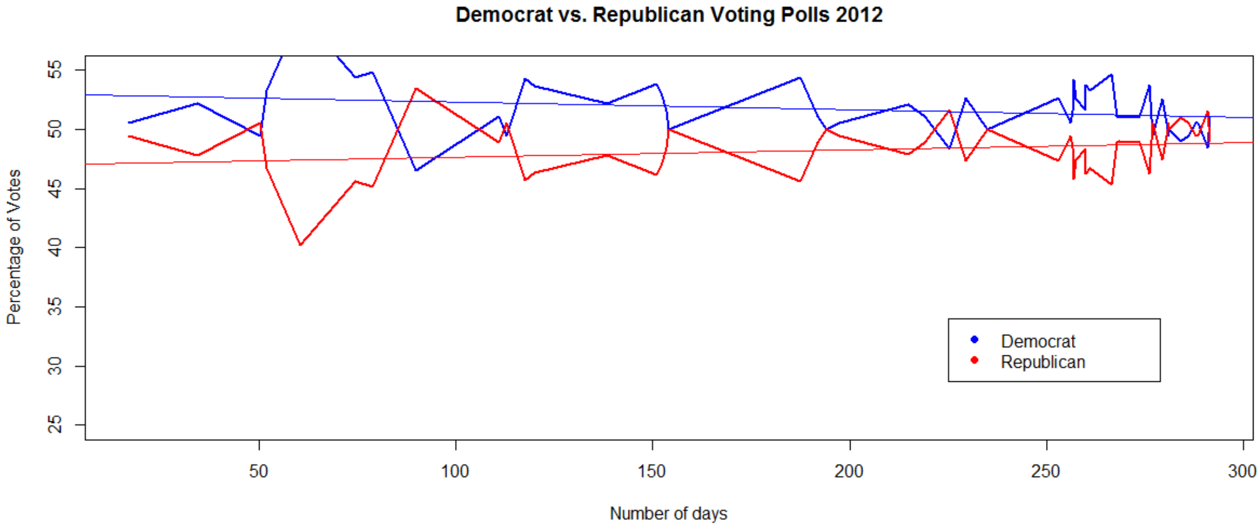
$$\hat{Y}_r = -.00322 X_r + 46.533 \tag{3}$$

Y: % Vote for Democrat or Republican
X: Number of days since the first poll

Once we had our primary model, we wanted to test its accuracy to ensure we could feel comfortable with the predictions we were making. To do this, we fed the polling data we had gathered from 2012[3] and 2008[1] into the model. Graphs of the results are below:

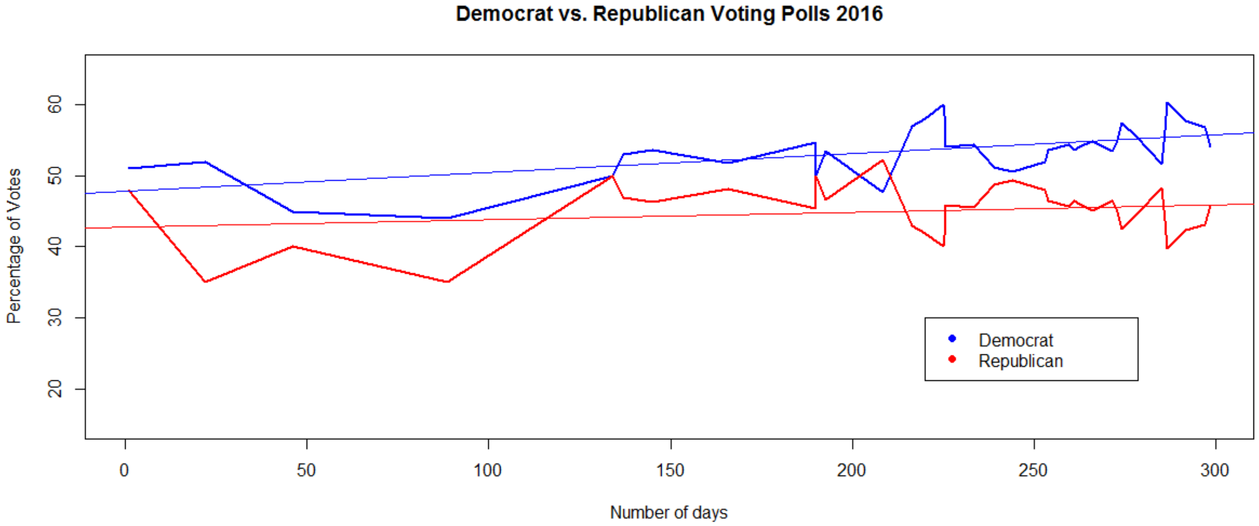


The model predicted 51.2468 % for the Democratic candidate in 2008 and 48.7523% for the Republican candidate, compared to the actual results of 52.6% for the Democratic candidate and 46.3% for the Republican candidate. In 2012, the model predicted 51.0039% to the Democratic candidate and 48.9961% to the Republican candidate, with the Democratic candidate receiving 51.2% in the election and the Republican candidate receiving 47.3% in the election. While our predictions were not exact, we were comfortable with how close they were to the actual results and therefore felt confident in using the model to predict the 2016 results.



5 Results

In the end, our final model predicted Hillary Clinton receiving 54.4% of the total vote in Virginia and Donald Trump receiving 45.6% of the total vote. The Old Dominions decided on using the value of 312 in our model for the X value due to the fact that it is election day[2].



6 Discussion

In reflecting on our model, it is important to note areas of potential bias or inaccuracy. First, our model does not take demographics into account at all, and relies only on polling data. This could be problematic if the polls are not representative or if individuals poll differently than they vote. If we wanted to improve our model, the first step we would take is to incorporate demographics as a predictor, most likely using a weighting system with polling data being weighted more heavily.

References

[1] N.A, *All presidential polls: 2008*, http://www.electoral-vote.com/evp2008/Pres/pres_polls.html, year = 2008.

[2] ——, *All presidential polls: 2016*, http://www.electoral-vote.com/evp2016/Pres/pres_polls.html.

[3] ——, *All presidential polls: 2012*, http://www.electoral-vote.com/evp2012/Pres/pres_polls.html, (2012).

[4] ——, *Historical presidential election information by state:virginia*, http://www.270towin.com/states/Virginia, (2016).