STAT 605 Group Project

Hanlin Tang(htang79),Hua tong(htong24) , Yuechuan Chen(ychen959), Yuhan Zhou(zhou453)

November 14, 2020

1 Data Set

1.1 Source and Background

The data we are going to explore is the ANES 2016 time series study and 2020 time series study(if possible) from American national election studies. The data is about the personal information and political position of the interviewed person.
We want to find a much more accurate and informative election prediction model by exploring this dataset.

1.2 Variable Description

This data file contains 1841 variables so we can't list all of them in this short proposal. However, we have made a rough plan about the variable we are going to use and will try to explore information about the following types of variables:

- 1. Internet/Radio/TV/newspaper usage;
- 2.Previous presidential choice;
- 3.Political standing(view of income gap, self-reported GOP/DEM,etc.);
- 4.Geographical politic status(state percent fpr DEM/GOP candidate,etc.);
- 5.Socioeconomic status(current income, money invested in Stock Market,etc.)

We plan to explore the data based on the variables we mentioned above. Details of these types of variables may change in during the exploration

2 Statistical Questions

- How does the silent majority affect the prediction of election? Can we provide a prediction model that correct for the silent majority?
- What correlates with people's political standing? Is it about ethnicity, socioeconomic status or other traits?

3 Research Plan

3.1 Statistical Methods

- Generally, we want to fit a generalized linear model model to the data set. Before that, we will do some correlation analysis of variables and exclude the outliers.

3.2 Computational Tools

- First, we plan to use Slurm to to grab the part of the data set we need
- Then use R to get the general characteristics of the data set.
- After that, since the data set is very large and the calculation will be huge, we want to do regression analysis through CHTC using R. If the workload is too huge for R, we may try to python to finish the rest fo analysis.
- For the model we are going to use, we would firstly try to use logistic model to explore the associations we desire to see. We may also try to use other methods like SVM to see what kinds of prediction model we can build.

4 Summary and Conclusions

As we can see, this projects may involve huge amount of pre-study of political science in the variable selecting and association exploration, however, we believe by combing the statistical methods and political science, we can explore the mechanism behind 2016 and 2020 election and try to see what can we do to improve and support our democracy.