Statistical Analysis Plan

Assessing the impact of socio-economic factors on Presidential Primary Election voting in the USA in 2016

Population

U.S.A. Population.

Primary Objective

- Find if there are specific socio-economic or demographic factors that are associates with an increased or decreased preference for a political party by county.

- Find associations between specific socio-economic or demographic factors and the fraction of people voting for a republican candidate by county.

Secondary Objectives

- Find state-wide factors that are associated with a preference for one political party over another.

- Prediction of the final state-wide outcome of the presidential election in 2016. For this an outcome by county will be needed and then grouped by state.

Data Collection

- Demographic data on counties from U.S.A. Census Bureau

- MIT Election Data and Science Lab, 2018, "County Presidential Election Returns 2000-2016", https://doi.org/10.7910/DVN/VOQCHQ, Harvard Dataverse, V2

Variables Under Consideration

Outcome variables

Could be either the data on winning party (binary variable) or the proportion of votes won (in %). Note if the proportion of votes is used the proportion of

total votes emitted may be needed as well.

Covariates

There are in total fifty explanatory variables, where: - Eighteen of them are demographic variables relating to the population and racial composition

- Two of them relate to educational attainment

- One of them correspond to the number of war veterans

- Seven relate to housing; fourteen relate to income and employment; five of them to sales

- Three of them to building permits, land area and population per square mile (density)

All of the explanatory variables are continous and most of them relate to the percentage of the population while some of them are raw counts.

Missing Data Procedures

Unlikely to be an issue. One of the states of the U.S.A. (Alaska) has different ways of getting census data and elections votes, so analysis may exclude Alaska.

Summaries to be presented

- Scatterplots showing the preference of the association for variables and the increase / decrease preference for a particular party.

- Maps (univariate/bivariate) showing the relationships of the data.

- Descriptive statistics for the variables of interest.

Models to be fitted

- Regression models may be used for the find associations for an increase or decrease preference for a particular party. However, feature selections will need to be done carefully using maybe PCA or hierarchical clustering.

- After investigation, it seems that SVM (Support Vector Machines) may be suitable for classification for the prediction objective.

Risks

- Dataset has the percentage of republican data. One of the assumptions is that there are only two parties in the election. However, this percentage would need to be re-escaled from republican / democrat if we exclude the percentage of the "other" parties. Other option is to take all the parties in mind.