Statistical analysis of US presidential election data (2)

Overall project description

In the two recent US Elections, there have been much speculation on whether various socio-economic and demographic factors, such as race, age and income levels (to name a few) played a role in the preference for political parties or candidates. Presidential elections in the USA occur every four years, with registered voters casting their ballots on Election Day, which is the first Tuesday after November 1 that year. The modern political system in the U.S. is a two-party system dominated by the Democratic Party and the Republican Party. These two parties have won every United States presidential election since 1852, alternating on a fairly regular basis. In this set of projects, you will be asked to make use of presidential election data and demographic data from the US Census Bureau to analyse potential associations between socio-economic-demographic groups and electoral results in various states and counties in the USA.

Individual project details

How many individual projects are available in this area: 2

Analysis of 2012 Presidential election data

The data set provided, `election2012.csv`, gives a number of demographic characteristics for each state (from the US Census Bureau web site, http://www.census.gov), along with the electoral outcomes in that state, for the 2012 Presidential election. The variables are listed in the following order:

- Column 1: `State` (name of State)
- Column 2: `State.ID` (2-letter ID for state)
- Column 3: `won` (which party won D- democratic; R- Republican)
- Column 4: `Sep12unempl` (Percent unemployed in September 2012)
- Column 5: `Unempl.changeJan09` (Change in percent unemployed between Jan 2009 and Sep 2012)
- Column 6: `PercPoverty` (Percent of population in poverty)
- Column 7: `UrbanPop2000` (Percent of population living in urban areas)
- Column 8: `Over65` (Percent of population aged 65 or higher)

- Column 9: `PercFemale` (Percentage of female population)
- Column 10: `High.school.or.less` (Percent who have a high school degree or less)
- Column 11: `Graduate.deg` (percent having graduate or professional degrees)
- Column 12: `No.health.insurance` (percent with no health insurance)
- Column 13: `African.American` (Percent African American or Black)
- Column 14: `Hispanic` (Percent Hispanic or Latino)

Question(s) of interest
Using this data, you will try to assess whether various demographic characteristics seem to have a possible influence on electoral outcome. In particular, the main questions of interest are:

- Are there any particular groups or clusters of states characterised by certain patterns of socio-economic and demographic factors?
- Are there specific combinations of social and/or economic factors that tend to favour either the Democratic or Republican party winning in a state?
- Which states seemed most important in decided the outcome of the 2012 election, and why?
- Can socio-economic-demographic factors alone be used to predict the electoral outcome in states? How well can the election results be predicted from these?

Relevant courses
We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models, Data Analysis, Multivariate methods or Machine Learning (main dissertation).
- Big Data Analytics, Multivariate methods or Machine Learning (advanced chapter).

Assessing the impact of socio-economic factors on Presidential primary election voting in the USA in 2016

Data available

You are provided a data set on Presidential election results for each US county in 2016 `PresElect2016R.csv` and socio-economic data from the US Census Bureau (until 2014), in the file `UScounty-facts.csv`. An additional file, `UScounty-dictionary.csv`, is provided, which lists the detailed descriptions of variables available in the county facts file. For the purposes of this analysis, you may assume that there was an election involving only two parties in each county: Republican and Democratic. A brief description of the variables in the files are listed below:

File 1: `PresElect2016R.csv`

- Column 1: `state`
- Column 2: `state.po` (2-letter state abbreviation)
- Column 3: `county` (county name)
- Column 4: `FIPS` (unique ID for county from US Census records)
- Column 5: `candidatevotesR` (number of votes cast for Republican presidential candidate)
- Column 6: `totalvotes` (total number of votes cast in the county)
- Column 7: `fracvotesR` (fraction of total votes received by the Republican Presidential candidate)
- Column 8: `partywonR` (binary variable that takes the value 1 if the Republican candidate won in that county; is otherwise zero)

File 2: `UScounty-facts.csv`

The columns of this file correspond to measurements on several variables for each county, described in `UScounty-dictionary.csv`. Variables 1-18 correspond to demographic variables relating to the population and racial composition of counties. Variables 19 and 20 correspond to educational attainment; variable 21 to the number of war veterans in the county; variables 22-28 relate to housing; variables 29-42 to income and employment; variables 43-47 to sales; and variables 48-50 to building permits, land area and population per square mile, respectively.

Question(s) of interest

The main questions of interest are twofold: first, are there any discernible associations between various socio-economic and other factors and the propensity of the county population to vote for a particular party? Second, can the relationship between various factors and primary election results by county be consolidated into a model that can forecast the actual 2016 presidential election results, by state? In particular, you may want to consider:

- Are there specific socio-economic or demographic factors that are associated with an increased or decreased preference for a political party, in a county?
- Is there an association between specific socio-economic or demographic factors and the fraction of people voting for a Republican Presidential candidate in a county?
- Are there state-wide factors that are associated with a preference for one political party over another?
- How well can your model associating socioeconomic factors with 2016 election results be used to predict the final state-wide outcome of the presidential elections in 2016? (For this question you might want to locate a data set listing the winning party in each state- this is available on numerous internet news sites, such as CNN.com or NPR.org; alternatively, you can consolidate data from within your existing data set.)

Relevant courses

We strongly recommend that you have taken the following courses to undertake this project:

- Generalised linear models, Regression Models, Data Analysis, Multivariate methods or Machine Learning (main dissertation).
- Big Data Analytics, Machine Learning or Bayesian Statistics (advanced chapter).