PollsterRegressionTutorial

Introduction

Using Cohn (2016) we will build a simplified version of their model. Afterwards, we will obtain some recent relevant data, estimate the model and discuss the choice between types of regression. For this paper, we will discuss the rationale behind using logistic regression, poisson regression and negative binomial regression for election data analysis.

Simplified Model

Polling data more often than not is based on categorical outcomes, as a result of this, logistic regression will be used to build the simplified model. We will create a simplified logistic regression model in R to model this properly.

The model outlined above utilizes age, gender, education , race and voter history to build a logistic regression model. This is a simplified model of the one built by Stanford University to determine the likeliness of a candidate winning in the 2016 election.

Data set

The data set we will be using is sourced from https://results.elections.myflorida.com/downloadresults.asp?ElectionDate=11/8/2022&DATAMODE= and tracks the general election results during the year 2022 in the state of Florida. Despite the format, we will manually convert it to a csv file so that we can read the contents fluently.

```
  ElectionDate PartyCode  PartyName RaceCode            OfficeDesc CountyCode
1    11/8/2022       REP Republican      USS United States Senator        ALA
2    11/8/2022       REP Republican      USS United States Senator        BAK
3    11/8/2022       REP Republican      USS United States Senator        BAY
4    11/8/2022       REP Republican      USS United States Senator        BRA
5    11/8/2022       REP Republican      USS United States Senator        BRE
6    11/8/2022       REP Republican      USS United States Senator        BRO
```

```
  CountyName Juris1num Juris2num Precincts PrecinctsReporting CanNameLast
1    Alachua        NA        NA        64                 64       Rubio
2      Baker        NA        NA         9                  9       Rubio
3        Bay        NA        NA        27                 27       Rubio
4   Bradford        NA        NA         0                  0       Rubio
5    Brevard        NA        NA       171                171       Rubio
6    Broward        NA        NA       355                355       Rubio
  CanNameFirst CanNameMiddle CanVotes
1        Marco                 39220
2        Marco                  9431
3        Marco                 51657
4        Marco                  8156
5        Marco                165233
6        Marco                238962
```

Estimating The Model

First, consider that the data used in the New York Times article is most likely proprietary or contains sensitive information not available to the public. The data outlined above is the closest public election data that we could find, we will tweak our original logistical regression model to fit this new data and run other regressions on.

Below, we build the models for the logistic, negative binomial and poisson regressions using the data set. We append an extra column called RepWin that counts the number of representatives for each party code. This is a binary indicator that we will use to prepare for using the logistic regression model.

Table 1: Predicted Probabilities of Democratic Winning by County

CountyName	PredictedProbabilityDemWin
Alachua	0.1175506
Baker	0.1106972
Bay	0.1214031
Bradford	0.1106972
Brevard	0.1076344
Broward	0.1283038

Discussion

The models above offer a very simplified view of the different regression techniques that can be applied to election data. Logistic regression is best used for binary outcomes, predicting

a win or lose situation. Poisson and negative binomial regressions could model aspects like number of votes. This particular example highlights that Logistic Regression models may be best to use for these types of scenarios.

However, this process also underscores the complexity involved in predicting elections, there are several different factors to account for in any prediction. Influenced by myriads of factors, elections cannot be modeled by simplistic models that will inadequately capture all determinants in an election. Due to the limited predictive power of the logistic model in this assignment, the predictor variables CountyName and OfficeDesc highly likely oversimplify the electoral dynamics in the 2022 general election of Florida. Critical evaluation of both models and data are necessary to undertake such a consequential forecast.