Project 1: US 2016 Presidential Primary Election

Chen Nan

Due date: Septermber 29, 2019

1 Introduction

This project studies predictive models for primary election in United States. The 2016 US Presidential Election data was colllected. The dataset includes a selected set of counties in united states and their demographic information. It also includes the votes received for the selected major candidates. For Democratic party, two candidates are selected: **Bernie Sanders** and **Hillary Clinton**. For Republican party, three candidates are selected: **Donald Trump**, **John Kasich**, and **Ted Cruz**.

In the datafile **"P1train.csv"**, 2498 counties are included. Each row include information for one unique county. The first column is the abbreviation of the state the county belongs to. The 2nd to 52nd columns summarize different demographic data in the county. The detailed explaination of each column is given in **"county_facts_dictionary.csv"**. The 53rd to 57th columns are raw votes for each candidate. Note that, due to differences in primary election policies, the votes for candidates in two different parties might not be directly comparable (For details, see https://en.wikipedia.org/wiki/Primary_elections_in_the_United_States). The last column is the **Response** for the project. It is calculated as the percentage of votes for Hilary Clinton in each county. Note that for this percentage, only the votes for Hilary Clinton and Bernie Sanders are relevant.

```
HilaryPercent=#Hilary Clinton/(#Hilary Clinton+#Bernie Sanders)
```

The testing dataset **"P1test.csv"** has the same data structurre for 300 counties. The last few columns (53rd to 58th) are empty (set to 0). You are expected to predict the output values (the column **HilaryPercent**) for each sample in the test dataset. The accuracies of your prediction will be evaluated based on the numbers you provided.

box cox transformation

2 Project Assignment

Your task is to build a regression model to predict the percentage of votes in primary election for each county in the testing data. For Step 1-3, **only the first 52 columns** should be used as predictors in the training data.

approch 1: choose model de modelselection avec 5 parmi 52. ou greedy search . 2nd method: scatter plot correlation level

2.1 Step 1: Simple Regression Model

In this part, you are required to develop a simple model that can be used for predicting the response. To reduce the difficulty, you are allowed only limited manipulations of the original data set. You are allowed to take power transformations of the original variables (square roots, logs, inverses, squares, etc), but you are *NOT* allowed to create interaction variables. Your model should include *NO* more than 5 predictors/covariates, but should explain as much variability as possible.

After obtaining the model with aforementioned features, you are required to diagnose and analyze the model and provide meaningful interpretations. Please focus your attention on the interpretation of the model. A strong analysis should include the interpretation of various coefficients, statistics, and plots associated with their model and the verification of any necessary assumptions.

2.2 Step 2: Full Regression Model

In this part, you are free to construct the "best" regression model for predicting the percentage. You are encouraged to experiment with any of the methods that were discussed during the semester for finding a suitable model. You are allowed to create any new variables you desire (such as quadratic, interaction, or indicator variables). Your model needs to be estimated based on the training data, and provides prediction on the testing data. Forecast errors will be evaluated as a component of your project score.

Note: You are allowed to construct multiple linear regression models to make the forecasting. Only the final forecasting results should be submitted for evaluation.

To evaluate forecasting accuracies, we will use the weighted mean squared errors to account for differences in vote numbers

$$\text{WMSE} = \frac{1}{n}\sum_{i=1}^{n} n_i(\hat{p}_i - p_i)^2, \tag{2.1}$$

where n_i is the total primary votes received for the Democratic candidates, $\hat{p}_i$ and p_i are your forecasted and true percentage of votes Hilary received.

2.3 Step 3: (Optional) Free Form Model

You can choose any arbitrary model, including but not limited to regression models, for prediction purpose. If you choose to do this part, you need to summarize the method you choose, report the

results, and compare the results with regression models in your report. The forecasting accuracy from this model will be evaluated. If your accuracy is better than that of the best regression model in the class, you will be awarded *3 bonus points*.

2.4 Step 4: (Optional) Election Prediction

After the primary election, Hilary Clinton and Donald Trump were elected as presidential candidates in each party. In the presidential election, each voter votes either Hilary Clinton or Donald Trump. Based on the primary results in the training data, can you predict the percentage of votes Hilary Clinton received during Presidential Election against Donald Trump? For students with reasonable approach and results, you will be awarded *3 bonus points*.

3 Submission

1. A report not more than 10 pages with 1.5 spacing (soft copies and hard copies), which documents the methods using, main findings, and interpretations. Codes and software printouts should NOT be included in the report.
2. Complete codes used for the analysis, with reasonable details of comments in Jupyter Notebook (Soft copies only). **Attention**: Make sure your results are replicable by the codes you submitted. Unreproducible results are considered cheating/plagiarism.
3. Forecasting results on the test dataset in a "csv" file with a single column, as shown in the following example

   ```
   A0001124H
   10.31
   8.5
   20.1
   ...
   11.5
   ```

   Only the Step 2, 3, and 4 require forecasting results submission.