Predict U.S Election Results By Logistic Regression Model

Yingyu Li   Chenxuan Ding   Jialu Xu   Yilin Wang

02/11/2020

Abstract

In this project, main topic is about prediction on U.S election results, which will be released on Tuesday, November 3, 2020. Logistic Regression model will be fitted based on survey data from [Nationscape] and then apply fitted Logistic Regression model on Post-Stratification data from [IPUMS]. At the same time, some important data properties will be presented for both survey data and Post-Stratification data. Finally, prediction will be shown that Biden will have larger probability than Trump to become the next U.S president.

Keywords: Trump, Biden, Survey, Logistic Regression Model, Model Diagnostics, Post-Stratification, Election, Age, Gender, Education, Employment, Household Income, Census Region.

1. Introduction

U.S election is one of the hottest topics recently in the world especially in North America. U.S election's results will be released on the Tuesday, November 3, 2020. Before revealing U.S election results to public, we would like to do a prediction on the U.S election. Currently, there are two president candidates have the highest votes, **Donald Trump** and **Joe Biden**. As a result, in this project, we are going to do prediction between these two candidates. In this report, we will do a deep investigation from three perspectives: *modeling data description*, *regression model selection* and *prediction*. And finally, we will also make conclusions from these three perspectives. At the same time, we will identify weaknesses of our methods and corresponding Next Steps.

2. Data Descriptions

Main target of this project is to predict election results between **Donald Trump** and **Joe Biden**. Hence, we denote **Donald Trump** by '**1**' and **Joe Biden** by '**0**' in all data sets for modeling convenience.

Survey data is collected from **Nationscape** and *post-stratification data* from **IPUMS**. Both of them need to be registered and take few days to get data sets. Full survey data sets are downloaded from **Nationscape** but only the latest data set (**June 25, 2020**) will be used. 2018 1-year ACS are selected, in order to reduce size of data set, 24 variables and 600k records are downloaded from **IPUMS**. These two data sets will be analyzed separately and shown in the results Section.

3. Model

In this big section, Logistic Regression model will be chosen to fit Survey data set and then predicting U.S election results by using Post-stratification data. Next, this section will be divided into two parts by explaining what/why Logistic Regression model, how to use Logistic Regression model.

3.1 Logistic Regression Model

Logistic Regression Model is a appropriate regression analysis to conduct when the dependent variable is binary, for example, probability of win and loss or probability of pass and fail. Linear regression can also be applied to binary dependent variable, but it is hard to interpret.

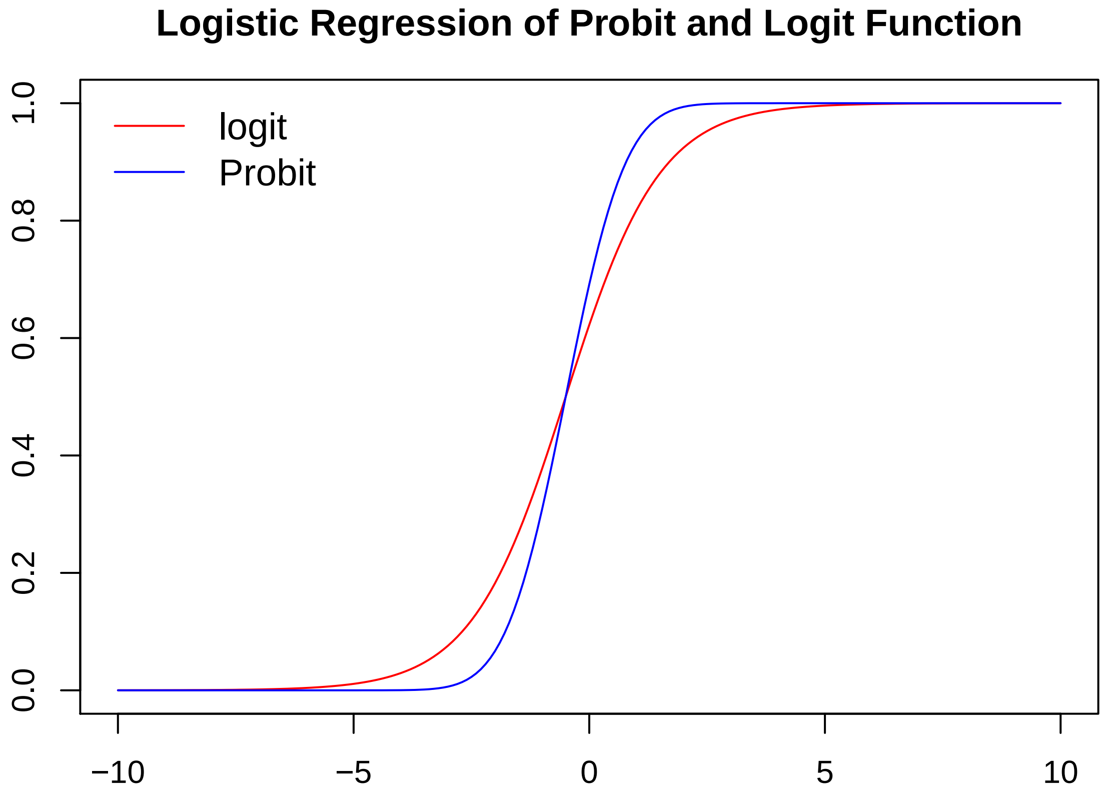
Logistic Regression can use *logit* function with the following expression, because log odds have a range $(-\infty, +\infty)$.

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \tag{1}$$

Also, Logistic Regression can use *probit* function with the following expression, because inverse of Normal Cumulative Distribution Function $(\varphi^{-1}(x))$ has a range $(-\infty, +\infty)$.

$$\varphi^{-1}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \tag{2}$$

In this project, Logistic Regression Model with Probit function will be used for prediction U.S Election results.



3.2 Modeling Survey Data Set

When fitting a Logistic Regression model, there are several items need to be considered before hand:

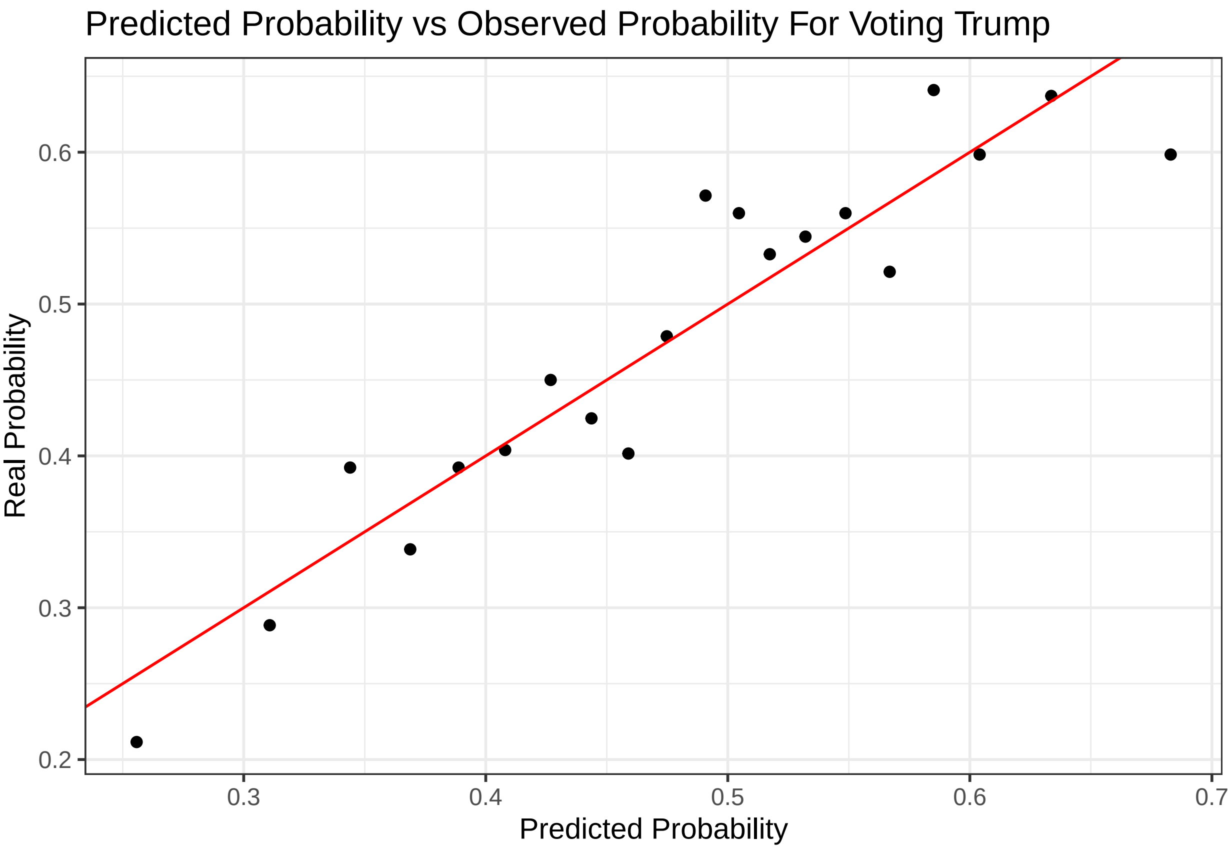
- Variable Selection: At the beginning, six predictors are selected from Survey data set for predicting election results. However, these six predictors are not necessary significant. Hence, Forward Stepwise model selection method will be used to select a model with the largest Akaike information criterion number. From the following table, one can find that all six variables are selected to fit logistic regression model.

Table 1: Attributes that selected by Akaike information criterion

Attributes
gender
age
census_region
household_income

Attributes
education
employment

- Model Accuracy Check: After fitting Logistic Regression model by using all these six predictors. The next step is doing model diagnostics by checking residual plot against fitted probability. Firstly, original survey data set is divided into 20 different groups, each group has around 250 records. Within each group, real probability that vote trump is calculated, real probability that vote trump is obtained by Logistic Regression model. From the following plot, one can observe that predicted probability is not far away from real probability, that mean the model can be used for the upcoming election results.

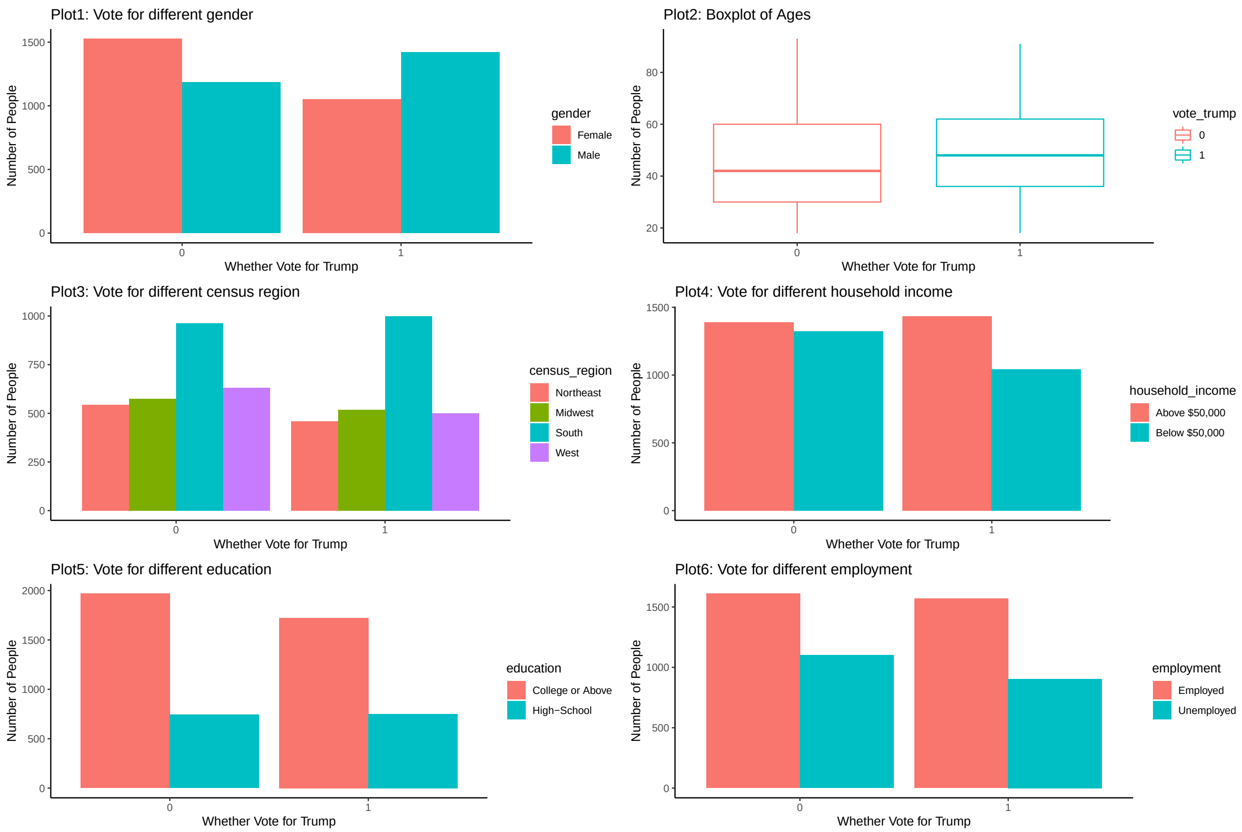


4. Results

In this section, data and model results will be shown. The first sub-section is about data analysis results for Survey data, the second sub-section is about data analysis results for Post-Stratification data, and the last section is about prediction results by using Logistic Regression model from Section 3.

4.1 Survey Data Analysis Results

For survey data, it is cleaned by *01-data_cleaning-survey.r* module, also, some multi-level variables (e.g. *employment*, *education* and etc.) are reduced into lower dimensions. There are six predictors (*gender*, *age*, *census_region*, *household_income*, *education*, *employment*) are selected for modeling in order to predict U.S election results. Below are plots regarding to these six variables against actual vote situation from survey data set.



From above plots, one can observe that:

- Plot1 is number of candidates votes by different genders. More female vote Biden rather than Trump, but more male choose to vote Trump rather than Biden. Both Biden and Trump have around 2500 votes in total.
- Plot2 is boxplot of voting ages. Range of age under 'Biden' is wider than under 'Trump' and median of age under 'Trump' is higher than under 'Biden'.

- Plot3 is number of candidates votes by different census regions. Number of votes from South area is the largest, and Trump get more votes than Biden in the south area. For other regions, Biden gets more votes than Trump.
- Plot4 is number of candidates votes by different household income level, there are lots of income levels in the original survey data set, but it is classified into two levels (Above `$50,000` and Below `$50,000`). In `Below $50,000` category, Biden has obvious advantages, in `Above $50,000` category, Trump has little advantages.
- Plot5 is number of candidates votes by different education level. People who receive college or above education give more votes to Biden.
- Plot6 is number of candidates votes by different employment status. In both 'Employed' and 'Unemployed', Biden has more votes.

4.2 Post-stratification Data Analysis Results

For Post-stratification data, it is cleaned by *01-data_cleaning-post-strat.r* module. In addition, Post-stratification data will be used for doing real prediction, as a result, its format need to be re-organized to keep consistent with Survey data set. Below are summary tables about selected predictors.

Table 2: Genders distribution in Census data

gender	Count	Proportion
Female	305472	0.5093
Male	294316	0.4907

Table 3: Education distribution in Census data

education	Count	Proportion
College or Above	303485	0.5059871
High-School	296303	0.4940129

Table 4: Region distribution in Census data

census_region	Count	Proportion
South	224844	0.3748725
West	141099	0.2352481
Midwest	128319	0.2139406
Northeast	105526	0.1759388

Table 5: Employment distribution in Census data

employment	Count	Proportion
Unemployed	305145	0.5087548
Employed	294643	0.4912452

Table 6: Household Income distribution in Census data

household_income	Count	Proportion
Below $50,000	360699	0.6013775
Above $50,000	239089	0.3986225

From Above table, one can observe some properties in the Post-Stratification data set:

- Female and Male are equally distributed in the Post-Stratification data set.
- People with College or above education and High School education are equally distributed in the Post-Stratification data set.
- About 37.5% people in Post-Stratification data set comes from South area, which is the most. And about 17.5% people in Post-Stratification data set comes from Northeast area, which is the lease.
- People with employed status and unemployed status are equally distributed in the Post-Stratification data set.
- Most people's household annual incomes (around 60%) are less than `$50,000` in the Post-Stratification data set.

4.3 Model and Prediction Results

Below is estimated parameters for fitted logistic regression model.

Table 7: Estimated Model Parameters

	fit1.coefficients
(Intercept)	-1.2349839
genderMale	0.5497878
age	0.0179216
census_regionMidwest	0.1369595
census_regionSouth	0.3382881
census_regionWest	0.0300151
household_incomeBelow $50,000	-0.2444160
educationHigh-School	0.3782508
employmentUnemployed	-0.3438327

For Post-stratification, formula to Use demographics to "extrapolate" how entire population will vote is:

$$\hat{Y^{ps}} = \frac{\sum N_j \hat{Y_j}}{\sum N_j}$$

Where $\hat{Y_j}$ is the estimate probability in the j^{th} cell and N_j is the population size of the j^{th} cell.

Table 8: Winning Probability for Trump and Biden

Candidates	Probability
Trump	46%
Biden	54%

Table 8 is winning probability for "Donald Trump" and "Joe Biden". Based on the Logistic Regression model in this project. Trump has less winning probability than Biden.

5. Discussion

In the project, the main target is to predict U.S election results by using the latest Post-stratification data set. People can find many efforts including data preparation, data analytic, model fitting and etc. Next is a summary about what this project did:

1. Register Democracy Fund + UCLA , and access *full survey data set* from there. Similarly, Register IPUMS, and access *American Community Survey (ACS) data set* after selection few interesting variables.
2. Do data cleaning and variable classification for both Survey data set and Post-stratification data set.
3. Present data analysis results and properties for both Survey data set and Post-stratification data set.
4. Select appropriate *logistic regression model* to fit Survey data set. In order to make sure *logistic regression model* is accurate, Akaike information criterion is applied to select most significant variables. In addition, comparison between real vote probability and predicted probability are made.
5. Use fitted *logistic regression model* and *Post-stratification* method to predict U.S election results.

All procedures and results are described in detail in the main body of the project. After doing all above steps, one can conclude that fitted *logistic regression model* is accurate to predict U.S and election results. Moreover, from prediction results, Trump (46%) has less winning probability than Biden (54%).

5.1 Weaknesses

However, when we are doing this project, we also identified some weaknesses that may impact U.S prediction results. Because of project words limitation, it will be notified here and next step solutions will also be provided.

- At the beginning, predictors are selected for our own interests instead of using a quantitative way. So, some important variables may impact votes may be ignored that cause less accurate prediction results.
- When fitting *logistic regression model*, model diagnostics are not enough. Since the focus of this project is not logistic regression model, When checking the accuracy of *logistic regression model*, only two items are diagnosed.
- Backtesting should be processed for logistic regression model: usually, in order to make sure model accuracy, statisticians divide *data set* into *Training data set (used for modeling)* and *testing data set (used for testing)*. But in this project, model is not tested by testing data set.

5.2 Next Steps

After identifying weaknesses of this project, next will be the list of corresponding *Next Steps* for future works.

- All variables (except identification variables) should be used to fit *logistic regression model*, and then use some statistical methods (e.g stepwise) to pick up most significant variables.
- More model diagnostics should be processed in order to make sure *logistic regression model* is a real fit in this case. If model assumptions are violated, then use other more accurate models instead.
- Divide *Survey Data Set* into *Training* and *Testing*. Training data set is used for model fitting, testing data set is used for model testing. If model predictions on testing data set have large difference than observations in testing data set, that means prediction can not be accurate when predicting Post-Stratification data set. Then we need to reconsider model.

Finally, there are no 100% accurate models in the world. Every statistical model has some biases than the real world. However, people is able to optimizing models and try to control biases in an acceptable range.

Also, based on the prediction results in this report, Biden has more chances to win the election to be the next U.S president. However, Trump and Biden probabilities have no huge differences, Trump still has big chance to win.

Reference

Wu, Thompson, Changbao. 2020. “Sampling Theory and Practice. Springer International Publishing.”

Tausanovitch, Chris, and Lynn Vavreck. 2020. “Democracy Fund + UCLA Nationscape” https://www.voterstudygroup.org/publication/nationscape-data-set.

Voter Study Group. 2020. “Democracy Fund + UCLA Nationscape User Guide.”

Steven Ruggles, Ronald Goeken, Sarah Flood, and Matthew Sobek. 2020. “IPUMS USA: Version 10.0” https://doi.org/10.18128/D010.V10.0.

Daniel Jurafsky & James H. Martin. 2019 “Speech and Language Processing, Chapter 5, Logistic Regression.”

Hamed Taherdoost. 2016 “Sampling Methods in Research Methodology, How to Choose a Sampling Technique for Research.”