Predicting the 2019 Canadian Federal Election Result if Every Citizen Had Voted

Aiting Zhang(1004926066)

21 December 2020

Abstract

This report aims to predict if every Canadian had voted in the 2019 Canadian Federal Election, how the outcome would have been different. For this investigation, I obtain the survey data set from the 2019 Canadian Election Study - Online Survey (CES), a program to record Canadians' preferences and participation in politics and the issues of social concerns. Also, I choose Education Highlight Tables, 2016 Census as population data set from Statistics Canada because this data set represents Canada's population from 2016 to 2019. I will build a multilevel logistic regression model and post-stratification based on the main variables in these two data sets: gender, age, province, education level, and vote intention. Then, using the model to predict how the election outcome would have been changed.

Keywords

Multilevel logistic regression model, post-stratification, Federal Election, Predict, Census

Introduction

There is no doubt that voting and elections are the most basic elements of democracy. In Canada, people can know the social and political life of Canada through the federal elections. Statistical analysis is very useful in the Canadian Election Study (CES). CES is a large-scale survey of citizens, and it will be conducted in every election year, and it enhances the understanding of Canadian electoral democracy. CES is the data set that records Canadians' preferences and participation in politics, as well as the issues of social concerns.

According to Statistics Canada, it states "just over three-quarters (77%) of Canadians reported voting in the 2019 federal election, unchanged from the 2015 election" (Government of Canada Reasons for Not Voting in the Federal Election, October 21, 2019). Since it was not everyone voted in the 2019 Canadian Federal Election, then this project aims to analyze how the 2019 federal election would have changed if everyone had voted. From the data sets of CES, the 2019 Canadian Election Study will be the survey data. I choose Education Highlight Tables, 2016 Census as population data set from Statistics Canada because this data set represents Canada's population from 2016 to 2019. Since the census is conducted every five years in Canada, the next population census will be conducted in 2021, and then we assume the population has no obvious changes between 2016 and 2019.

These two data sets will be used to investigate how the result of the 2019 Canadian Federal Election would be changed if 100% of Canadians voted in the 2019 federal election by creating a multilevel logistic regression model and post-stratification. In the Methodology section, I describe the data, the process of post-stratification, the model. Furthermore, I describe estimating voter intention in the Results section. Other parts regarding the model will be included in the Discussion section.

Methodology

Table 1

```
## Rows: 17,856
## Columns: 5
## $ Province   <chr> "Quebec", "Quebec", "Ontario", "British Columbia", "Onta...
## $ Age        <chr> "25 to 34", "55 to 64", "25 to 34", "55 to 64", "55 to 6...
## $ Gender     <chr> "Female", "Female", "Female", "Male", "Female", "Male", ...
## $ Education  <chr> "University or above", "College", "University or above",...
## $ Votechoice <chr> "Green Party", "Liberal Party", "ndp", "Conservative Par...
```

Table 2

```
## Rows: 520
## Columns: 5
## $ Province    <chr> "Newfoundland and Labrador", "Newfoundland and Labrador...
## $ Age         <chr> "25 to 34", "25 to 34", "25 to 34", "25 to 34", "25 to ...
## $ Gender      <chr> "Male", "Male", "Male", "Male", "Male", "Female", "Fema...
## $ Education   <chr> "No degree", "High School", "College", "Some university...
## $ Total_count <int> 2615, 6440, 7030, 645, 5800, 1970, 6345, 8910, 620, 991...
```

Data

From the data sets of CES, the 2019 Canadian Election Study will be the survey data. It records the situation of the 2019 federal election in Canada. There are 37,822 observations and 620 variables. These respondents are all Canadian citizens and live in Canada. They were invited to take this survey and answer some questions about their personal information, especially their vote intention. Then, I choose Education Highlight Tables, 2016 Census as population data set from Statistics Canada because this data set represents Canada's population from 2016 to 2019. Since the census is conducted every five years in Canada, the next population census will be conducted in 2021, and then we assume the population has no obvious changes between 2016 and 2019. In this census data, there are 252 observations and 20 variables. This census data set's each row represents a class of respondents, which means they have the same characteristics. It is mainly based on the highest education certificate that they obtained. And, according to their age, gender, residence to divide them into groups. Both of these two data sets are cleaned up because some repeated data are purposeless for analysis. For instance, I combine the variables which relate to education levels in census data into one variable. Then this education variable is able to illustrate each type of respondents' education level. Meanwhile, I only select some survey data variables to make two data sets' variables are matched with each other. I also delete the census data set's Age variable, which says "All ages, 15 plus" and "25 to 64" since these data do not have any meaning to study and are repeated. Also, some respondents did not provide specific answers, such as in the census data, they said they lived in Canada, not a specific province, and for their education level, they answered they do not know. All like these data are deleted. Table 1 and Table 2 illustrate the characteristics of survey data and census data. Table 1 shows the main variables that we selected from the raw survey data, which are Province, Age, Gender, Education, and Votechoice. Education means the highest education certificate for each citizen obtained. Votechoice is each citizen's vote intention for the 2019 federal election. Table 2 shows the main variables which we selected from the census data. These variables are the same with survey data except for the variable, Total_count. Total_count represents that a class of people, who lives in the same province, has the same gender and education level, and has a similar age.

Model

In the survey data set, I only reserve "Liberal Party" and "Conservative Party" for the Votechoice variable to build a logistic regression model. Firstly, I partition the survey and census data sets into cells based on Gender and Age. Then, the variables Gender and Age will combine into one variable. There are 8 different groups for Age and Gender. Then, I create two logistic regression models that are predicting the probability of response variable by the random coefficient and predictors in R. The aim of this project is to predict the probability of a vote intention for the Liberal Party. For the first model, model 1, (1 + Gender + Province | cell) is the random coefficient, Education and Age are the predictors. The second model, model 2, (1 + Gender + Education | cell) is the random coefficient, Province and Age are the predictors. Then, the AUC for model 1 is 0.692, which means the area under the ROC curve is 0.692, and it has the probability of 69.2% that the model 1 will predict correctly. By contrast, there is 68.9% that model 2 will predict correctly. Model 2 has a lower accuracy than model 1. Thus, I choose model1 to be the logistic regression model. The model formula is
$\log(\frac{Prob_{Liberal}}{1-Prob_{Liberal}}) = -1.26023 + a_j - 0.09440 * EducationHighSchool + 0.44828 * EducationNodegree + 0.40881 * EducationSomeuniversity + 0.50394 * EducationUniversityorabove - 0.09495 * Age35to44 - 0.26110 * Age45to54 - 0.46790 * Age55to64$

$\log \frac{Prob_{Liberal}}{1-Prob_{Liberal}}$ is log odds. $Prob_{Liberal}$ is the expected proportion of vote for Liberal Party. -1.26023 is the fixed baseline intercept. If all predictors equal 0 and random coefficient equals 0, then log odds is equal to zero. Also, a_j is the random coefficient,

$a_j = 0.022657592 * Female25to34 - 0.057049701 * Female35to44 - 0.031960567 * Female45to54 + 0.018207352 * Female55to64 + 0.074884073 * Male25to34 - 0.061177851 * Male35to44 + 0.033829570 * Male45to54 + 0.009929694 * Male55to64$

Moreover, Education and Age group are all dummy variables, which means if one of respondents meet one of these criterions, then the corresponding criterion will be equal to one, but if they do not meet one of these criterions, the corresponding criterion will be qual to 0.

Post-stratification

After that, I conduct a post-stratification to predict the proportion of votes for the Liberal Party and the Conservative Party. Post-stratification aggregates the value of cell-level by weighting every cell by its relative proportion in the population. In this project, the census data has the variable Total_count, representing the number of respondents under different personal conditions, such as their age, gender, education level, and the province they lived in. Firstly, I divide the census into 8 different cells based on respondents' gender and age. Then, applying the logistic regression model on the census data will estimate the probability of voting in every province. After that, I sum the Total_count for every cell, and then I weight each proportion estimate by the corresponding population of each cell and sum them together and divide that by the total population of each cell. This process can be presented as $\widehat{Y}^{PS} = \frac{\sum_j N_j \widehat{y_j}}{\sum_j N_j}$, N_j is the population size of the j^{th} cell, $\widehat{y_j}$ is the estimate proportion for votes in each cell.

I also use another method to predict the vote probability on the Liberal Party and the Conservative Party, applying the logistic regression model on the census data and converting each respondent's vote intention to a probability. If the probability of their vote intention is greater than 0.5, I predict they will choose the Liberal Party. However, if their probability is lower than 0.5, I predict they will choose the Conservative Party. Then, I use these probabilities to calculate total votes based on the population for each cell. Finally, I calculate the total votes for each province. If the Liberal Party's total votes in a province are higher than the Conservative Party, then I predict this province will choose the Liberal Party. Otherwise, they will

choose the Conservative Party. Thus, we can know which party will have higher amounts of votes among 13 provinces in Canada.

Result

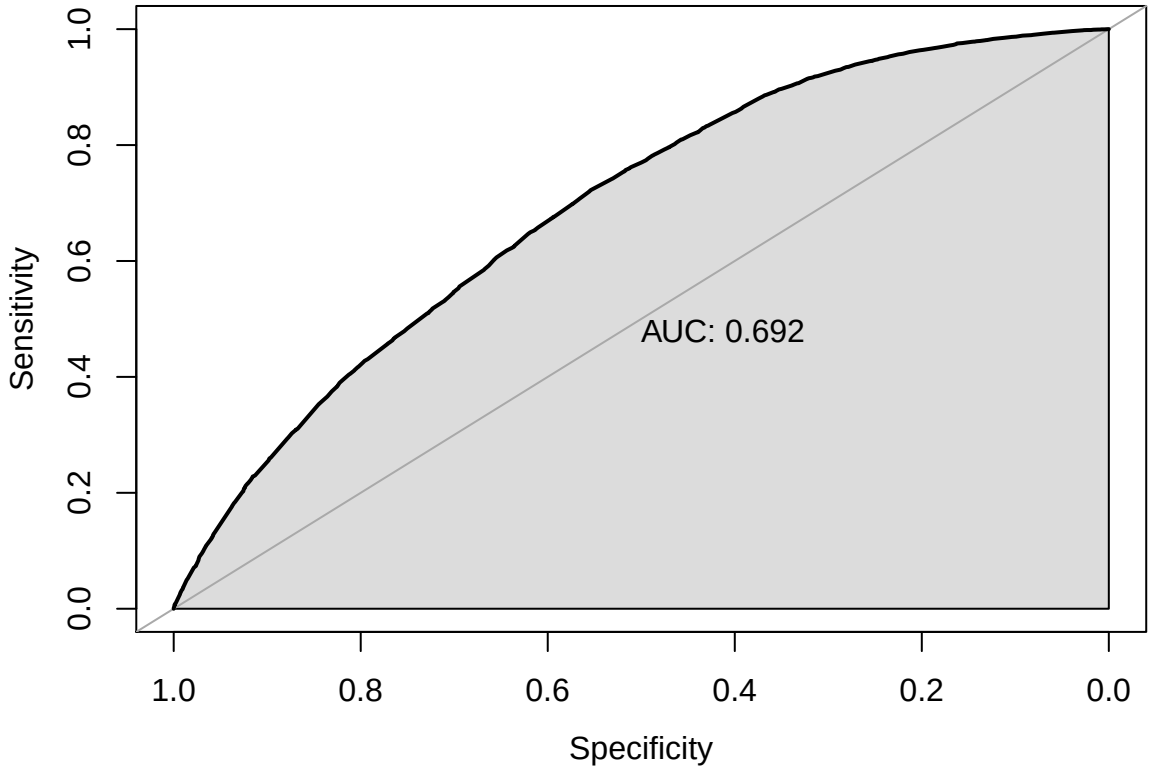

Figure 1

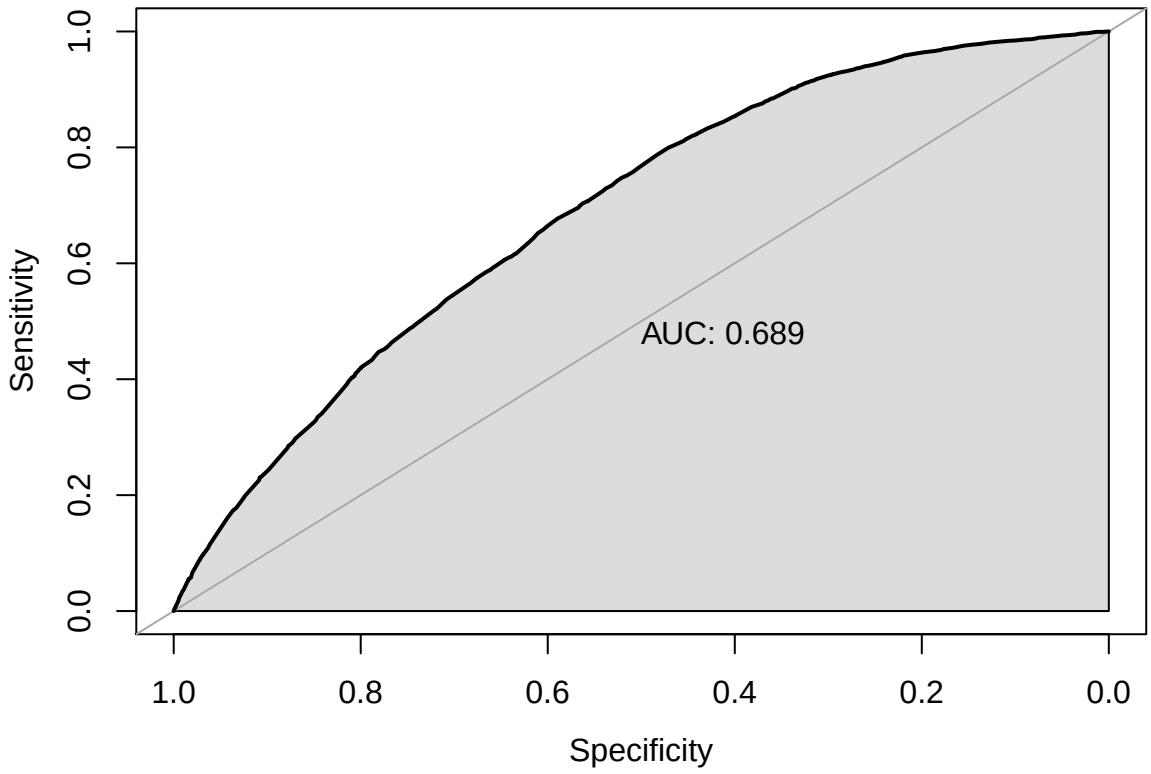

Figure 2

I build two logistic regression models to see which one will predict better. These two models are different on random coefficients. Figure 1 and Figure 2 shows these two models' area under the curve. It is obvious that Model 1 has a larger area, 0.692, which means Model 1 will have a higher accuracy on prediction than Model 2. So, I choose Model 1 to predict. When I conduct post-stratification, I divide the census into 8 different cells based on respondents' gender and age. Then, applying the logistic regression model on the census data will estimate the probability of voting in every province. After that, I sum the Total_count for every cell, and then I weight each proportion estimate by the corresponding population of each cell and sum them together and divide that by the total population of each cell. Based on the formula $\hat{Y}^{PS} = \frac{\sum_j N_j \hat{y_j}}{\sum_j N_j}$, I get $\hat{Y}^{PS} = 0.494$, which means the prediction of voting probability for the Liberal Party is 49.4%. Moreover, I also apply the logistic regression model on the census data, convert each respondent's vote intention to a probability, and calculate each province's total votes.

Table 3

```
## # A tibble: 2 x 2
## # Groups:   Winner [2]
##   Winner                n
##   <chr>             <int>
## 1 Conservative Party     5
## 2 Liberal Party          8
```

```
## [1] 0.3846154 0.6153846
```

From Table 3, among all 13 provinces in Canada, there are 5 provinces vote for the Conservative Party, and 8 provinces vote for the Liberal Party. The proportion of the Conservative Party's vote is 38.46%, and the proportion of the vote for the Liberal Party is 61.54%.

Discussion

Summary

This report aims to predict if every Canadian had voted in the 2019 Canadian Federal Election, how the outcome would have been different. Since not all Canadians voted, only 77% of Canadian citizens voted. In this investigation, I build the logistic regression model with post-stratification based on the main variables in both survey and census data: gender, age, province, education level, and vote intention to predict. I apply the post-stratification on the census data to predict the proportion of voting for the Liberal Party. The result is $\hat{Y}^{PS} = 0.494$. Besides, I apply the logistic regression model on the census data, convert each respondent's vote intention to a probability, and calculate each province's total votes. The result is that 5 provinces vote for the Conservative Party and 8 provinces vote for the Liberal Party.

Conclusion

During the process of post-stratification, it predicts the vote proportion for the Liberal Party is 49.4%, which is almost near 50%. So, we cannot predict the winner only by this method, and we do not have stronger evidence to conclude which party is the winner since the proportion is very nearly 50%. By contrast, according to Table 3 in the Results part, 5 provinces vote for the Conservative Party and 8 provinces vote for the Liberal Party. The proportion of the vote for the Liberal Party is 61.54%, and the proportion of the Conservative Party's vote is 38.46%. Also, based on the ROC curve, the accuracy for this logistic regression model is 69.2%. So, the model is feasible and reliable. We can then conclude that if every Canadian had voted in the 2019 Canadian Federal Election, the winner is still the Liberal Party. This is the same as the actual outcome in the 2019 Canadian Federal Election. Therefore, in this election, the Canadian government has not changed much.

Weakness and Next Steps

There are also some limitations to this investigation. First of all, when cleaning both survey data and census data, I dropped some repeated data and some NAs, such as if the respondents do not indicate their province specifically, or they do not show their education level correctly, I delete these data. This may cause the observation size becomes smaller, and then the accuracy of the prediction results will be influenced. Besides, when creating the logistic regression model, I only keep the Liberal Party and the Conservative Party, but there still hace other parties that I do not consider. This will also influence the prediction results. And the survey data from the 2019 Canadian Election Study - Online Survey (CES), but I choose Education Highlight Tables, 2016 Census as population data set from Statistics Canada. Since the census is conducted every five years in Canada, the next population census will be conducted in 2021. Then we only assume the population has no obvious changes between 2016 and 2019. So, the assumption is too idealistic; this population data set is a little bit outdated.

The next step for this investigation is to fix the weakness in order to make the prediction more accurately. And I can continue to find more real-time data that can represent the total Canadian population in 2019. Also, I can consider how to be cleaning the data sets again. For example, if the respondent does not indicate their province specifically, I may make these people belong to "other province", then the observation size will not decrease significantly. Then, the result of our prediction will be more accurate.

References

1. Stephenson, Laura B, et al. "2019 Canadian Election Study - Online Survey." Harvard Dataverse, Harvard Dataverse, 1 May 2020, dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FDUS88V.

2. Government of Canada, Statistics Canada. Reasons for Not Voting in the Federal Election, October 21, 2019. 26 Feb. 2020, www150.statcan.gc.ca/n1/daily-quotidien/200226/dq200226b-eng.htm.

3. Government of Canada, Statistics Canada. "Education Highlight Tables, 2016 Census." Government of Canada, Statistics Canada, 27 Nov. 2017, www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/edu-sco/index-eng.cfm.

Appendix

Code and data supporting this analysis is available at: https://github.com/aitingzhang47/STA304_Final-Project.git