Prediction of proportion of vote rusult of 2019 Canada election

Kefan Cai,1004819949

December 22, 2020

#I. Abstract

This report is aimed to predict the result of 2019 Canadian Federal Election. Data would be fron CES and GGS. A logistic regression model is build and post-stratification is used to do analysis. Age, gender, language and province are the predictors chosen in this analysis. The estimated result is that Liberal would win, which is the same result as the real election outcome.

#II.Keywords

2019 Canadian Federal Election, election prediction, logistic regression, post-stratification, Canada census.

#III.Introduction

In the current era of big data, there are many areas where statistical knowledge is used to solve problems. Statistical analysis is a useful way to make predictions, that is, to use available data and information to predict future results. For example, previous data can be used to analyze the results of the next year's election.

By learning more statistical knowledge, a interesting question that whether the result would change if different method of statistical analysis changes and what about the differences between analysis result and real outcome exists. To solve these questions, some analysis of prediction of 2019 Canadian Federal Election has been made.

A good statistical analysis requires a large amount of existing data and sampling to proceed to the next level of analysis. However, simple random sampling can skew the results significantly. So post-stratification can be used to conduct more detailed and comprehensive sampling to increase the accuracy of the results.

In this analysis, data from CES 2019 phone and GGS will be used to predict the case that how the 2019 Canadian Federal Election would have been different if 'everyone' had voted. In the Methodology section, I will do post-stratification and build a MRP model to make analysis. Results will be provided in the Results section. The conclusion and other discussion will be in the discussion section.

#IV.Methodology

#Data & Model

To build the model, age, gender, language and province have been chosen. Firstly, the age is divided into 3 stages. The reason why people who are less than 18 have not been removed is that in this analysis, an assumption that 'everyone' should vote has been made. So even though they don't have vote authority in real life, they are counted in this analysis. Language and province can show which place that citizen lives.

In this part, a Multiple logistic Regression Model will be build to take the analysis.(y) is a binary response variable of predictors(xi).p is denoted as the case when $y = 1$, represents the probability of an event occurring. $Beta0$ is the intercept(constant) and $Beta1$ to $Beta4$ represents the coefficient to each (xi).

The mathematic notation is

$$log(p/1-p) = \beta_0 + \beta_1 x_{agegroup} + \beta_2 x_{gender} + \beta_3 x_{language} + \beta_4 x_{province}$$

```
## # A tibble: 15 x 5
##    term                              estimate std.error statistic p.value
##    <chr>                                <dbl>     <dbl>     <dbl>   <dbl>
##  1 (Intercept)                        -0.219      0.627    -0.349   0.727
##  2 age_groupsenior                    -0.0586     0.353    -0.166   0.868
##  3 age_groupyouth                     -0.411      0.391    -1.05    0.293
##  4 genderMale                         -0.143      0.340    -0.420   0.674
##  5 genderTransgender                  -0.368      0.433    -0.851   0.395
##  6 languageFrench                     -0.0954     0.484    -0.197   0.844
##  7 provinceBritish Columbia            0.515      0.734     0.702   0.483
##  8 provinceManitoba                    0.401      0.740     0.542   0.588
##  9 provinceNew Brunswick               0.131      0.755     0.173   0.862
## 10 provinceNewfoundland and Labrador   0.489      0.763     0.641   0.522
## 11 provinceNova Scotia                 0.486      0.786     0.618   0.536
## 12 provinceOntario                     0.302      0.727     0.416   0.678
## 13 provincePrince Edward Island        0.399      0.755     0.529   0.597
## 14 provinceQuebec                      0.392      0.753     0.521   0.602
## 15 provinceSaskatchewan                0.344      0.774     0.444   0.657
```

#Post-stratification

Post-stratification technique is a good choice to do further analysis. To use post-stratification, the data should be cleaned at first. In this case, the CES data is the survey data and GGS data is the census data.

Firstly, for age data, they are classified into 3 groups: youth(younger than 24), adult(25~59), and senior(older than 60). The new data with 3 cells named age group was created to take place of age.

Then, the province data is classified into 13 groups. They are Newfoundland and Labrador, Prince Edward Island, Nova Scotia, New Brunswick Quebec, Ontario, Manitoba, Saskatchewan, Alberta, British Columbia, Northwest Territories, Yukon, and Nunavut.

After that, the language type is divided into 3 cells of English, French and others. Gender is classified into 2 cells, which are male and female.

After that, by multiplying each cells as $(3 * 13 * 3 * 2 = 234)$, a result of 234 cells of total can be drawn. By applying the model, the ratio can be estimated and the final proportion can be calculated.

The mathematic notation is

$$\frac{\sum N_j \hat{y}_j}{\sum N_j} = \hat{y}^{PS}$$

where $\sum N_j$ is the population size of j^{th} cell, $\hat{y}_j$ is the estimate in each cell that were constructed and $\hat{y}^{PS}$ is the estimate y.

```
## # A tibble: 1 x 1
##   alp_predict
##         <dbl>
## 1       0.622
```

#V.Result

The estimated result is that the proportion of voters in favour of voting for Liberal to be 0.622, which means that it is estimated to be the result that 62.2% people would vote for Liberal. This is based off the post-stratification analysis of modelled by a logistic linear regression model.

#VI.Discussion

#Summary In this analysis, the data from CES and GGS are used to build a logistic regression model and a post-stratification of age, gender, language and province to predict the 2019 Canada election result in the

case of 'everyone' vote. To make 'everyone' assumption, all age people includes the under-ages are considered in the analysis.

#Conclusion

The result of estimation is that Liberal would win, which is consitent to the truth. It shows that the model and post-stratification steps are mostly correct and the data is reliable to a certain extent.

#Weakness & Next Steps

One weakness is that the data was collected by phone so there may be many problems such as nobody answers the phone. That would lead to an incompletement of data. Also while doing analysis, missing value issue has been found. There are some NA in data and it may affect the correctness. Moreover, in this analysis, the only thing to make 'everyone' assumption is adding the under-ages into voters. It is worth to consider that whether it is enough to do so.

For next step, more data on online can be used to improve the data. Also, more variables can be added into the model to find the most suitable combination. In this analysis, only logistic regression model has been used. Other model can also be used to find the best one.

#VII.Reference

1. CES data Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, '2019 Canadian Election Study - Online Survey', https://doi.org/10.7910/DVN/DUS88V, Harvard Dataverse, V1 Stephenson, Laura, Allison Harrel, Daniel Rubenson and Peter Loewen. Forthcoming. 'Measuring Preferences and Behaviour in the 2019 Canadian Election Study,' Canadian Journal of Political Science. LINK: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DUS88V
2. GSS data Beaupre??, P. (2020). General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User???s Guide (Vol. 2019001). Ottawa: Authority of the Minister responsible for Statistics Canada. Retrieved October 19, 2020. General Social Survey - Family (GSS). (2019, February 06). Retrieved October 19, 2020, from https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey General Social Survey: An Overview, 2019. (2019, February 20). Retrieved October 19, 2020, from https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2019001-eng.htm

3.Kenton, W. (2020, September 21). How Multiple Linear Regression Works. Retrieved October 19, 2020, from https://www.investopedia.com/terms/m/mlr.asp