Using MRP to estimate the 2019 Canadian Federal Election result when 'everyone' had voted

Yuxin Xie

2020/12/21

Using MRP to estimate the 2019 Canadian Federal Election result when 'everyone' had voted

Yuxin Xie

21 December 2020

Github: https://github.com/Rachel-xie/final-project.git

Abstract

To estimate the 2019 Canadian Federal Election result, one can analyze a eligiable Canadian Citizen's voting preference based on age and gender. Our study is to estimate the voting preference of all Canadian citizens that have the right to vote. After generating the preferences and personal informations of some citizens by phone survey, the voting result of survey data can be investigated by buliding a multilevel regression model, grouping the selected citizens by their living province. Then apply the model to all the citizens, obtain the final estimate result by processing post-stratification.

Keywords

Canadian Federal Election, Multilevel Regression with Post-stratification, turnout, vote, 2019, estimate, province, age, sex, Liberal, Conservatives

Introduction

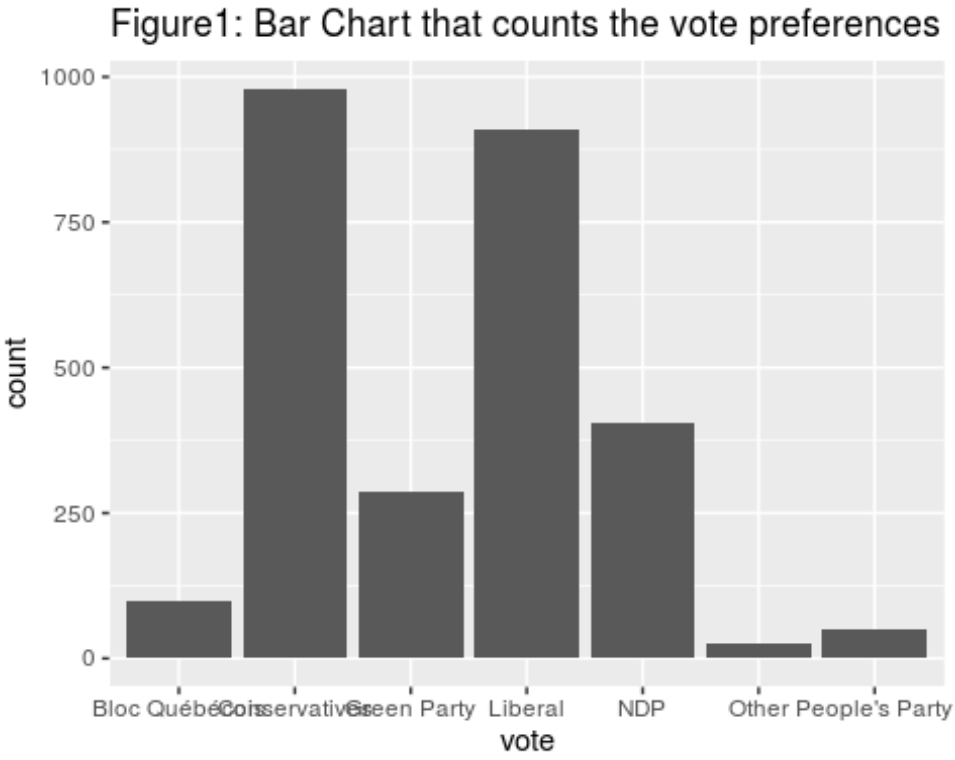
Statistical analysis is widely used in estimating the election results based on the historical data. Even after the election, we could always use statistic to estimate the election result base on the data of voters' personal information. Comparing the estimation to the reality, the government can identify which kind of factors would bring significant impacts during the election process. Therefore, the government could use the outcomes to take actions and prevent the insecure factors.

Canadian Federal Election is always a big event to all over the world, especially for Canadian citizens. However, there are still some citizens being indifferent and

renounced voting. Post-stratification is a widely used statistical technique in analysis when it is hard to generate information of the target population. It can use survey data to estimate the outcomes for the entire target population. In this report, I will apply multilevel regression with post-stratification on the estimation of the 2019 Canadian Federal Election result if all eligible voters in Canada votes.

I will use two data sets in this project, one is the survey data, another is the census data. Detailed informations about the data sets will be explained in the next section (Data). The model and post-stratification technique I used will be described in the Model section, and the results of this estimation are in the Result section. The summary and conslusions will present in the Discussion section.

Data



The survey data set is obtained from the 2019 Canadian Election Study by phone survey. I selected the questions about gender, year of born, province, and which party to vote to build the survey data set for this study. The variable age, is the age of the citizen in 2019. The variable sex, is either male or female. For the living place of the citizens in this phone survey, there are some provinces in Canada that does not exist. Therfore, this study only estimate the voting results for citizens in province Newfoundland and Labrador, Prince Edward Island, Nova Scotia, New Brunswick, Quebec, Ontario, Manitoba, Saskatchewan, Alberta, and British Columbia. And I removed the cases that do not have answer for at least one question in my survey data. Viewing the figure1, one can tell that most od the Canadian citizens are tend to vote for Liberal party or Conservatives party. Hence, I decided to only

discuss the cases that are planing to vote for these two parties. Finally, I am using a cleaned survey data set with 1889 observations of 5 variables.

The census data is obtained from Gernal Society Survey of families in 2017. Assuming there is no significant changes in population from 2017 to 2019. Selected age, sex, weight, and province from the whole data set to building census data set for this study, I get a cleaned data set with 20602 observations of 4 variables.

Model

Multilevel logistic regression can estimate the probability of a certain event to happen based on a set of predictors. Thus, I use a multilevel logistic regression model with random intercept to predict the probability of a voter that will vote for Conservatives. The model have two predictors age and sex, where age is numerical and sex is set to two levels, either male or female. All cases are grouping by provinces. The multilevel logistic regression model using in this study is

$$log(P/(1-P))_{ij} = \alpha + a_j + \beta_1 * age + \beta_2 * sexMale + \epsilon_{ij}$$

Where $log(P/(1-P))_{ij}$ represents the logarithm of the odds, where P is a probability. Hense, the probability of a voter that will vote for Conservatives P can be culculate from the log odds $log(P/(1-P))_{ij}$. α is the coefficient mean, it is the baseline of the model. a_j indicates random variable, which follow normal distribution. j indicates the provinces, is the group level. β_1 means for a voter's age increased by 1 unit, we expect to have a β_1 increase in the probability of voting for Conservatives. β_2 represents the slope of sex, which means when the sex changed from female to male, we expect a β_2 increase in the probability of voting for Conservatives. ϵ_{ij} represents the residual of the model, it is a random variable and follows normal distribution.

Results

Using the post-stratification technique,

$$\hat{y}_{ps} = \sum N_i/N * \hat{y}_i$$

where $\hat{y}_{ps}$ is the estimated proportion of voters that will vote for Conservatives party in the population, N is total population size N_i is the number of people in each provinces in the population, $\hat{y}_i$ is the proportion of each provinces in favour. The result for is $\hat{y}_{ps}$ 0.6157213, which means the Conservatives party will win the 2019 Canadian Federal Election base on this study.

Discussion

Time will be a big effect to the estimation, we cannot forcast the events that happens between the survey date and the voting day. There might have some events or

policies that make people do not want to vote for there previous preference, so the estimation result will be different.

References

-Cycle 31, General Social Survey: Families, Public Use Microdata File Documentation and User's Guide. https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf

-2017 datafile, https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/subsda3

- Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, '2019 Canadian Election Study - Phone Survey', https://doi.org/10.7910/DVN/8RHLG1, Harvard Dataverse, V1, UNF:6:eyR28qaoYlHj9qwPWZmmVQ== [fileUNF]

- 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.

-R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

-RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/

-Samantha-Jo Caetano(2020), data cleaning code (gss_cleaning.R).https://q.utoronto.ca/courses/184060/files/9827182?module_item_id=1913033

-Samantha-Jo Caetano(2020), data cleaning code (gss_dict.R).https://q.utoronto.ca/courses/184060/files/9827182?module_item_id=1913033

-Samantha-Jo Caetano(2020), data cleaning code (gss_labels.R).https://q.utoronto.ca/courses/184060/files/9827182?module_item_id=1913033

-Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

-Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686