Analysis of the Outcome of 2019 Canadian Federal Election If “Everyone” Votes

Date

December 22, 2020

Author

Haichuan Xue

Github Link

https://github.com/HaichuanXue/Analysis-of-the-Outcome-of-2019-Canadian-Federal-Election-If-Everyone-Votes.git

Abstract

According to the 2019 Canadian Federal Election, the Liberal Party wins the election by winning 157 seats against the Conservative with 121 seats and others with fewer seats. However, comparing with the total vote result, the Conservative has 6,150,177 votes, whereas the Liberal Party has 5,911,588 votes. In other words, the Conservative has more votes in total (CBCNEWS, 2019). It is of interest to investigate the election outcome if every qualified Canadian citizen votes. In this report, we will use a Multilevel Regression with a Post-stratification method to estimate the 2019 Canadian Federal Election result. It is concluded that if all the qualified Canadian citizens vote in the election, the Liberal Party could end up having more votes in total.

Keywords

Multilevel Regression with a Post-stratification, 2019 Canadian Federal Election, Election Study, Analyze past elections

Introduction

National elections are one of the most important events for a country. Generally, the winner of the election will be the president of the nation until the next election. The final choice of the president will directly impact the future development of the nation. All the qualified citizens have the right and opportunity to decide their next president among all the candidates. However, it is possible that not all qualified citizens will vote. Also, there might exist some discovered or undiscovered measurement errors. For example, during the 2020 U.S. Presidential Election, approximately 6000 uncounted votes were found in Georgia due to humans' unintentional error (BBC NEWS, 2020). Without the uncounted votes found, the vote difference between the two parties is 12284 counts. This is the result when ignoring all the measurement errors. Thus, the problem is that it is possible the final voting result is not consistent with the true result if every qualified citizen is voted, and all votes are recorded without errors.

It is usually unrealistic to have no errors during a large data collection such as a presidential election, or the final result fails to represent the entire population's will due to a number of people not voting. However, a statistical analysis can be offered to address these problems by assuming every qualified individual has voted and to predict the result based on the assumption. In this report, the 2019 Canadian Federal Election will be researched to address the problem since the competition between the Liberal Party and the Conservative Party is intense. One of the ways is to use a Multilevel Regression with Post-stratification (MRP). The brief idea is to estimate a result from a representative population through a model generated from survey data. In this report, a statistical inference via MRP is provided.

Another way is to analyze specific census data. The data should include the vote intention and any other potential covariates. This way is one of the ideal ones, and the result is expected to be very close to the true value. However, completing a census data collection usually requires a high cost and is time-consuming. Thus, this way may not be efficient for the research topic in this report.

For this report, the objective is to make a statistical inference to estimate the 2019 Canadian Federal Election result if all the qualified citizens have voted and their votes are correctly recorded. Two data sets will be used to construct this analysis via MPR. In the methodology section, the summary and modification of the data, generation and evaluation of the proposed model will be included. Then, the result section will present the analysis and evaluation results. Lastly, the discussion section contains the summary, conclusion, weakness, and the next steps of the research.

Methodology

Data

As mentioned previously, two data sets are used to complete this research. They are the Campaign Period Survey (CPS) data from the 2019 Canadian Election Study (CES19) and 2017 General Social Survey (GSS17) data. The CES data can be accessed fromhttps://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DUS88V and the GSS17 data can be accessed from http://www.chass.utoronto.ca/. The selection and weakness of the two data sets can be viewed in the discussion section. Here is a description of them.

Summary and modification of CES19 data

The CES19 data is an online survey with a targeting population of Canadian citizens and permanent residents aged 18 and over. The data is collected separately during two periods, including the campaign period from September 13th, 2019 to October 21st, 2019, and the post-election period. In the campaign period, 37822 participants are surveyed about which party they intent to vote for and other relevant information before they actually vote. For our research, we only consider the high-quality responses, which are qualified voters (i.e., citizens and aged 18 and over) from the campaign period in the clean data since all the qualified Canadian citizens are the targeted population. The high-quality responses refer to the responses that are unique and finished within a reasonable time from 8.3 mins to 60 mins. The fast, over-slow, or duplicated responses may cause measurement errors or answering biases to reduce the data quality.

In the final clean data without missing values, it contains 4881 observations with seven covariates and one the response variable. The response variable is *vote Liberal*, which is a binary variable (Yes or No). The vote intention has more factors in the original data. However, a binary response variable is required for our model (see Model section). In addition, the Liberal Party is the winner of the election, and it is more critical to estimate whether the Liberal Party still has a higher chance to win if satisfying our assumption. The seven covariates are summarized as the following,

1. Demographic covariates:
   these covariates can identify an individual's social status which is expected to impact whether they will vote for the Liberal or not.
   a) Sex: Male or Female
   b) Age: 18 – 29, 30 – 39, 40 – 49, 50 – 59, 60 – 69, 70 – 79, 80+
   c) Education: Grade 12 and under, Complete high school, Complete college/associate degree, Some university or bachelor's degree, Above bachelor's degree
   d) Family income: level 1 (1~60,000), level 2 (60,001~110,000), level 3 (more than 110,000)
2. Geographic covariates: these covariates are selected to distinguish the culture or custom difference among people which is expected to affect an individual's political point of view.
   a) Birth in Canada: Yes or No
   b) Province: Ontario, Saskatchewan, Quebec, British Columbia, Manitoba, Nova Scotia, Alberta, Prince Edward Island, Newfoundland and Labrador, New Brunswick
3. Cells: a combination information of province, age, education, and family income of each individual

*Cells are designed for Post-stratification (see Model section for more)

*The factor decisions for each demographic and geographic covariate are based on the purpose of matching the factors in two data sets.

Summary and modification of GSS17 data

The GSS17 data collect information about social living conditions in various aspects from 20602 participants from February 2nd, 2017 to November 30th, 2017. Its target population is all the people aged 15 and over in Canada, excluding people who are residents of institutions full-time or who are from Yukon, Northwest Territories, and Nunavut. Similarly, we only consider the information from the individuals who are Canadian citizens and aged 18 and over.

In the final clean data without missing values, it contains 18743 observations with seven covariates. After the modification from the original GSS17 data, the seven covariates are the same as the seven covariates from the final clean data of CES19. The factor levels in each covariate are the same as well to satisfy the condition of using MRP.

Model and Evaluation

We propose an MRP model for our statistical analysis in R software. MRP is a technique that can justify the difference between the target population and the study population when estimating results. In our case, the study population is the people from the CSP clean data since the data is used to build a specific model so that we can use the model onto the target population to get a better estimation. The target population is the people from the GSS clean data, which is the post-stratification data assuming all the qualified Canadian citizens. The multilevel model built to estimate the odds of voting for the Liberal Party is a logistic linear regression model shown as the following,

$$Y_i \sim Binomial(N_i, P_i)$$

$$log\left(\frac{P_i}{1-P_i}\right) = \mu + X_i\ \beta_i$$

Where:

- Y_i represents the result of whether an individual will vote for the Liberal Party. If yes, then $Y_i = 1$. Otherwise, $Y_i = 0$.
- N_i is the population size and P_i is the probability of voting for the Liberal Party.
- μ is the intercept, which is the baseline, referring to the odds of voting for the Liberal Party for a female citizen who is aged 18-29, was born outside of Canada, earns a degree above Bachelor's, lives in Alberta with a level 1 condition of family income.
- X_i is the vector of covariates, including age, sex, education, family income, birth in Canada, and province.
- β's are the coefficients of each covariate to show its impact.

The key part of MRP is the creation of cells. The cells are used to partition the population into each call/group so that we will have informative knowledge based on the study population. Given the post-stratification data, the probability of voting for the Liberal Party y will be estimated for each cell via multilevel modeling. The statistic value $\hat{y}^{ps} = \frac{\sum N_i \hat{y}_i}{\sum N_i}$ is estimated to represent how the target population will vote in proportion, where $\hat{y}_i$ is the estimate for each cell and N_i is the population size for i^{th} cell.

The model evaluation is based on the significance of β's for each covariate and the area under the curve (AUC) - receiver operating characteristics (ROC) curve. An estimated β with a p-value smaller than 0.05 is considered as a significant covariate associating with the response variable. A high value of AUC indicates that the model has a good discriminate ability to distinguish the results between voting for the Liberal and not voting for the Liberal.

Results

We estimate the proportion of voters who are likely to vote for the Liberal Party to be 0.86, which is $\hat{y}^{ps} = 0.86$. This estimation is calculated through the method of MRP given a logistic linear regression model that contains the covariates of sex, age, education, birth in Canada, province, and family income.

Model Evaluation

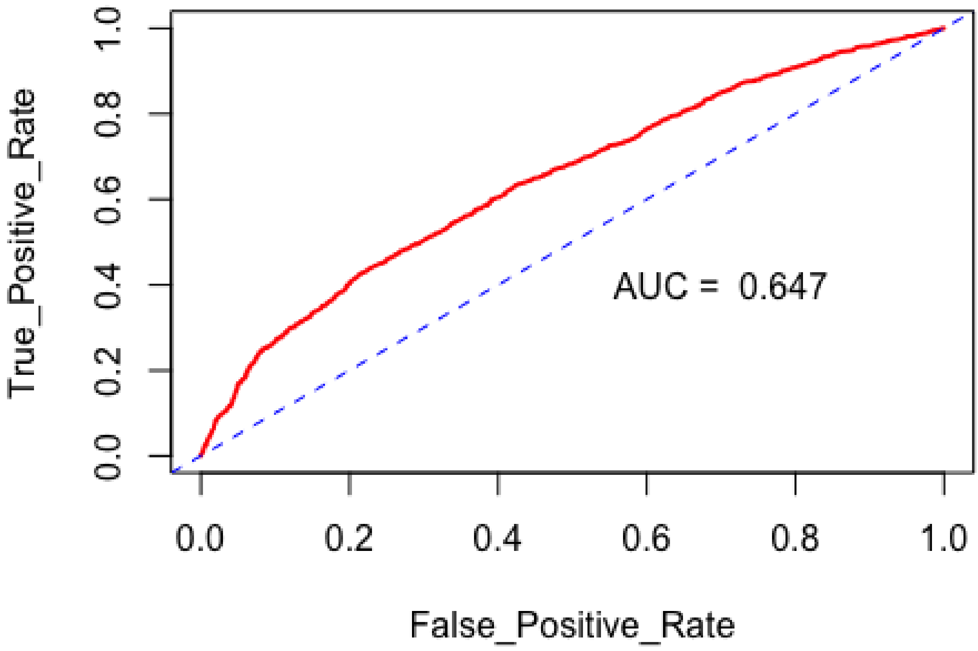
Table 1 shows the summary of the proposed model that can estimate the odds of voting for the Liberal Party. The intercept is the baseline with a coefficient of 1.534. It means that the odds of voting for the Liberal for a female citizen, who is aged 18-29, was born outside of Canada, earns a degree above Bachelor's, lives in Alberta with a level 1 condition of family income, is 1.534. The odds difference between a female who is aged 60-69 and another female aged 18-29 is -0.348 (see Age60-69 in Table 1) when other conditions of covariates are the same. In other words, a 60-69 aged woman is less likely to vote for the Liberal comparing with a woman who is 18-29 years old under the same conditions. Similar interpretations can be made for any other estimated coefficients. From the table, we can see that the sex difference (see SexMale in Table 1) is not a significant covariate to estimate the odds of voting for the Liberal since its p-value of 0.12 is larger than 0.05. For a similar reason, we can find that *birth in Canada* is not a significant covariate, and there exit some factor levels which are not significant in age, education, province, and family income.

Table 1. Summary of the multi-level Model

	Estimated β	**P value**		**Estimated β**	**P value**
Intercept	1.534	< 0.05			
SexMale	0.102	0.12	educationBachelor&under(uni)	-0.183	0.11
Age30-39	-0.04	0.97	educationCollege/Associate Degree	0.152	0.18
Age40-49	0.023	0.83	educationGrade12&under	0.323	< 0.05
Age50-59	-0.161	0.11	educationHighschool	0.258	< 0.05
Age60-69	-0.348	< 0.05	birth_in_canadaYes	0.084	0.32
Age70-79	-0.499	< 0.05	provinceBritish Columbia	-0.755	< 0.05
Age80+	-0.289	0.22	provinceManitoba	-0.489	< 0.05
			provinceNew Brunswick	-1.022	< 0.05
			provinceNewfoundland and Labrador	-1.346	< 0.05
			provinceNova Scotia	-1.224	< 0.05
			provinceOntario	-1.104	< 0.05
			provincePrince Edward Island	-0.943	< 0.05
			provinceQuebec	-1.470	< 0.05
			provinceSaskatchewan	0.419	0.09
			income_familylevel 2	-0.010	0.89
			income_familylevel 3	-0.257	< 0.05

Figure 1 is the plot of AUC for the model. The y-axis refers to the rate of truth-positive, meaning the model predicts an individual will vote for the Liberal and the individual actually wants to vote for the Liberal. The x-axis is the rate of false-positive, meaning the model predicts an individual will not vote for the Liberal, but the individual actually wants to vote for the Liberal. We estimate the AUC value of the model, which is the area under the red curve. It equals 0.647. The dotted line is s helper line that shows an AUC value at 0.5 for the purpose of comparison. The interpretation of the AUC value can be viewed in the weakness of the model section.

Figure 1. AUC of the model



Discussion

Summary

To estimate a result of the 2019 Canadian Federal Election if every qualified citizen votes, we carry out a statistical inference based on the method of Multilevel Regression Post-stratification. The method requires two datasets, including sample data and post-stratification data, and partition the population into cells. In this analysis, the Campaign Period Survey data from the 2019 Canadian Election Study is used as the sample data to build a logistic regression model to estimate the odds of voting for the Liberal Party with the covariates of age, sex, education, family income, birth in Canada, and province. The 2017 General Social Survey is used as the post-stratification data to estimate the proportion of voters in favor of voting for the Liberal Party based on the model prediction result for each cell-level.

Conclusion

From the result section, we estimate the proportion of voters who are likely to vote for the Liberal Party is to be 0.86. It means that based on our statistical inference, we expect that 86% of the entire population, referring to all the qualified Canadian citizens, will vote for the Liberal Party in the 2019 Canadian Federal Election. In reality, the Liberal Party has 33.1% of the total votes (CBCNEWS, 2019). Thus, it is expected that if every qualified citizen votes in the election, the Liberal Party should have more than 33.1% of the total votes.

This finding supports us to conclude that the result is possible to be alternated if all the votes from the entire qualified voter population are counted in an election. Thus, encouraging all the qualified people to participate in the election is essential so that the final result can truly represent the entire population.

Weakness

Weakness of GSS17 data

The drawback of using GSS17 data as the post-stratification data can be that this data is not an ideal data to represent the entire population of qualified Canadian citizens in 2019 for two reasons. The first aspect is that GSS17 data is collected in 2017 instead of 2019. Thus, the information might be old for analyzing the 2019 Canadian Federal Election. Another aspect is that the GSS17 data does not have any observations of individuals from Yukon, Northwest Territories, and Nunavut. Qualified citizens in those places in Canada are exclusive in our analysis.

Weakness of model

The drawbacks of the proposed model are some insignificant covariates and a median-quality value of AUC. The covariates of *sex* and *birth in Canada* are not significantly associated with the odds of voting for the Liberal Party. They will reduce the prediction accuracy of the model. We keep them for study purposes. The AUC value of the model is 0.65. It means that the model can distinguish between voting for the Liberal and not voting for the Liberal for an individual at 65% of the time. Usually, the higher the AUC value is, the better the model prediction accuracy will be. If the AUC value is higher, we are more confident about our conclusion.

Next step

It is suggested to use better post-stratification data such as Canadian census data. At present, there is only the 2016 Canadian census data available. However, the 2016 Canadian census data may not suitable since it is old data. The next Canadian census data will be collected in May 2021. It is better to use the 2021 version. Another suggestion could be that propose a better model by finding other more significant covariates.

Reference

BBC NEWS. (Nov. 20, 2020). Retrieved from https://www.bbc.com/news/election-us-2020-55006188

CBCNEWS. (2019). Retrieved from https://newsinteractives.cbc.ca/elections/federal/2019/results/

Hadley Wickham and Evan Miller (2020). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 2.3.1. https://CRAN.R-project.org/package=haven

Sam Firke (2020). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.0.1. https://CRAN.R-project.org/package=janitor

Statistics Canada. (n.d.). Retrieved from https://census.gc.ca/index-eng.htm

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

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77 <http://www.biomedcentral.com/1471-2105/12/77/>