Predicting the 2019 Canadian Federal Election by MRP

Jiayang Wang (1003790935)

December 21, 2020

Abstract

The Liberal party won the 2019 Canadian federal election with 6,018,728 votes of total 18,170,880 valid votes. This paper focuses on exploring the 2019 Canadian federal election results assuming the whole Canadian population will vote as valid votes. By the datasets from Canadian Election Study (CES) and the General Social Survey (GSS), to predict the overall popular votes of the 2019 Canadian federal election, using a regression model with post-stratification (MRP) which is a statistical method to predict population by sample data, and then compare the results of the prediction with the true results of 2019 election to find out the performance of the model and study the importance of difference in results.

Key words

Canadian federal election; Logistic regression; Post-stratification; Canadian election study; General social survey

Introduction

The Liberal party won the 2019 Canadian federal election with 33.12% of votes of total 18,170,880 valid votes which is approximately only half of the total Canadian population, according to Elections Canada (2019). The current first-past-the-post (FPTP) election method is convenient in execution and statistics, but it is often criticized for its unrepresentative, and may consequently discourage response rates (77.87% for the 2019 election), leading to investigation of the results of the 2019 Canadian federal election if the total population will vote[1].

One popular method for predicting the federal election results by sample data is multilevel regression with post-stratification (MRP) which can adjust non-representative data to better analyze opinion and other survey responses and predict population. In this report, only voting results for the Liberal party or Conservative party will be considered, and since the CES data is randomly sampled and it is representative (CES survey codebook, 2019), a logistic regression model with post-stratification will be conducted.

The datasets are obtained from the Canadian Election Study (CES) and the General Social Survey (GSS), which will be demonstrated in the Data section. Further narrowing down the scope to the variables that are useful, a demographical analysis on CES data is developed to see which variables are correlated with voting intention, and fit them into a logistic regression model to estimate the probability of voting intention existing by explanatory variables in Model section. To further estimate the voting proportion, a post-stratification analysis is conducted based on the results of the logistic regression in the Post-Stratification section. Results of predicting the federal election are provided in the Results section. How the 2019 Canadian federal election can be different if the total population voted will be discussed in the Discussion section as well as the conclusion, and analyze the weakness of the data and model, and discuss what can be improved in the next step.

Methodology

In this section, we analyze the datasets, fit a logistic regression model, and conduct a post-stratification analysis.

Data

Our survey data is obtained from the Canadian Election Study (CES), a dataset that focuses on Canadians' political behaviour and attitudes, measuring preferences on key political issues, in our study, is the 2019 Canadian federal election. The CES data is collected by a stratified random sampling method(STRS)[2] by region via the online survey, and the population is all Canadian citizens and permanent residents aged 18 or older. The CES survey is consists of the campaign period survey(CPS) and post-election survey (PES), where the CPS survey is used as our survey data since it contains more observations and voting intention. The CES survey has strengths in its data quality because incomplete responses, duplicate responses, and other low-quality data have all been removed from the data. However, the CES survey balances the percentage of gender and age within each region, which might be non-representative; therefore, we create agegroup based on age and use agegroup as the variable instead of age to reduce the influence of potential non-representative, and our survey data is obtained by cleaning and summary analysis from the CES data, including 17230 observations and 5 variables:

Voting intention: voting preference of the respondent, variable of interest

Sex: sex of the respondent, predictor

Province: province of residence of the respondent, predictor & cell

Agegroup: age group of the respondent at the time of the survey interview, predictor

Education: education background of the respondent with the highest certificate diploma or degree, predictor

Table.1.

	Conservative Party (N=8492)	Liberal Party (N=8738)	Overall (N=17230)
factor(province)			
Alberta	1981 (23.3%)	552 (6.3%)	2533 (14.7%)
British Columbia	909 (10.7%)	910 (10.4%)	1819 (10.6%)
Manitoba	482 (5.7%)	342 (3.9%)	824 (4.8%)
New Brunswick	162 (1.9%)	221 (2.5%)	383 (2.2%)
Newfoundland and Labrador	103 (1.2%)	207 (2.4%)	310 (1.8%)
Nova Scotia	148 (1.7%)	318 (3.6%)	466 (2.7%)
Ontario	3152 (37.1%)	4002 (45.8%)	7154 (41.5%)
Prince Edward Island	20 (0.2%)	42 (0.5%)	62 (0.4%)
Quebec	1020 (12.0%)	2015 (23.1%)	3035 (17.6%)
Saskatchewan	515 (6.1%)	129 (1.5%)	644 (3.7%)
agegroup			
20 or less	148 (1.7%)	272 (3.1%)	420 (2.4%)
21 to 35	1575 (18.5%)	1854 (21.2%)	3429 (19.9%)
35 to 50	2178 (25.6%)	2221 (25.4%)	4399 (25.5%)

	Conservative Party (N=8492)	Liberal Party (N=8738)	Overall (N=17230)
50 to 65	2802 (33.0%)	2611 (29.9%)	5413 (31.4%)
65 to 80	1789 (21.1%)	1780 (20.4%)	3569 (20.7%)
sex			
Female	4325 (50.9%)	5058 (57.9%)	9383 (54.5%)
Male	4167 (49.1%)	3680 (42.1%)	7847 (45.5%)
education			
Bachelor's degree (e.g. B.A., B.Sc., LL.B.)	1954 (23.0%)	2566 (29.4%)	4520 (26.2%)
College, CEGEP or other non-university certificate or di...	3090 (36.4%)	2513 (28.8%)	5603 (32.5%)
High school diploma or a high school equivalency certificate	1448 (17.1%)	1066 (12.2%)	2514 (14.6%)
Less than high school diploma or its equivalent	417 (4.9%)	324 (3.7%)	741 (4.3%)
Trade certificate or diploma	318 (3.7%)	418 (4.8%)	736 (4.3%)
University certificate or diploma below the bachelor's level	718 (8.5%)	904 (10.3%)	1622 (9.4%)
University certificate, diploma or degree above the bach...	547 (6.4%)	947 (10.8%)	1494 (8.7%)

Table 1 is the baseline characteristics table of the CES data that shows the number and percentage of voting intention for people in different province, agegroup, sex, and education. We see that some province has strong voting preferences for a specific political party, for example, people in Alberta tend to vote for the Conservative party, while other variables have less preferences, and the total voting intention is similar.

Our census data is obtained from the General Social Survey (GSS), which is designed as an independent, annual, and cross-sectional sample survey, containing several survey contents for each respondent, such as entry component, conjugal history, fertility intentions, etc (GSS user guide, 2017). A stratified random sampling design(STRS) is carried out for sampling, dividing the population into 27 groups known as strata by geographic areas, and then randomly sampled by telephone interviews within each strata. The target population is all non-institutionalized people above 15 years old living in Canada, and the sampling frame combines both telephone numbers and the Address Register[3](AR) to ensure good coverage of all households. Then the data is collected by telephone interviews, and for those non-response or refused, more calls will be made to contact or explain the importance of the survey as well as encourage participation. The STRS is an appropriate method in this case because the strata are formed based on respondents' shared attributes, geographic areas, including not only 10 provinces but 17 Metropolitan areas, and also it involves the random sampling from the target population, so it is random enough that each sample is equally likely to be selected, and the sample population can be the best representative of the target population. However, this method can be time-consuming and high-cost because of the large sample size and survey contents. The GSS dataset contains 20602 observations and 461 variables, which is good for developing many different aspects of research due to its large quantity of attributes, and we can use the GSS dataset as the census data that used for predicting the federal election results taking the advantage of its large sample size. However, it is not clean enough, resulting in difficult code reading. When cleaning the data, we filter people who are eligible to vote that age is greater than 18 and citizenship status is whether by birth or by naturalization. Therefore, our census data is obtained by cleaning the 2017 GSS data, including 4 useful variables:

Sex: sex of the respondent, predictor

Province: province of residence of the respondent, predictor & cell

Agegroup: age group of the respondent at the time of the survey interview, predictor

Education: education background of the respondent with the highest certificate diploma or degree, predictor

When mapping our survey and census data, in order to conduct MRP, we rename the variables of sex and education; we filter out age greater than 80 in the CES data since the GSS data only contains people with age less than 80; we create age groups in both dataset; we filter people only with sex male or female in the CES data since the GSS data only contains people with sex male or female; we summarise the 11 education levels into 7 levels in the CES data to map GSS data; and we remove province of Northwest Territories, Yukon and Nunavut to map GSS data.

Model

MRP uses a regression model to relate individual-level survey responses to various characteristics and then rebuilds the sample to better match the population. In this way, MRP can not only allow a better understanding of responses, but also allow us to analyze data that may otherwise be unusable.

We firstly build a logistic regression model that is used to model the probability of response variable existing by explanatory variables. In our case, we want to predict the probability of a voter voting for the Liberal party by other predictors, including agegroup, sex education and province in our survey data, and we build a logistic regression model in R with the following formula:

$$\log(\frac{\hat{Liberal}}{1-\hat{Liberal}}) = - 0.05913 - 0.68057\mathbf{Age21to35} + ... - 0.81500\mathbf{Age65to80} - 0.36286\mathbf{SexMale} - 0.46933\mathbf{EduCollege} + ... + 0.25367\mathbf{EduAboveBachelor} + 1.30788\mathbf{ProvinceBC} + ... - 0.10837\mathbf{ProvinceSask}$$

Where $\hat{\mathbf{Liberal}}$ is the expected proportion of vote for the Liberal party. - 0.05913 is the intercepts which means when all other predictors are 0, the log-odds is - 0.05913, where we define $\log(\frac{\hat{Liberal}}{1-\hat{Liberal}})$ is the log-odds, and then we can calculate the probability of voting for Liberal party by log-odds. Besides, $\mathbf{SexMale}$ is a dummy variable of sex, where female is the baseline for variables of sex, which means when other predictors are unchanged (i.e. agegroup, province, and sex) and the sex of the voter changed from female to male, the log-odds will decrease by 0.36286. Similarly, agegroup, education and province are all dummy variables that can be interpreted as when other predictors are unchanged, if one of the variable changes from 0 to 1, it means the voter changes from the baseline to the corresponding variable with corresponding coefficients, and we expect the log-odds changes by that coefficient.

We use agegroup, sex, education and province as a predictor for voting probability because we believe they have influence on voting proportion; When testing whether the coefficients of predictors of agegroup and sex equals to 0, the p-values of these two predictors are all less than the significance level of 5%, which means they are statistically different from 0, so that they are good predictors. When testing whether the coefficients of variables of education equal to 0, the p-value of education level of trade certificate and level of below bachelors are greater than the significance level of 5%, which means these two education level are not statistically different from 0, so they cannot predict the probability of voting for the Liberal party well; however, the p-values of other education level are less than the significance level of 5%, so we still consider education as a good predictor. Similar with predictor of province, the p-value of province Saskatchewan is greater than the significance level of 5%, while other provinces have p-values less than the significance level of 5%, therefore, province is considered as a good predictor. Additionally, we build an alternative logistic regression model without province in order to check model overfitting[4], and we find that the AIC of model with province is 1635 less than the alternative model, where a smaller AIC value means better model performance; therefore, the model with province as predictor is considered as a better model.

Post-Stratification

We continue our study by conducting a post-stratification analysis to estimate the proportion of voting for the Liberal party, where post-stratification aggregates cell-level value by weighting each cell by its relative proportion in the population, the sampling weights are adjusted so that they sum to the population sizes within each post-stratum. This calculation results in removing bias because of nonresponse and underrepresented groups in the population.

In our report, we use province as cell. We firstly divide census data into 10 cells by province, and then we apply our logistic regression model on census data to predict the probability of the vote in the census data in each cell by predictors of agegroup, sex, education and province.

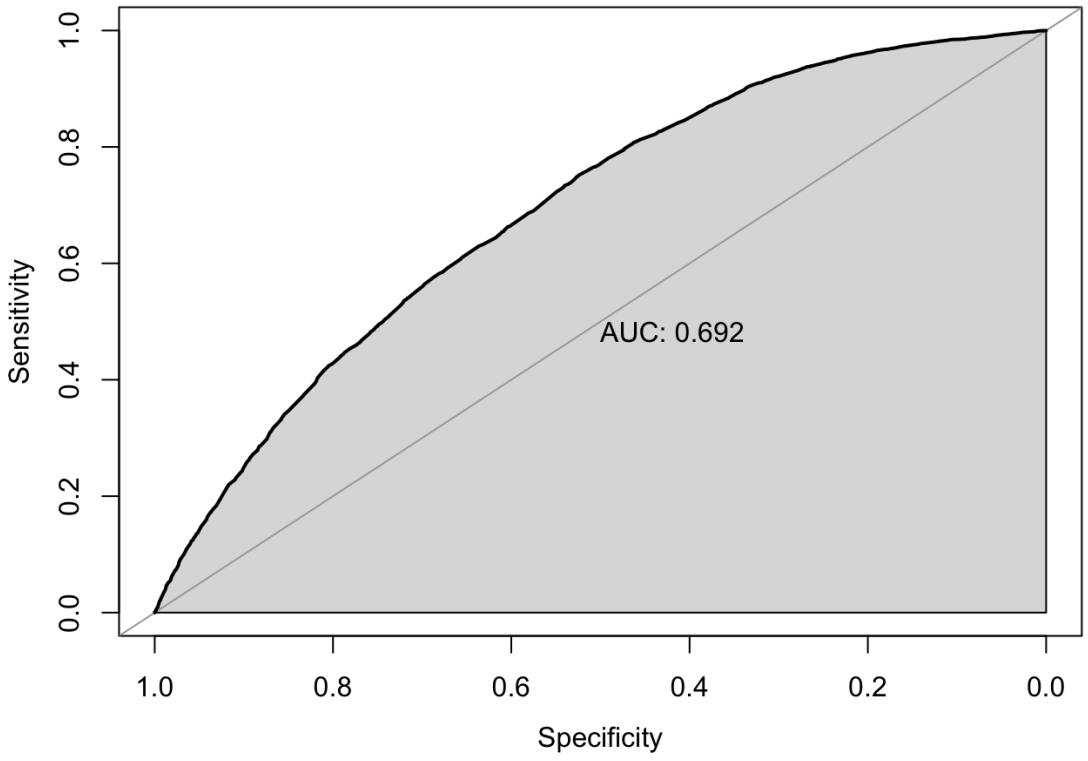
Then we weight each proportion estimate by the corresponding population of that province and sum them together and divide that by the total population. This process can be denoted as $\hat{Y}^{PS}$, which also can be shown in formula: $\hat{Y}^{PS} = \frac{\sum_j N_j \hat{y_j}}{\sum_j N_j}$, where N_j is the population for the cell, and $\hat{y_j}$ is the estimated weight of the vote proportion for that cell.

Besides, we calculate the total estimate popular votes by applying our logistic regression model on census data to predict the probability of the vote in the census data in each province, then we convert this probability into a prediction of a specific candidate, when the probability is greater or equal to 0.5, it is considered to vote for the Liberal party, and otherwise the Conservative party. Then group our data by province and calculate the total votes, and we have the estimated election poll result.

Results

We plot an AUC curve(Area under the curve) to visualize the model performance:

Fig.1.



We see that the area under the curve is 0.692, which means there is 69.2% probability that our model will predict the true result.

In addition, when conducting the post-stratification analysis, we divide census data into 10 cells based on province, and apply our logistic regression model on census data. Then we weight each proportion estimate by the corresponding population of that cell and sum them together and divide that by the total population, and finally we calculate our $\hat{Y}^{PS}$ = 0.63 by applying formula $\hat{Y}^{PS} = \frac{\sum_j N_j \hat{y_j}}{\sum_j N_j}$, So we can predict the voting probability for the Liberal party by $\hat{Y}^{PS}$, meaning the estimated votes proportion in favour of the Liberal party is 63%, where the estimation is based on our post-stratification analysis of the proportion of voters in favor of the Liberal party modeled by a logistic regression model, which accounted for agegroup, sex, education and province. Besides, when predicting the total popular votes, we have 11840 votes for the Liberal party in census data, and 6910 votes for the Conservative party. Combining the estimated votes proportion with the whole Canadian population, we have 22,197,145 people votes for the Liberal party.

Discussion

Summary

Our study uses a logistic regression model with post-stratification(MRP) to predict the voting results of the 2019 Canadian federal election, assuming the whole Canadian population will vote. We firstly build a logistic regression model by the CES data with predictors agegroup, sex, education and province, then we conduct a post-stratification analysis on the GSS data to estimate the proportion of votes for the Liberal party, where logistic regression model is applied on census data for each province to predict the votes proportion. Finally, we calculate $\hat{Y}^{PS}$ = 0.63, and we apply this estimated proportion to the whole Canadian population to compute the total popular votes for the Liberal party.

Conclusion

Based on the t-test results for the predictors of the logistic model that we diagnosed in the Model section, we consider agegroup, sex, education and province are good predictors. Based on the AIC compared to the alternative model, we consider our logistic model is a good model. By the ROC curve, the probability of our logistic regression model captures the true proportion of votes for the Liberal party is 69.2%, meaning that our model is reliable. As we predicted in the proportion of votes for the Liberal party by post-stratification, our $\hat{y}^{PS}$ is 0.63, which is used to estimate the true voting proportion, meaning that the proportion of voters will vote for the Liberal party is 63%. Therefore, it is estimated that 22,197,145 Canadian citizens will vote for the Liberal party, assuming the whole Canadian population vote, and we can safely conclude that the liberal party is predicted to win the election by our model.

While comparing the true voting results in the 2019 Canadian federal election, the Liberal party won 6,018,728 votes and the Conservative party won 6,239,227 votes, where the voting proportion for the liberal party is 49.1%. What causing the difference between estimated proportion and true voting results might be the difference of population of the voters, since the 2019 Canadian federal election did not have the whole Canadian population as electors, whereas our study focuses on the whole Canadian population as voters. Besides, we only consider a two-party political system, while Canada has a multi-party political system, which may cause errors in estimated results, and we will fully discuss in the limitation section.

Weaknesses

There are some limitations in our study. Firstly, we only consider the case that all voters vote for either the Liberal party or the Conservative party; However, Canada has a multi-party political system, so that the estimated results may be larger than the true proportion. For instance, we estimate the voting proportion is

63% for the Liberal party, and 37% for the Conservative party, while the true 2019 voting proportion is 33.12% and 34.43%, respectively, where the estimated proportion is indeed larger than the true voting proportion. When mapping the CES data and the GSS data, we filter out age greater than 80 in the CES data, we filter people only with sex male or female in the CES data, we summarise 11 education levels into 7 levels in the CES data, and we remove province of Northwest Territories, Yukon and Nunavut, to map the two datasets. This may lead to the prediction result to be idealistic, which may affect the accuracy of the prediction result. Similarly, when cleaning the dataset, we drop some missing data (NAs) for better modeling, but it decreases the observation size, which may also lead to the inaccuracy of our model and predicted results. Additionally, when collecting data, the CES survey balances the percentage of gender and age within each strata, which might be non-representative and cause imperfect sampling population; therefore, we create agegroup based on age and use agegroup as the predictor instead of age to reduce the influence of potential non-representative. Also, there are some non-response problems when collecting the data. For example, when collecting the GSS data, there are non-response or refused; therefore, in order to fix this problem, more phone calls will be made to contact or explain the importance of the survey as well as encourage participation. Besides, the population size of the GSS data is not large enough to be predicted as census data. Although it contains 20602 observations, the sample size is relatively small to use as census data. Thus, this may affect the accuracy of the prediction result, and may be the reason of the error between the estimated proportion and the true voting results. What's more, the census data is collected in 2017, so our prediction lacks of timeliness, and may cause prediction results inaccurate. Another limitation is, when fitting the logistic model, we use four predictors, agegroup, sex, education and province, to reduce the probability of model overfitting; however, fewer predictors may have the problem of not fully explaining the response variable by explanatory variables. Finally, when estimating the total popular votes, we did not combine the Canadian election policy, such that there are 338 electoral districts in Canada, and within each electoral district, one Members of Parliament(MP) will be voted to win the seats for Canada's House of Commons, where the political party who win the majority seats is the winner. This is due to lack of variable of electoral district in the dataset. Therefore, due to the limitations of our study, we might not able to predict true voting results of the 2019 Canadian federal election the best, assuming the whole Canadian population will vote, but we can improve our study by several methods, which will be discussed in the Next step section.

Next Steps

We are going to focus on some aspects for future improvement. Firstly, in terms of Canada's multi-party political system, we can build 5 logistic models for each big political party, Liberal, Conservative, New Democratic, Bloc Québécois and Green, creating 5 new variables in survey data record whether the participant votes for a specific party or not. Then, we can use these 5 variables as the response variables to fit in 5 logistic models separately, where each model predicts the voting proportion for a specific party. Therefore, we can have the estimated proportion of each 5 political parties. Secondly, when considering non-representative and imperfect sampling population that might be caused by the balance of percentage of gender and age within each strata in the CES data, we can build a multilevel regression model instead of a logistic model to adjust non-representative samples. Additionally, we can use some larger size dataset as census data to improve the model accuracy and reduce the estimated errors, for example, the Canadian Census Data. Besides, to find a better regression model, we can use some regression methods to find the best predictors for the model, for example, we can apply backward elimination with AIC/BIC to find a better-performed model, so that we can have a more accurate estimated voting proportion by post-stratification. When considering combine the Canadian election policy, we can use province to replace electoral district, such that we use the number of electoral district in each province multiply by $\hat{y_j}$, the estimated weight of vote proportion for each province, and we can obtain the estimated number of MPs for different parties in each province, and finally compute the number of seats won of the Canada's House of Commons for each party, so that we can know which party wins the majority seats. Finally, in terms of timeliness, we can add 2 years on the age variable in census data because the GSS data is collected in 2017 but we are actually predicting

the 2019 election. This can make our model become more reliable to predict the 2019 election results; however, this may not be considered as a good solution, so we can make a survey and collect new data for future elections.

References

Elections Canada. 43rd General Election. (2019). https://www.elections.ca/content.aspx?section=ele&document=index&dir=pas/43ge&lang=e (https://www.elections.ca/content.aspx?section=ele&document=index&dir=pas/43ge&lang=e)

General social survey on Family (cycle 31), 2017. [dataset] https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm (https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm)

Stephenson, Laura B., Allison Harell, Daniel Rubenson and Peter John Loewen. (2019). The 2019 Canadian Election Study – Online Collection. [dataset] http://www.ces-eec.ca/ (http://www.ces-eec.ca/)

Appendix

GitHub repo: https://github.com/austinwjy/sta304-final-report (https://github.com/austinwjy/sta304-final-report)

---

1. Suppose the results of 2019 federal election is unknown↩
2. A random sampling method by dividing the population into groups and randomly sample↩
3. List of all dwellings, used to group together telephone numbers associated with the same address↩
4. Too many predictors may cause model overfitting↩