How The 2019 Canadian Federal Election Would Have Been Different If Everyone Had Voted

Zuoyu Wang

19/12/2020

Keywords

Canadian Federal Election, Post-stratification, Multilevel Regression Model, Election forecasting

Code and data supporting this analysis is available at: https://github.com/zuoyuwang/2019-Vote-Prediction.git

Abstract

Due to the low vote turnout in the 2019 Canadian Federal Election, this study utilizes logistic regression model and post-gratification technique to make a prediction for the popular voting intention of the Liberal party by assuming everyone will vote. The predicted result indicates the popular vote outcome for the Liberal party might be dropped to 29.3% if everyone has voted. Therefore, those who did not vote might have great influence to the election result, and everyone should respect and treasure the rights to vote.

1. Introduction

In the 2019 Canadian federal election, the Liberal party led by Justin Trudeau had won 157 seats to form the minority government though losing 20 more seats compared with last election. Also, the popular vote outcome for the Liberal party had dropped from 38.5% to 33.1% (CBC News, 2019). Nevertheless the Liberal party won the election, the voting statistics had demonstrated a significant decrease in people who support the Liberal party.

However, the vote turnout was only 67%, which means there were almost 9 million citizens did not vote in 2019 (Elections Canada, 2019). Since those missing decisions took about one third of the voters' population, such huge amount of people may contain valuable information that were not taken into consideration. Therefore, whether the outcome will be different if everyone had voted is an interesting and important topic since it may reveal the potential influence of those who did not vote and encourage Canadian citizens to vote in the next election.

Fortunately, the 2019 Canadian Election Study(CES) had constructed an online survey investigating voters' preference for political parties even though they decided not to vote in the election day. In this paper, the 2019 CES online survey data will be applied to fit a logistic regression model regarding the voting intention for the Liberal party. Moreover, the post-stratification technique will be introduced to estimate the popular vote outcome for the Liberal party by assuming everyone has voted using the census data generated from the 2017 General Social Survey(GSS).

In the methodology section (Section 2) the two datasets, the detailed process of the logistic regression model and post-stratification will be described. Result for the model and estimation will be provided in the result section (Section 3) Furthermore, the discussion section (Section 4) will conclude the major finding of this study, point out any weakness, and provide suggestions for future research.

2. Methodology

2.1 Data

This study involves two datasets. The survey dataset is extracted from the 2019 Canadian Election Study(CES) and the census data is based on the 2017 General Social Survey(GSS). The CES data is an online survey conducted during the election campaign period whose population includes Canadian citizens and permanent residents, aged 18 or older. According to its codebook, it aims for 50% men and 50% women, and 28% of respondents aged 18- 34, 33% aged 35-54 and 39% aged 55 and higher. The survey instrument was presented on the Qualtrics online platform (CES codebook, 2019). The GSS data's target population includes all non-institutionalized persons 15 years of age and older, living in the 10 provinces of Canada. Its sampling frame is created by lists of telephone numbers in use and list of all dwellings within the ten provinces. The survey involves stratified sampling by dividing the country into 27 stratas, and it is conducted by computer assisted telephone interviews (GSS codebook, 2017). Both datasets create additional values for each variable to represent missing/invalid/skip answers.

The CES data also involves a post-election recontact. Ideally, using the post-election statistic will yield more accurate prediction since it reveals the actual vote results of individuals. However, post-election data has huge amount of missing observations compared with data collected during election campaign period. As a

result, this study only considers the data collected before the election day. Fortunately, one excellent feature of the CES data is that it not only asks citizens' voting intention, but also asks people who are unlikely to vote about their preference of political parties by assuming they will vote. These two variables can be combined to be the response variable for building the logistic model since it can represent opinions for both people who decide to vote and those who are unlikely to vote. The potential predictor variables initially consisting of 12 variables relating to one's educational, social, economic information. Those 12 variables are shared in both datasets, which can be applied for post-stratification calculation.

For producing valid model and prediction, all samples who are not eligible to vote, such as people who aged below 18 and permanent residences, and all missing observations are filtered out. Furthermore, the variables from two datasets are reformatted so that they have the same variable names and categories for each variable. After doing so, the survey data and census data now contain sample size of 15677 and 12393 respectively. Details of all formatted variables are shown in Table 1. Notice that the GSS data is conducted in 2017, in order to utilize it to make valid prediction for the 2019 election, the variable 'age' is increased by 2 before filtering and dividing into different age groups.

Table 1: Variables Summary Table

Names	Types	Levels
VoteLiberal	Binary	[Yes/No]
Gender	Categorical	[Male/Female/Other]
Province	Categorical	[Alberta/British Columbia/Manitoba/...]
Education	Binary	[Below University Degree/University Degree and Above]
Religion	Binary	[Important/NotImportant]
BornInCA	Binary	[Yes/No]
MotherTongue	Categorical	[English/French/NonOfficial/Multiple]
Employment	Binary	[Employed/Other]
HaveChild	Binary	[Yes/No]
FamilyIncome	Categorical	[Below $50,000/$50,000-$99,999/$100,000 and Above]
MarriageStatus	Binary	[Married/Other]
HouseholdSize	Categorical	[One/Two/Three/Four/Five or More]
AgeGroup	Categorical	[18-34/35-54/55 and Higher]

2.2 Logistic Model

Since the objective of this study is to predict the popular vote outcome for the Liberal party in 2019 by assuming everyone in the census data will vote, the result is a binary variable indicating yes or no. Therefore, a logistic regression model can be applied to predict the probability of individuals' voting intention towards the Liberal party. Below is the full model based on the 12 potential predictor variables mentioned in the Data section:

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1\text{Gender} + \beta_2\text{Province} + \beta_3\text{Education} + \beta_4\text{Religion} + \beta_5\text{BornInCA} + \beta_6\text{MotherTongue}+$$

$$\beta_7\text{Employment} + \beta_8\text{HaveChild} + \beta_9\text{FamilyIncome} + \beta_{10}\text{MarriageStatus} + \beta_{11}\text{HouseholdSize} + \beta_{12}\text{AgeGroup}$$

The variable p stands for the probability of voters to vote for the Liberal party. Since each predictor is a categorical variable, the β coefficient for each represents the degree of change when switching categories, and β_0 is the bias term.

To perform further variable selection from the full model, this study will consider Akaike information criterion (AIC) stepwise selection method. AIC can be interpreted as a measure of the quality of a model and it provides penalty for different number of predictors in order to reduce overfitting or underfitting. The

stepwise selection method by AIC involves iteratively removing or adding variables and compare AIC to choose the best model. For alternative variable selection methods, Bayesian Information Criterion(BIC) also estimates goodness of fit for a model by performing a stronger penalty on the number of predictor variables compared with AIC. More technically, the model selected by AIC usually involves more variables and perform better predictions than BIC since BIC is more strict on the number of variables. Therefore, AIC is considered since the logistic regression model will be used for future prediction.

To make valid model diagnostic after variable selection, this study checks the multicollinearity between variables, the significance of each estimated coefficients, as well as the goodness of fit of the model. Multicollinearity happens when a predictor variable is high correlated with another, which can produce relatively large standard error for the model. The variance inflation factor(VIF) measures how a predictor variable is affected by the others. In this study, a value of VIF under 4 will indicate there is no significant correlation between predictor variables. The estimated coefficient could be interpreted as the odds ratio of the interest, and the 95% confidence intervals will be used to assess the significance of the estimates. If the confidence interval does not contain 1, it indicates this variable is significant. Finally, the AUC-ROC curve can be introduced to check how much the final model is capable of distinguishing between classes. Higher AUC value demonstrates the model gives a better performance on predicting voters' intention.

2.3 Post-stratification

Post-stratification technique involves dividing the census data into different demographic cells based on all the variables. Each cell will be weighted by their population when used for making predictions, then combine all cells will produce the final prediction on the entire census data. Following is the formula for post-stratification calculation:

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

Each Nj represents the population size of each cell, and $\hat{y}_j$ is the prediction for such cell based on the constructed logistic regression model. N is the total population size, which can also be written as the summation of all the cells' population ($\sum N_j$). To apply this formula, the census data will be grouped by all the predictor variables to form each cell. Each cell can be treated as a new input to the logistic regression model and yield an estimated result. Then add all the weighted predictions for each cell will produce the final predication.

2.4 Software

The software programming language R is utilized to perform all the data formatting and calculations mentioned above. Details of the code can be accessed by the provided GitHub link in the appendix.

3 Results

After performing model selection by AIC, the final logistic regression model predict the probability of people voting the Liberal party can be expressed as following:

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 \text{Gender} + \beta_2 \text{Province} + \beta_3 \text{Education} + \beta_4 \text{Religion} + \beta_5 \text{BornInCA} +$$
$$\beta_6 \text{MotherTongue} + \beta_7 \text{Employment} + \beta_8 \text{HaveChild} + \beta_9 \text{FamilyIncome} + \beta_{10} \text{AgeGroup}$$

Table 2 indicates the odd ratio of each predictor variables and corresponding 95% confidence intervals. For variables contain multiple categories such as "Province" and "MotherTongue", the confidence intervals for certain categories appear to contain 1, which indicates that there may not be significant different in

voting intention between those categories. However, in general the variable is significant since there are more categories whose confidence interval do not contain 1. Therefore, those variables are kept for the final logistic regression model.

Table 2: Odds Ratio and 95% Confidence Interval of Model Coefficients

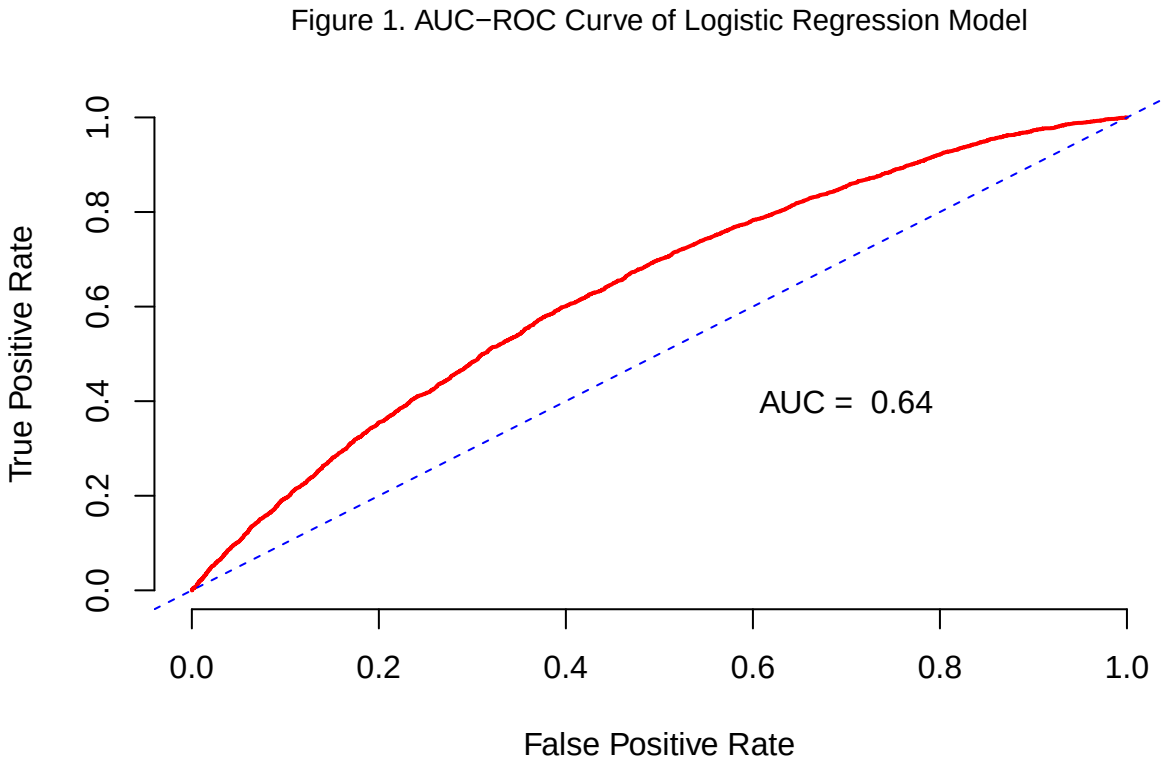
	est	2.5	97.5
(Intercept)	0.1572559	0.1272850	0.1942837
GenderMale	0.9244899	0.8599804	0.9938383
GenderOther	0.9377872	0.5769017	1.5244275
ProvinceBritish Columbia	2.1819785	1.8296467	2.6021582
ProvinceManitoba	2.2314227	1.7961465	2.7721832
ProvinceNew Brunswick	3.7213717	2.8907654	4.7906370
ProvinceNewfoundland and Labrador	4.2320512	3.2431010	5.5225715
ProvinceNorthwest Territories	2.5493292	0.7831730	8.2983959
ProvinceNova Scotia	3.9960166	3.1648784	5.0454225
ProvinceNunavut	4.1486936	1.4164416	12.1513364
ProvinceOntario	3.0363287	2.6268038	3.5096994
ProvincePrince Edward Island	4.3041366	2.4563915	7.5417914
ProvinceQuebec	3.4846940	2.9170191	4.1628430
ProvinceSaskatchewan	0.8220348	0.6154058	1.0980418
ProvinceYukon	2.2858273	0.7051814	7.4094503
EducationUniversity Degree and Above	1.5629335	1.4511445	1.6833341
ReligionNotImportant	1.0975185	1.0204748	1.1803789
BornInCAYes	0.7803651	0.7006643	0.8691318
MotherTongueFrench	0.6607419	0.5715359	0.7638713
MotherTongueMultiple	1.1780471	1.0756371	1.2902075
MotherTongueNonOfficial	1.0810225	0.9042347	1.2923742
EmploymentOther	1.0652888	0.9803239	1.1576176
HaveChildYes	0.8232943	0.7605983	0.8911583
FamilyIncome$50,000-$99,999	0.9345152	0.8579642	1.0178965
FamilyIncomeBelow $50,000	0.8874880	0.8032078	0.9806117
AgeGroup35-54	1.1278004	1.0127092	1.2559713
AgeGroup55 and Higher	1.3322852	1.1915988	1.4895817

Table 3 shows the VIF value for each predictor variables. Each variable has a VIF less than 4, so there does not exist multicollinearity in the model.

Table 3: VIF Table of Predictor Variables

	GVIF	Df	GVIF^(1/(2*Df))
Gender	1.053331	2	1.013074
Province	2.225737	12	1.033899
Education	1.116189	1	1.056498
Religion	1.064457	1	1.031725
BornInCA	1.236892	1	1.112156
MotherTongue	2.705916	3	1.180463
Employment	1.348279	1	1.161154
HaveChild	1.150971	1	1.072833
FamilyIncome	1.206588	2	1.048069
AgeGroup	1.492851	2	1.105361

Figure 1 is the AUC-ROC curve for the constructed logistic regression model. The curve shows that the constructed logistic regression model is able to distinguish whether people will vote for the Liberal party 64 percent of the time.



By performing all the model diagnostics, the constructed logistic regression model is accurate enough for making prediction on the census data, where post-stratification will be applied. According to all the variables and their categories, the census data is been grouped into 3340 cells, each cell is corresponding to a unique combination of the categories for predictor variables. Then input such combination will yield the estimated probability for voting the Liberal party within each cell. Table 4 shows a summary of population size and voting estimations for the 3340 cells. Finally, sum all the estimations weighted by their population produce the final prediction. As a result, the probability for voting the Liberal party is around 0.293 if everyone in the census data will vote. Compared with the real popular vote come in 2019, this prediction illustrates a nearly 4 percent decrease for supporting the Liberal party(33.1% vs 29.3%).

Table 4: Summary of Population Size and Voting Estimation for the Cells

Measure	Min	Mean	SD	Max
Population size	1.0000000	3.7104790	5.6892429	59.0000000
Vote Probability Estimation	0.0610723	0.3023394	0.1122076	0.5929244

4. Discussion

4.1 Conclusion

This study's objective is to predict the popular vote outcome for the Liberal party in the 2019 Canadian Federal Election by assuming everyone will vote since the actual vote turnout is only 67%. To generate valid prediction, the 2019 CES dataset is applied to build a logistic regression model and the 2017 GSS dataset is considered as the census data to introduce the post-stratification calculation. By performing AIC stepwise election method, the final logistic regression model is consist of ten predictor variables. After all the model diagnostics and post-stratification calculation, the final prediction indicates the vote rate for the Liberal party might drop from 33.1% to 29.3% if everyone will vote. It may raise voters' attention that the election results might be different if everyone can seize the opportunity to vote. However, there still exists certain weaknesses in the whole prediction procedure. Any potential problem will be addressed in the following sub sections. For further research interest, some suggestive steps will also be discussed below.

4.2 Weakness

The weaknesses of this study exist in both survey and estimation procedure. The CES data is collected through the Qualtrics online platform. "However, due to platform constraints, it is not possible to go back to a previous block. Due to the complex design of the survey, many small blocks of questions were used, making this a common restriction."(CES codebook, 2019) As a result, the technical constraints might cause bias in the dataset. Moreover, since the survey is conducted online, it involves random sampling, which may cause the sample to be less representative of the population. Besides the online survey dataset, the CES website also provides a phone survey dataset with much fewer samples. Nevertheless, both datasets cannot truly represent the whole voting population, and this study only considers the online dataset due to its large sample size. For the GSS data, though it is conducted through stratified sampling, it only has a response rate of 52.4%, which is less representative when considering as the census data (GSS codebook, 2019). Both datasets also contain lots of missing values which might be caused by respondents' partial response or misunderstanding, and those missing data further reduce the accuracy of the model result and the final prediction. As mentioned above, the GSS data is collected in 2017, though the variable 'age' has been increased by 2 to accommodate the 2019 election before building the model, it is not realistic and will influence the predicted results as well.

Besides the inherent defects in the survey and datasets, the procedure leading to the final prediction also contain potential drawback. To perform valid post-stratification calculation, only variables that are shared in both datasets are considered predictor variables. By doing so, other significant variables that may increase the model prediction is neglected. Also, when formatting variables such as 'Gender' and 'Province', the two datasets do not fit perfectly. For example, the GSS dataset only considers female and male since it originally requires sex of the respondent, while the CES dataset offers 'Other' options since it originally requires gender of the respondents. As a result, when doing post-stratification calculations, the census data does not include the category 'Other' for the variable 'Gender', which leads the results to be imprecise. Similarly, when counting respondent's residence, the GSS data only offers answer for the ten provinces, while the CES data also collects answers of the three territories. Therefore, the variable selection and formatting procedure causes the census data to be less representative. Moreover, when performing post-stratification technique, it is obvious that the census data does not cover all the demographic cells. Table 5 shows the categories of all the chosen variables for the final logistic regression model. There should be 89,856 different combinations of variables' categories, but the census data only forms 3340 cells. As mentioned above, part of the reason is the lack of certain categories in the census data, while the major problem is the insufficient sample size of the census data. Table 4 indicates the mean population size is merely 3.7 for all the cells formed by the census data. Hence the post-stratification may not be accurate enough.

Last but not least, this study only predicts the popular vote outcome, and does not predict the seats won for the Liberal party. The actual election result is determined by the majority of seats winning rather than the popular vote. Therefore, to check whether assuming everyone will vote can indeed influence the final election results requires future research.

Table 5: Model Variables Summary Table

Names	Types	Levels
VoteLiberal	Binary	[Yes/No]
Gender	Categorical	[Male/Female/Other]
Province	Categorical	[Alberta/British Columbia/Manitoba/...]
Education	Binary	[Below University Degree/University Degree and Above]
Religion	Binary	[Important/NotImportant]
BornInCA	Binary	[Yes/No]
MotherTongue	Categorical	[English/French/NonOfficial/Multiple]
Employment	Binary	[Employed/Other]
HaveChild	Binary	[Yes/No]
FamilyIncome	Categorical	[Below $50,000/$50,000-$99,999/$100,000 and Above]
AgeGroup	Categorical	[18-34/35-54/55 and Higher]

4.1 Next Steps

Considering the weaknesses mentioned above, any subsequent work can be introduced to enhance the prediction. First of all, the survey methodology can be improved to reduce imperfect sampling population and non-response issues. The designer should consider more representative sampling frame, and spread out the survey in more various channels, such as online, phone and in-person surveys. Study groups may also seek for cooperation with the government to gain citizens' attention since the federal election is a national topic. To get more accurate prediction results, future census survey might focus more on the election-related questions so that more valuable and significant variables can be considered for building the model. Researchers can also use alternative estimation procedures such as using different models and techniques to predict and compare the result made in this study. Furthermore, more data and ideas can be involved to make valid prediction for the seats winning for the Liberal party by assuming everyone can vote. As a result, the prediction will be more accurate and this topic will be more influential. By advocating the significance of voting, further vote turnout is wished to be increased, and the election results will be more conformed to citizens' wishes.

5. Reference

[1] CBCNews. Canada Votes 2019. Retrieved Dec 9, 2019 from:https://newsinteractives.cbc.ca/elections/federal/2019/results/

[2] Elections Canada. 2019 Voter Turnout at Federal Elections and Referendums. Retrieved Dec 9, 2019 from: https://www.elections.ca/content.aspx?section=ele&dir=turn&document=index&lang=e

[3] 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

[4] 2017 General Social Survey (GSS): Families Cycle 31. Retrieved Dec 9, 2020 from: http://www.chass.utoronto.ca/

[5] Wu, Changbao, and Mary E. Thompson. "Basic Concepts in Survey Sampling." Sampling Theory and Practice. Springer, Cham, 2020. 3-52.

[6] Alan Agresti "An Introduction to Categorical Data Analysis" Second Edition. John Wiley & Sons, Inc. 2007. 99-163.

Appendix

Code and data supporting this analysis is available at: https://github.com/zuoyuwang/2019-Vote-Prediction.git