Who Would Win the 2019 Canadian Federal Election If Every Canadian Voted: A MRPs Analysis

Mo wanchen (1003989035)

Due: December 22, 2020

Code and data supporting this analysis is available at: https://github.com/wanchen-mo/sta304_final

Abstract

Canada's first-past-the-post system is a special electoral system since the most proportion of direct voting from citizen does not shape the electoral result. The in this analysis, I will investigate the situation when every Canadian aged above 18 could directly vote for their preferable party during the election. This analysis will simulate logistic MRPs models based on the CES2019 website data and impose post-stratification analysis based on the GSS data to estimate the proportion of the observations who would vote for Liberal Party and Conservative Party in the 2019 Canadian Federal election.

Key Words: 2019 Canadian Federal Election, MRPs, Logistic Regression, Liberal Party, Conservative Party

Introduction

2019 Canadian Federal Election was an important political event, which had drawn many attentions from the public. Domestically, the public's doubts toward the leadership of Liberal Party is mounting as Justin Trudeau, the leader of Liberal Party, continuously broke his promises since he became the Prime Minister. Thus, this was a change for Conservative Party to come into the power. Statistics shows that, the preliminary number of votes gained by Liberal Party in 2019 Election was 5,915,950 (about 33.1%), which was slightly lower than that for Conservative Party with 6,155,662 vote, accounting for 34.3% (Statista, 2019). However, because of Canada's special electoral system, gaining the most popular votes does not mean winning the election. The result turned out that Liberal Party won the 2019 election, and Justin Trudeau re-elected the Prime Minister, because Liberal Party had gained the most number of seats in the House of Common.

According to Canada's electoral system, voter does not directly vote for specific candidate they support. Instead, they would vote for the member who they believe could represent their electoral district to enter the House of Commons during the general election. It is noticeable that there are in total of 338 seats in the House of Common, corresponding to 338 electoral districts in Canada (Election Canada, 2009). After the general election, the leader of the party whose members win the most seats in the House of Common will form the government and become the Prime Minister, while the second majority party will become the official opposition (Election Canada, 2009). The result of the 2019 Canadian Federal Election shows that 157 elected representations belong to Liberal Party while Conservative Party gained 121 seats, and the remaining 60 seats were distributed among New Democrats (24), Bloc Quebecois (32), Green (3), and Other (1) (Support The Guardian, 2019). Thus, although Conservative party gained the most votes from the public, Liberal Party still gained the victory in the election.

Therefore, this paper aims to study the result of 2019 Canadian Federal Election by applying MPRs to simulate the prediction based on the assumption that if everyone aged above 18 could directly vote for their preferable party. According to Keller (2019), multilevel regression and post-stratification (MRPs) has always been a widely used approach for predicting the result of the coming election based on sample data from opinion survey collected from the public because this is a low-cost and efficient approach. MRPs would group survey participants by specific demographic parameters, like age, education and gender, and create a model to simulate the relationship between election outcome and individual demographic factors. For example, MRPs could predict the probability of a female aged between 20-30 voting for Conservative Party. I will also compare MRPs model to simple logistic models to investigate which specific model performs better in predicting the result.

Methodology

Data

The datasets using as survey data and census data for this analysis are 2019 Canadian Election Study and 2017 General Social Survey respectively. The 2019 Canadian Election Study, specifically the online survey version, is a survey conducted by the CES team to investigate and enhance citizens' understanding on electoral democracy in Canada. This series of datasets collect citizens' basic information, attitude toward parties, political issue, and social issue. This survey is conducted through online platform, which enables the CES team to has the access to larger sample size since the population of internet users is huge and the ability of internet allows the team to design more complex and comprehensive survey. The following table is a glimpse at the baseline characteristic of this dataset, specifically the characteristics of the variables that will be used for building models.

Table 1: Baseline Characteristics of 2019 Canadian Election Study

Total # of observations: 37822 **Total # of variables: 620**		
Variables	**Category**	**Details**
Cps19_votechoice	Liberal Party	9510 obs.
	Conservative Party	8058 obs.
	Don't know	4908 obs.
	Other (e.g. ndp, green…)	15252 obs.
Cps19_gender	Female	15551 obs.
	Male	21980 obs.
Cps19_education	Bachelor's degree	9192 obs.
	Completed college	7702 obs.
	Completed high school	5865 obs.
	Other (9 more edu. levels)	15063 obs.
Cps19_province	Ontario	14808 obs.
	Quebec	8399 obs.
	Alberta	4481 obs.
	Other provinces	10134 obs.
Cps19_Age	Minimum	18
	Maximum	99
	Median	49

As it is listed above, variables including gender, education, province and age will be used as the independent variables while votechoice will be the respond variable. It is noticeable that for the variable votechoice, 4908 observations choose the answer "Don't know/ Prefer not to answer", which will impact the final result predicted by model and drive down the proportion of voting for Liberal and Conservative Party. Also, among 15252 obs. in other (e.g. ndp green...), 6258 obs. are NA, which will be filtered out while building the model and the sample size largely will be reduced largely. Moreover, the variable cps19_age is not in the original dataset, instead, the dataset records participants' year of birth, and cps19_age is mutated based on their yob. However, we will mutate this variable again into age group like "25 or younger" and "older than 75" when using is as the independent variable.

Moving on to the census data, the 2017 General Social Survey (GSS) is a survey about Canadian Family, conducted through phone interview to collect family's basic information, living standard and family member's well-being across 10 provinces in Canada. This dataset will be used for generating a post-stratification to estimate the proportion of the population who would vote for Liberal Party and Conservative Party respectively.

Model

For the analysis, I will use models combining logistic regression and multilevel regression to simulate the prediction over participant's vote choice based on the 2019 Canadian Election Study dataset. For better understanding, a logistic regression is a used when we need to model the relationship between a single dependent variable and various independent variables. It is noticeable that logistic regression has similar characteristics to linear regression but the major different is that a logistic regression requires the output to be a categorical and binary respond variable, which mean the output needed to be able to transform into 0 or 1 in rstudio. A multilevel regression allows us to group observations by assigned variables combination, which might help to simulate a better prediction over the relationship between dependent and independent variables.

I will first build two simple logistic regression models without considering multilevel regression using variables gender, education, age group and province, and the models are model #1: lib_model1 and con_model1. One problem is that the respond variable cps19_votechoice is not binary and it includes many parties, but I want to predict the results for conservative party and liberal party only since they are the two most competitive and popular parties in Canada. Thus, I mutated two new variables "Liberal_or_others" and "Conservative_or_others" to be the output for each simple logistic regression model. Liberal_or_others will return 1 if participants choose liberal party and return 0 if they choose others, and the same logic applies to Conservative_or_others. Comparing to an alternative model that filtered all other parties out, the prediction of this model will be more realistic by taking other parties into consideration, since it is impossible for all Canadian to vote between liberal party and conservative party only.

Based on the two respond variables, Conservative_or_others and Liberal_or_others, I build two logistic multilevel regression models with a cell group, named model #2: lib_model2 and con_model2. The cell would group observations in the dataset by their education level and gender. For example, observations will fall into categories like "Above Bachelor's degree Female" and "Some university/college Male".

To compare model #1 and model #2, we will look at their ROC curve and BIC score. In a graph of ROC curve, there is a score named AUC which is ranged from 0 to 1, and it examine the probability of the model prediction's correctness. In other words, the closer the AUC score to 1, the better the model since AUC = 1 indicates that this model's prediction is 100% correct. The AUC for model #1 is 0.614 for lib_model1 an 0.685 for con_model1. For model #2, the AUC is 0.615 (lib_model2) and 0.685 (con_model2). The figures for model #1 and model #2 are pretty close, but the scores for model #2 are slightly higher than that of model #1. Turning to the BIC score which demonstrates the model's accuracy and penalize the model's complexity heavily. The lower the BIC score, the more realistic the model prediction is. BIC scores for model #1: 36443 (lib_model1), and 34296 (con_model2); BIC score for model #2: 36431 (lib_model2), and 34283(con_model2). By comparing scores for both models, the BIC score for model #2 that are grouped by

cell groups are generally lower than the scores for model #1. Thus, we can conclude that model#2 perform better predictions of participants' vote choice comparing to model #1, and I will mainly discuss model #2 in the following sections, and the equations generated by the two logistic multilevel regression models for liberal party and conservative party are shown below:

Level 1: Individual Level:

Liberal Party:

$$log(\frac{y_{jLib}}{1-y_{jLib}}) = \beta_{0jLib} + \beta_{age} * x_{age} + \beta_{province} * x_{province} + \epsilon$$

Conservative Party:

$$log(\frac{y_{jcon}}{1-y_{jcon}}) = \beta_{0jcon} + \beta_{age} * x_{age} + \beta_{province} * x_{province} + \epsilon$$

The above equations are very similar, and the only different will be the output and the simulated β_i input for age and province. The outputs for both equations, y_{jLib} and y_{jcon} are the predictions over the probability of voters in each cell group (j^{th} group) voting for Liberal Party and Conservative Party respectively. β_{0jLib} and β_{0jcon} are the intercepts, which means they represents the random effect of each cell group in the corresponding model. β_{age} and $\beta_{province}$ represents the coefficients of each corresponding independent variable. For example, when holding the input of province as zero, if the observation is aged from 26 to 45, then their probability for voting for a party would increase or decrease by the coefficient of $\beta_{age26to45}$. Finally, the ϵ is the random error term.

Level 2: Group Level:

Liberal Party:

$$\beta_{0jLib} = r_{00} + r_{0jLib} * W_{jLib} + u_{0jLib}$$

Conservative Party:

$$\beta_{0jcon} = r_{00} + r_{0jcon} * W_{con} + u_{0jcon}$$

The outputs of above group level equations are the intercepts of individual level's equations. r_{00} is a non-random intercept, which is generated though fixed effect.
r_{0jLib} and r_{0jcon} are the log-probability of participants from each cell group voting for Liberal Party and Conservative Party respectively, and the W's terms are the j^{th} cell group. Similarly, the last components u_{0jLib} and u_{0jcon} are random errors.

Post-Stratification

After simulating the logistic multilevel regression models, I have conducted a post-stratification analysis to estimate the proportions of population who would support Liberal Party and Conservative Party respectively. By conducting a post-stratification analysis, we would adjust the weight of estimates for each cell group, summing up all figures and dividing by the population size. During this process, possible biases in the survey data could be reduced. In this analysis, I create the cell group based on two variables, education and gender. For example, conservative would be grouped by categories including Bachelor's Degree Female or Above Bachelor's Degree Male. Some researches reveal that voting decision might be affected by education and gender largely. Education directly impacts the voter turnout positively since this group of people might have deeper understanding and receive more information about political life comparing with those with less education (Feess, 2001). Also, as more female enter political system, gender is influencing the vote significantly due the different political stand and the desire for gender equality (Goodyear-Grant, E, Bittner,

A, 2018). Thus, that are the reasons why I chose education and gender as the cell. After then, I estimate the proportion of each cell group voting for Liberal Party or Conservative Party. Taking the summation of all proportions from each cell group and dividing by the total number of participants in the census data, I could obtain the final results of the predictions toward proportions of the voting.

Results

Table 2 shown below displays all estimated coefficients (β_{age} and $\beta_{province}$), for the equations of individual level simulated through the logistic multilevel regression models. For example, if the observation is aged from 46 to 65 and lives in Ontario, and their probability of voting for Liberal Party would increase by 0.009 (age_group) and 0.999 (province). It is noticeable that the estimated intercepts in this sectors are r00 for the equations of β_{0Lib} and β_{0Lib} in group level.

Table 2: Estimated Coefficients ($\beta_{age,province}$) for Individual Level Equations

Estimated Coefficients ($\beta_{age, province}$) for Individual Level			
Liberal Party		**Conservative Party**	
Intercept (r_{00})	-1.73199	**Intercept (r_{00})**	-0.41917
Age_group (26 to 45)	-0.03333	**Age_group (26 to 45)**	0.54107
Age_group (46 to 65)	0.00918	**Age_group (46 to 65)**	0.70295
Age_group (66 to 75)	0.17378	**Age_group (66 to 75)**	0.73938
Age_group (older than 75)	0.16505	**Age_group (older than 75)**	0.99165
Province (B.C)	0.65949	**Province (B.C)**	-1.24814
Province (Manitoba)	0.61372	**Province (Manitoba)**	-0.81962
Province (New Brunswick)	0.93685	**Province (New Brunswick)**	-1.41870
Province (Newfoundland and Labrador)	1.29099	**Province (Newfoundland and Labrador)**	-1.63241
Province (Northwest Territories)	1.12153	**Province (Northwest Territories)**	-1.94409
Province (Nova Scotia)	1.16728	**Province (Nova Scotia)**	-1.78691
Province (Nunavut)	1.25234	**Province (Nunavut)**	-0.79024
Province (Ontario)	0.99947	**Province (Ontario)**	-1.20896
Province (Prince Edward)	0.98633	**Province (Prince Edward)**	-1.89239
Province (Quebec)	0.83181	**Province (Quebec)**	-2.00023
Province (Saskatchewan)	-0.25849	**Province (Saskatchewan)**	-0.34189
Province (Yukon)	0.44081	**Province (Yukon)**	-1.34867

Table 3 shows all outputs of the slopes r_{0j} for each cell group in the equations β_{0Lib} and β_{0con} respectively. For instance, if the observation is a female with above Bachelor's degree, then her log-probability of voting for Liberal Party increases by 0.3071 but her log-probability of voting for Conservative Party decreases by 0.5262.

Table 3: Output of the Slopes (r0j) for the Group Level Equations

Output of the Slope (r_{0j}) for the Group Level			
Liberal Party		**Conservative Party**	
Above Bachelor's degree Female	0.3071	**Above Bachelor's degree Female**	-0.5262
Above Bachelor's degree Male	0.2899	**Above Bachelor's degree Male**	0.0257
Bachelor's degree Female	0.2499	**Bachelor's degree Female**	-0.3563
Bachelor's degree Male	0.2276	**Bachelor's degree Male**	0.2038
College Female	-0.1827	**College Female**	-0.1256
College Male	-0.2103	**College Male**	0.4166
High school Female	-0.3427	**High school Female**	0.0381
High school Male	-0.2464	**High school Male**	0.3742
Less than high school Female	-0.1840	**Less than high school Female**	-0.1388
Less than high school Male	0.2245	**Less than high school Male**	0.0385
Some university/college Female	-0.0397	**Some university/college Female**	-0.1751
Some university/college Male	-0.0845	**Some university/college Male**	0.2400

Result for Post-Stratification Analysis:

The post-stratification is conducted to estimate the proportions of participants who would vote for Liberal Party and Conservative Party. I simulated two MRPs models (Liberal and Conservative) based on the cell group consisting of education and gender, and factors including age group and province as the individual level. The post-stratification analysis is done based on the two MRPs, and finally calculating the estimation using the formula of yhatˆps. The results turn out the Liberal Party would gain 28.38% of total votes while Conservative Party would gain 27.88% of votes, which presents different results from the reality, and I will discuss this results as well as the reasons leading to this results in the discussion section.

Discussion

In this analysis, multilevel regression and post-stratification (MRPs) are used to simulate models predicting the result of 2019 Canadian Federal Election if every Canadian aged above 18 votes, specifically the proportions of votes gained by Liberal Party and Conservative Party. I first build two simple logistic regression models based on the 2019 Canadian Election Study dataset for both parties and compare them with multilevel regression models grouping observations by cell by their AUC and BIC scores, which indicates that multilevel regression models perform slightly better predictions over simple logistic regression models. Then, a post-stratification analysis is conducted using the 2017 General Social Survey to predict the proportion of total number of votes each party would get. The results turn out that Liberal Party and Conservative Party would gain 28.38% and 27.88% of total votes respectively. Thus, Liberal Party would eventually win the election in 2019, which matches the reality but in different way. This results indicate that 28.38% of population would vote for Liberal Party and 27.88% would vote for the Conservative Party if every Canadian aged above 18 could directly vote for the party they support.

Weaknesses and Next Steps

There are some limitations for this analysis in terms of the biasness of dataset and impractical assumption. Firstly, as it is displayed in table 1, under the variable cps19_votechoice, 4908 observations vote for the choice "Don't know/Prefer not to answer", which affect the prediction of final result and decrease the real number of votes each party would obtain. Thus, the predictions from my model show two figures of 28.38% and 27.88%, which are about 5% lower than the real results from the popular votes of 2019 Election. Also, the preliminary number of votes gained by Liberal Party (33.1%) was actually lower than that of Conservative Party (34.3%), and the results from my prediction are inconsistent with that. Since the survey dataset is collected through online platform, which makes the survey more complex to complete, and people who are unable to use internet could not participate, causing some bias in the dataset. Secondly, because of the restriction from the dataset, Canada's electoral system is not considered while building the MRPs models. Thus, this analysis assumes that if very voters above 18 could vote directly to nominated party, which is actually not realistic and impractical, since due to the special electoral system, Liberal Party finally win the 2019 Election with the most seats in the House of Common even though Conservative Party might have supports from citizens.

Limitations cause the MRPs models to simulate predictions that are less accurate and precise. For future analysis, I will consider using data collected by telephone interview which remains the survey simple but efficiently includes all necessary demographic information and allow more citizen to participate since telephone is more well-pervading than internet. Also, first-past-the-post electoral system needed to be considered while simulating to model in order to make the predictions more realistic and practice.

References

1. Access the Canadian Election Study Datasets a Little Easier. GitHub. Retrieved from https://hodgettsp.github.io/cesR/
2. Election Canada. (2009). The Electoral System of Canada. Retrieved from https://www.elections.ca/content.aspx?section=res&dir=ces&document=part1&lang=e
3. Dziak. J, Coffman. D, Lanza. S, Li. R. (2012). Sensitivity and Specificity of Information Criteria. AIC vs. BIC Sector. Retrieved from https://www.methodology.psu.edu/resources/AIC-vs-BIC/
4. Feess, S (2001). Does education influence voter turnout? Munich, GRIN Verlag. Retrieved from https://www.grin.com/document/101356
5. Goodyear-Grant, E, Bittner, A. (2018). How sex and gender influence how we vote. Retrieved from https://theconversation.com/how-sex-and-gender-influence-how-we-vote-106676
6. Keller, B. (2019). Federal elections: how the pollsters predict. Retrieved from https://www.horizons-mag.ch/2019/09/05/federal-elections-how-the-pollsters-predict/
7. Stephenson, Laura B., Allison Harell, Daniel Rubenson and Peter John Loewen. The 2019 Canadian Election Study – Online Collection. [dataset] http://www.ces-eec.ca/
8. SDA. (2017). General social survey on Family (cycle 31). Retrieved from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm
9. Support The Guardian. (2019). Canada election 2019: Full Results. Retrieved from https://www.theguardian.com/world/2019/oct/22/canada-election-2019-full-results
10. Statista. (2019). Preliminary number of total votes for each party in the Canada federal election held on October 21, 2019. Retrieved from https://www.statista.com/statistics/1062234/canada-election-2019-preliminary-results-for-popular-vote/
11. gss_cleaning Codes