If 'every one' Had Voted In the Election, How Does the Popular Votes of 2019 Canadian Federal Election Change?

Xiaoya Li

22/12/2020

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

Abstract

According to Elections Canada (2020), the voting rate in the 2019 Canadian Federal Election is 66.0%. Also, the popular votes for the Conservative party are higher than Liberal party, which is the winner of 2019 Canadian election Considering with low turnout rate and inconsistent result, this paper aims to analyses if every eligible voter in Canada have voted in 2019 Federal Election, in which the voter turnout rate is 100%, will the popular votes for Liberal higher than the Conservative party. The logistic multilevel regression models with post-stratification are used to simulate the result of whether the Liberal or Conservative Party can win, in which the education and province is the individual level, gender and age group is group level. This study concludes that the result of the popular votes for Liberal is higher than the Conservative party if the voter turnout rate is 100%.

Keywords

2019 Canadian Federal Election, multilevel logistic regression, post stratification, Conservative party, Liberal Party, turnout rate, popular vote rate.

Introduction

Multilevel regression with stratification is widely used in election result predictions (Public Opinion Quarterly, 2018), it can correct the bias caused by different observation weights between sample population and target population. Multilevel regression with stratification can be used when using a local dataset to predicate the 2019 Federal Election. For example, the training dataset has 90% of Ontario citizens, while 38% of electors are in Ontario in the 2019 Federal Election. Specifically, A post-stratification adjusts the estimated weight of each cell in the sample data, and it can adjust the model that builds with non-probability sampling data set. Furthermore, post-stratification allows non-representative sample data to analyze the population results. By using the non-population sampling dataset, the expenditure of collecting data can be reduced significantly.

The 2019 Canadian Federal Election used first-past-the-post voting (FPTP). Under this voting system, electors choose the members of Parliament (MPs). There are 338 MPs in the 2019 Canada Federal Election, and the votes of MPs decide who will win the election. The candidate who gets the most votes from the MPs' vote wins the election. However, there is a potential disadvantage for the voting system. That is, the winner of the election may not have the highest national popular vote. 2019 Canadian Federal Election

is an example. The Canadian Liberal Party led by Prime Minister Trudeau Jr. received only 5.85 million votes, which is less than the 6.1 million votes of the Conservative Party. But, due to the election policy, the Liberal Party won 157 seats (more than the Conservative Party's 121 seats) and became the winner of the election (Chris, 2020). Chris (2020) states that a low turnout rate may be a factor that contributed to the inconsistent result. According to Elections Canada (2020), 27,112,358 people are on the electors on the list in the 2019 Canadian Federal Election, but not all of the electors voted. The voting rate in the 2019 Canadian Federal Election is 66.0%. this study aims to analyze if the voter turnout rate is 100%, will the popular votes result will change.

In this paper, two logistic multilevel linear regression (MLR) models are built base on the survey dataset (2019 Canadian Election Study) to estimate the popular vote rates for the Liberal and Conservative Party respectively, and a census data set (2017 General Social Survey) for the post-stratification. In the Method section, I describe the data, the two MLR models, why the variable in the model are chosen, and the cell for post-stratification. In the result section, the formulas for both MLR models are shown, and output for formulas are organized in tables. Also, the popular vote rate for the Liberal and Conservative parties are compared in this section. In the discussion section, the paper connects the data with real world and conclude the factors leads to the estimated. Pointing out the weakness and next steps. Overall, this paper aims to analyze how the popular votes result different is of the 2019 Canadian Federal Election, if the turnout rate is 100%, and will the popular vote result and MP's vote result remain inconsistent?

Data

The logistic multilevel regression model (MLR) is used to simulate which candidate is the winner in the 2019 Canadian Federal Election, and stratification is used to adjust the estimates' weight of each cell. In terms of data set, 2019 Canadian Election Study(Stephenson,2019) is used to build the MLR models, and it is an online survey to collect the attitude of Canadians about the 2019 Federal Election, which has 37882 observations. It provides personal information about electors. The age, education, province, and gender elections choice are used to build the model. As Table 1 shows, 9510 observations (25.1%) in the survey data choose the Liberal Party in the election. The difference between the Liberal party and the Conservative Party is small, nearly 8058 observations vote for Conservation Party (21.3%). Other parties are not competitive in the election (Statistic Canada, 2018). Also, to determines the result under 100% turnout rate, the model not only analyzes people who "certain to vote", but also the observations who "likely to vote" and "unlikely to vote" are included as well.

Table1 2019 Canadian Election Study

total number obs. 37822		
number of variables: 620		
variables	category	observations
cps19_votes	Liberal Party	9510
	Conservative Party	8058
	Others	13902
cps19_gender	female	9510
	male	15551
cps19_education	Bachelor's	9192
	completed collage	7702
	others	20928
cps19_province	Alberta	4481
	Quebec	8399
	Ontario	14808
cps19_age	max	99
	min	18
	median	49

The 2017 General Social Survey (GSS) is a census data set and used for the post-stratification, which has 20602 observations. The target population for GSS is all non-institutionalized people over 15 years old and

living in 10 different provinces. GGS data set includes all variables that we used to build the model, though the age-weighted is different with all eligible electors in Canada.
Model

As mentioned above, the logistic MLR is used to estimate the final election result. since more than 2 parties in the election competition (Liberal party, conservative party Bloc Québécois, Green, New Democratic, etc) and the logistic MLR model only can produce a binary outcome, one logistic MLR model is not able to predicate the election result. Two logistic MLR are built to estimate the election result. The first model is used to predicate how many observations will vote for the Liberal party in percentage. Under the Liberal model, observations are divided into two groups, and a new variable is mutated to indicate that the observations voting Liberal party or not voting Liberal party. The second model is used to predicate how many observations will vote for the Conservative party in percentage. Similar to the Liberal model, observations are divided into two groups, and a new variable is mutated to indicate that the observations voting Conservative party and not voting Conservative party. The reason why only Liberal and Conservation party are analyzed because other parties are not likely to win according to on 2019 Federal election results and figure 1, and it is meaningless to analysis them. Only the Liberal and Conservation are the most competitive parties in the election.

This paper compares two logistic MLR models with the same individual level and different group levels to have a more accurate estimation. Model #1 combines gender and age group as group level, for example, female 25 or fewer years old, male 26 to 40 years old. The reason why this paper chose gender and age as group-level variables rely on the fact that different gender or age present different vote intentions. Also, Sean Simpson states that "If only young women were voting, the NDP and the Conservatives would be tied," (Stephenson, 2019). Furthermore, this paper build the MLR prediction model base on the data set that collected from the online survey, the average age of the observations in the data set may younger than the eligible electors' mean age. Model#2 combines gender and province, for example, female live in Ontario, male life in Alberta. There are significant gender and province gap in the election. According to EKOS Politics (2019), the Conservative party got the most votes in Alberta and Saskatchewan. Thus, the different states' weight between training data and the target population can cause a significant bias.

By compare the AIC, BIC of both model#1 and model#2. AIC is a value that indicates the accuracy of the model under the consideration of complexity. Thus, the smaller the AIC means the model is less complex and accurate. The AIC for model#1(gender and age as group level) is 36246.2 for the Conservative model and 34425.5 for the Liberal model while model#2 (states and gender as group level) is 37215.3.2 for the Conservative party and 35103.3 for the Liberal party, which is higher than model #1. So, model #1 performs better in terms of accuracy and less complexity. AUC scores range between 0 to 1, and it indicates the probability of the correctness of the model, and the AUC score for the prefect estimate model is 1. In our simulation, the AUC for model#1 (gender and age as group level) is 0.68 and 0.63 for conservative and liberal party, while the AUC for model #2 (gender and province as group level) is 0.67 and 0.61 for conservative and liberal, which indicates that model #1 has a better prediction. Thus, both AIC and AUC scores indicate model #2 which is the multilevel regression model with age and gender as the cell is a better choice. After organizing and cleaning the training data, the formula for the multilevel regression model is :

Level 1: Individual Level:

$$log(\frac{y_{jLiberal}}{1-y_{jLiberal}}) = \beta_{0jLiberal} + \beta_{provinceLiberal} * x_{province} + \beta_{eduLiberla} * x_{edu} + \epsilon$$

$$log(\frac{y_{jConservative}}{1-y_{jConservative}}) = \beta_{0jConservative} + \beta_{provinceConservative} * x_{province} + \beta_{eduConservative} * x_{edu} + \epsilon$$

in terms of individual level, in the first formula, $y_{jLiberal}$ is the proportion of electors in the j^{th} group level vote for Liberal Party, and $\beta_{0jliberal}$ is the random effect and intercept of the jth group. For example, the log-probability of voting Liberal party for less than 25 years old male is 0.0963 higher. $\beta_i Liberal$ is the

coefficients of corresponding independent variables, which are province and education. for example, with a unit change in Xi, the log probability for voting Liberal party will change $\beta_i Liberal$. The outcome for the first formula is the log probability of voting Liberal party. The only different between the first formula the second one is the second formula estimates the probability that electors will vote for Conservative party, others are the same.

Level 2: Group Level:

$$\beta_{0jLiberal} = r_{00Liberal} + r_{0jLiberal} * W_j Liberal + u_{0jLiberal}$$

$$\beta_{0jConservative} = r_{00Conservative} + r_{0jConservative} * W_j Conservative + u_{0jConservative}$$

In terms of group level, r00_Liberal is intercept of the intercept in the model, anf $\beta_0 jLiberal$, is fixed effects under the group, and it shows the log-probability of people who live in Alberta and has bachelor's degree. $r_{0jLiberal}$ is the log-probability of $jLiberal^{th}$ group of voting for Liberal. Wj_Liberal represents individual in the cell groups. For both group and individual level, error follows the standard normal distribution which represent as $\epsilon_L iberal$ and $u_{0jLiberal}$ respectively. The second formula is used for estimating the probability of voting Conservative and has the same logic as the first one.

Post stratification A post-stratification can adjust the weight of each cell in the sample data. With re-weighting and post-stratifying, some bias caused by the difference in weight between training data and the target population can be correct, in this case, the target population is all eligible Canadian electors. To consistent with the group level in the model, this paper creates cells based on different genders and age groups for post-stratification as well. After grouping by cells, the paper uses the Liberal MLR model estimates the proportion of voting Liberal in every single cell respectively. After that, each estimated proportion for every single cell is weighted based on the corresponded population size. Finally, adding the estimated proportions together and dividing by the total number of observations in the cleaned GSS (census) data set. The post-stratification process for the Conservative model is the same as the Liberal one, in which the cell is based on the gender and age group as well.

Result

Results from Liberal MLR model The table2 shows the output of fixed effect. r00 =0.01917 is the non-random intercept for group level in the formula. The other numbers are the values for $\beta_i Liberal$, where the i is the independent variable in the mode, including education and province.

Table2 Fixed Effect Output for Liberal Party

fixed effect for Liberal MLR model r00_Liberal, β_ Liberal education and province			
intercept	0.0598	Yukon	0.43801
education		British Columbia	0.65929
Less than high school diplo	-0.22479	Manitoba	0.61395
high school or equivalency certificate	-0.57207	New Brunswick	0.93582
higher than high school but lower than university	-0.44368	Newfoundland and Labrador	1.29517
University certificate, diploma or degree above the bachelor's level	-0.16973	Northwest Territories	1.11476
bachelor's diploma	baseline	Nova Scotia	1.1699
province		Quebec	0.83152
Nunavut	1.23267	Prince Edward Island	0.97418
Ontario	0.99965	Saskatchewan	-0.26085
		Alberta	baseline

The table3 shows the values for $r_{0j_Liberal}$, which is the slope of each in the group level.

Table3 Fixed Effect Output for Liberal Party

fixed effect for Conservative MLR model r00_conservative, β_conservative education and province			
intercept	0.0598	British Columbia	-24.354
education		Manitoba	-12.396
Less than high school diplo	-0.0027	New Brunswick	-14.819
high school or equivalency certificate	0.2983	Newfoundland and Labrador	-14.277
higher than high school but lower than university	0.2264	Northwest Territories	-3.141
University certificate, diploma or degree above the bachelor's level	-0.1709	Nova Scotia	-18.582
bachelor's diploma	baseline	Prince Edward Island	-7.838
province		Quebec	-41.702
Nunavut	-1.623	Saskatchewan	-4.907
Ontario	-30.626	Yukon	-3.083
		Alberta	baseline

Results from Conservative MLR model The table4 shows the output of fixed effect. r00 =0.01917 is the non-random intercept for group level in the formula. The other numbers are the values for $\beta_i Liberal$, where the i is the independent variable in the mode, including education and province.

Table4 r0jLiberal for Liberal Party

output for r0j_Liberal			
Female 25 or less	-0.12139	Male 25 or less	0.096306
Female 26 to 40	-0.08946	Male 26 to 40	0.014098
Female 40 to 55	-0.1168	Male 40 to 55	-0.11918
Female 55 to 70	0.074832	Male 55 to 70	0.003061
Female Above 70	0.212017	Male Above 70	0.049452

The table5 shows the values for $r_{0jLiberal}$, which is the slope of each in the group level.

Table5 r0jLiberal for Conservative Party

output for r0j_Conservative			
Female 25 or less	-0.80199	Male 25 or less	-0.19917
Female 26 to 40	-0.29421	Male 26 to 40	0.111742
Female 40 to 55	0.107557	Male 40 to 55	0.431205
Female 55 to 70	-0.06956	Male 55 to 70	0.57718
Female Above 70	-0.00649	Male Above 70	0.049452

Base on the Liberal MLR model which uses province and education as individual level, gender, and age group as group level, the post-stratification result shows that 28.70% of observations in the GSS data set votes for the Liberal party. Base on the Conservative MLR model which has the same individual and group level as the Liberal one, the post-stratification result shows that 27.56% of observations in the GSS data set will vote for the Conservative party. In conclusion, the popular vote rate for the Liberal Party is a little higher than the Conservation part, which is different from the popular vote rate in the 2019 Federal Election. In

the election, the popular vote rate for the Liberal is 33.12%, while the Conservative party is 34.34%, which higher than the Liberal party.

Discussion

In conclusion, the result of the 2019 Federal elections is dramatic. That is, the popular votes for the Conservative party are higher than the winner, the Liberal party. People argue that the Liberal party wins the trust of the government, but not the people. This paper aims to analyze if every eligible voter in Canada has voted in the 2019 Federal Election, in which the voter turnout rate is 100%, will the popular votes for Liberal higher than the Conservative party. Base on the 2019 Canadian Election Study data set, two logistic multilevel regression models with province and education as individual level, gender, and age group as group level are built. Post-stratifications with gender and age group as cells are used to simulate the proportion of electors voting for the Liberal and Conservative party respectively. According to the Post-stratification results, this study concludes that 28.70% of observations in the GSS data set votes for the Liberal party and 27.56% of observations in the GSS data set will votes for the Conservative party if the voter turnout rate is 100%. The estimated popular votes, with a 100% turnout rate, are different from the 2019 Federal Elections popular votes. One of reason contributes to the indifferent result is:

There is biased from the unequal turnout rate out rate through the provinces. According to Statistics Canada (2020) and Statista (2019), the proportion of popular votes for conservative in Alberta and Saskatchewan is the highest in Canada which is 69.2% and 64.3% respectively, and the turnout rata in Alberta (3%) and Saskatchewan (4%) increased the most compared to other provinces, and. Thus, with the increased turnout rate and the high support rate, Alberta and Saskatchewan populations increase the popular votes for the conservative party, which is one of the factors that cause the inconsistent between the MP's votes and popular votes. As the paper shows, if the turnout rate in each province is the same, in which 100%, the Liberal has higher popular votes. Thus, the past-the-post voting corrects the bias of uneven turnout rate in provinces.

Limitation &Next step

Firstly, there is a selection bias. The target population in the GSS data set (census data) are all non-institutionalized people, but the institutionalized people also have the right to vote. Secondly, there is no indigenes province in the census data. Therefore, there are three votes are missing. The result only has 335 votes in total. However, the 3 missing vote has no effect to the final result. That is, if Conservation win all of 3 votes is %, which still lower the Liberal Part. In further study, we can use the data set that indicates the electoral districts of the observation, build the models used electoral districts as the group level, and post stratified base on electoral districts as well to study how will the MP's votes change with the 100% turnout rate.

Reference

1. Chad P Kiewiet de Jonge, Gary Langer, Sofi Sinozich, Predicting State Presidential Election Results Using National Tracking Polls and Multilevel Regression with Poststratification (MRP), Public Opinion Quarterly, Volume 82, Issue 3, Fall 2018, Pages 419–446, https://doi.org/10.1093/poq/nfy023

2. Chris Erl. (2020) The People and The Nation: The "Thick" and the "Thin" of Right-Wing Populism in Canada. Social Science Quarterly 99.

3. Statistics Canada (2017). General Social Survey. Retrieved from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/hsda?harcsda4+gss31

4. Stephenson, Laura B., Allison Harell, Daniel Rubenson and Peter John Loewen (2019). The 2019 Canadian Election Study – Online Collection. Retrieved from

a) Election Canada (2020). Voter Turnout at Federal Elections and Referendums. Retrieved from https://www.elections.ca/content.aspx?section=ele&dir=turn&document=index&lang=e

5. Statistics Canada. (2019). Reasons for not voting in the federal election. Retrieved from https://www150.statcan.gc.ca/n1/daily-quotidien/200226/dq200226b-eng.htm

6. Statista. (2019). Canada election 2019: preliminary results for the popular vote, October, 22, 2019. Retrieved from https://www.statista.com/statistics/1062234/canada-election-2019-preliminary-results-for-popular-vote/

7. EKOS Politics. (2020). Update on the Political Landscape and the Issues of Race, Policing, and the Three Ms in the Canada-China Affair. Retrieved from https://www.ekospolitics.com/index.php/2020/06/update-on-the-political-landscape-and-the-issues-of-race-policing-and-the-three-ms-in-the-canada-china-affair/

8. Statistics Canada (2018). Population and Dwelling Count Highlight Tables, 2016 Census. Retrieved from https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Table.cfm?Lang=Eng&T=101&S=50&O=A

9. Samantha-Jo Caetano.(2020).gss_cleaning Codes.

10. SDA. (2017). General social survey on Family (cycle 31). Retrieved from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm

11. 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/