How could 2019 Canadian federal election be different if all population participated

Hyunseok Rha

Dec 21st, 2020

Abstract

Object: To predict the result of 2019 Canadian federal election with assumption of 100% voting participation rate and see how it would differ from the real result.

Methodology: Build multinomial multilevel regression model with the 2019 Canadian Election Study Stephenson et al. [2020] data and use poststratification with the 2016 Statistics Canada Census StatisticsCanada [2017] data to predict the election result.

Result: With the assumption that 100% of elidgible voters had voted, the gap between the number of votes that Libertal party got and Conservative party got gets closer and the number of votes NDP would get decreased.

Conclusion: It is important for every citizens to vote since the vote result can be distorted with less participation rate.

Keywords: 2019 Canadian federal election, Canadian politics.

Introduction

Election is one of the key components of democracy since it keeps a democratic country functions as it should by giving citizens the right to select their own government. Hence, every citizen who belongs to a democratic government has the right and duty to vote.

In 2019 Canadian federal election, 65.95% of the elidgeble voters out of 27.1 million have practiced the election. Liberal party got 6,018,728 votes(33.1%), Conservative party got 6,239,227 votes(34.3%) and New Democratic party got 2,903,722 votes(16.0%). And one question comes to my mind that "What if every population had participated, how would the result differ?".

To answer this question, I will use 2019 Canadian Election Study - Online Survey Stephenson et al. [2020] data and 2016 Statistic Canada Census data StatisticsCanada [2017] with multinomial multilevel regression with poststratification model to estimate the result of a scenario where 100% of eligible voters had voted. Then I will talk about how my model's result data differs from the real result, which has less participation rate, and why it is important for the citizens to practice their right by showing a possible mis-representation of their government that could be caused by less participation rate of voting which could be not a true representation of the whole citizen.

Finally, the code behind this paper can be found at https://github.com/142k/statistics

Data

The data used for this analysis are 2019 Canadian Election Study Stephenson et al. [2020] and 2016 Census StatisticsCanada [2017] data.

The 2019 Canadian Election Study Stephenson et al. [2020] is sourced from the Harvard Dataverse that can be accessed by the link in the references. The sampling populaton and frame of the survey is the entire Canadian who is eligible to vote for the 2019 Canadain federal election. The sample of the survey is the individuals who had participated the online survey. The data has been processed so it only contains relevalent data. It has total of 37012 records of individuals' year of birth, sex, age group, which province they are living in and which party that the individual are likely to vote for the 2019 Canadian federal election as you can see in Table 1. I have added columns such as age_group and sex to make bins for latter analysis. The strengths of this data is that it provides directly which political party that each individual would vote so we can directly use it to predict how the population would vote.

The 2016 Census StatisticsCanada [2017] data is sourced from Statistics Canada and can be accessed by the link in the references. Since it is a Census data, it covers most of all the Canadians which gives us the population of all Canadians. As you can see in Table 2, I have processed the census data so it would form bins and make prediction easiler in Methodology and Model part. The data covers 27,277,855 Canadians who are older than 20 years old, and I have seperated them by their province, age group and sex then count how many individual falls into each criteria. The strenght of this data is that since it is census data, it is one of the most ideal scenario we can work with that our data well represent the population that we are interested in.

And the last data is 2019 Canadian federal election result data. This data will not be used for modeling nor predictions, but instead it will be used to see the difference between the prediction and truth. I have modified the data in a way that only counts Liberal party, Conservative party and New Democratic party as you can see at Table 3.

Table 1: 2019 Canadian Election Study Data

yob	gender	province_territory	sex	age_group	which_party_to_vote_for
1989	A woman	Quebec	female	30 to 34 years	Green Party
1998	A woman	Quebec	female	20 to 24 years	Don't know/ Prefer not to answer
1998	A man	Ontario	male	20 to 24 years	Conservative Party
1999	A woman	Ontario	female	20 to 24 years	Liberal Party

Table 2: 2016 Canadian Census Data

province_territory	province_num	age_group	sex	count
Newfoundland and Labrador	5	20 to 24 years	male	13915
Newfoundland and Labrador	5	20 to 24 years	female	13785
Newfoundland and Labrador	5	25 to 29 years	male	14095

Table 3: 2019 Canadian Federal Election Result

name of party	number of votes	ratio(%)
Liberal Party	6018728	39.69698
Conservative Party	6239227	41.15130
ndp	2903722	19.15172

Methodology and Model

The analysis consist two steps.

First, we will work with 2019 Canadian Election Study Stephenson et al. [2020] data to build multilevel multinomial logistic regression model. The formual behind the regression is as following.

$$Pr(Y_{i,k}) = Pr(Y_i) = k|x_i; \beta_0, \beta_{age\_group}, \beta_{sex}$$

Where $\beta_0 = W_{province}$.

I will explain the formula and model in high level. With the 2019 Canadian Election Study data Stephenson et al. [2020], we would like to know for each individual who belongs to certain province, sex and age_group what it the probality of each individual would vote for a certain party. On top of that, I assume that the individuals from the same province tends to act similarly compare to who is from a different province. And that is why we have $\beta_0 = W_{province}$ to give the model another level. From the formula $Pr(Y_{i,k})$ represents the probalility of vote for a certain party i given information k.

With this model, we can predict how each factor would affect individual's voting result. And we will use this model to perform poststratification with 2016 Canadian Census StatisticsCanada [2017] data and that would be the second step.

For the poststratification, I have already modified the 2016 Canadian Census StatisticsCanada [2017] data so it would contain the facotrs that the multilevel multinomial logistic regression model uses. For each record of 2016 Canadian Census StatisticsCanada [2017] data, we will predict the probalility of voting for all the parties we have. Then, we will choose a party with the maximum probability and assing it to the corresponding records. This way, we can predict which party would win for each bin and how many votes they would have.

Results

The predicted number of votes for each party would get which is based on the multilevel multinomial logistic regression model with poststratification from the previous section can be found with Table 4.

Also you can see with Table 5. that the gap of ratio of votes between Liberal party and Conservative party has been decreased into less than 1% with the predicted result. Also the ratio of votes that NDP would get decreased by about 15% with the predicted result.

The interesting thing with the prediction is that the result is more extreme and the popular parties get more votes and less popular party tends to get less votes.

Table 4: Predicted result

name of party	number of votes	ratio(%)
Liberal Party	12971040	47.55154
Conservative Party	13208625	48.42252
ndp	1098190	4.02594

Table 5: Prediction versus real result

name of party	predicted ratio(%)	real ratio(%)	difference(%)
Liberal Party	47.55154	39.69698	7.854559
Conservative Party	48.42252	41.15130	7.271222
ndp	4.02594	19.15172	-15.125781

Discussion

First, I want to discuss the limitation of my study. Initially, I wanted to build bins and model with more factors such as individual's family income, education level, ethnicity and seperate them at city level instead of province level. With this approach, I believe the prediction would be more accurate and fine detailed. However, I could not do it since the Census data I could get is already aggreated at province, age group and sex level and I could not find a way to seperate the data with more detailed categories. Another limitation is, with my method of choosing which party would get the most of the vote, it only can replicate the system of the winner takes it all that can leads to ignoring minorities or a case of really close win.

For the next step, I would like to collect Census data that has more fine detail interms of possible bin category and improve my multilevel multinomial logistic regression model so it can handle more factors. On top of that, the way of choosing winner party can be improved as well. Instead of following the winner takes it all approach, I can calculate the number of votes by taking the probability of each party gets a vote as a weight.

Finally, as you can see from the result of this study, the voting result can be differ as more voters participate and it allows the democratic government to work at its best. So please excersice your right to vote, and encourage others to do so as well for the better future.

References

StatisticsCanada. Census Profile, 2016 Census, Canada, provinces and territories, 2017. URL https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/download-telecharger/comp/page_dl-tc.cfm?Lang=E.

Laura B Stephenson, Allison Harell, Daniel Rubenson, and Peter John Loewen. 2019 Canadian Election Study - Online Survey, 2020. URL https://doi.org/10.7910/DVN/DUS88V.