Analysis of how turnout affect the results of the 2019 Canadian election

Xinran Gu

09/12/2020

Abstract

This report is focused on how the turnout rate will influence the results of the 2019 Canadian election. According to Election Canada, only 67% of legal Canadian citizens participated in the 43rd Canadian general election. In order to investigate whether the results will change if all legal Canadian citizens have voted in 2019 election, a multilevel regression with post-stratification models with two independent variables age and gender is conducted in the report. The data set Canada Election study 2019 was used to fit the logistic regression model. Then the data set 2017 General Social Survey is used for post-stratification technique. The results of the models suggests that the liberal party will win the election if all citizens who are over 18 have voted in the 2019 election. Therefore, the conclusion of the report is that the results of the election will not change because of the turnout.

keywords

election, election turnout, multilevel regression model, post-stratification, logistic model,multilevel regression with poststratification model, 43rd Canadian general election

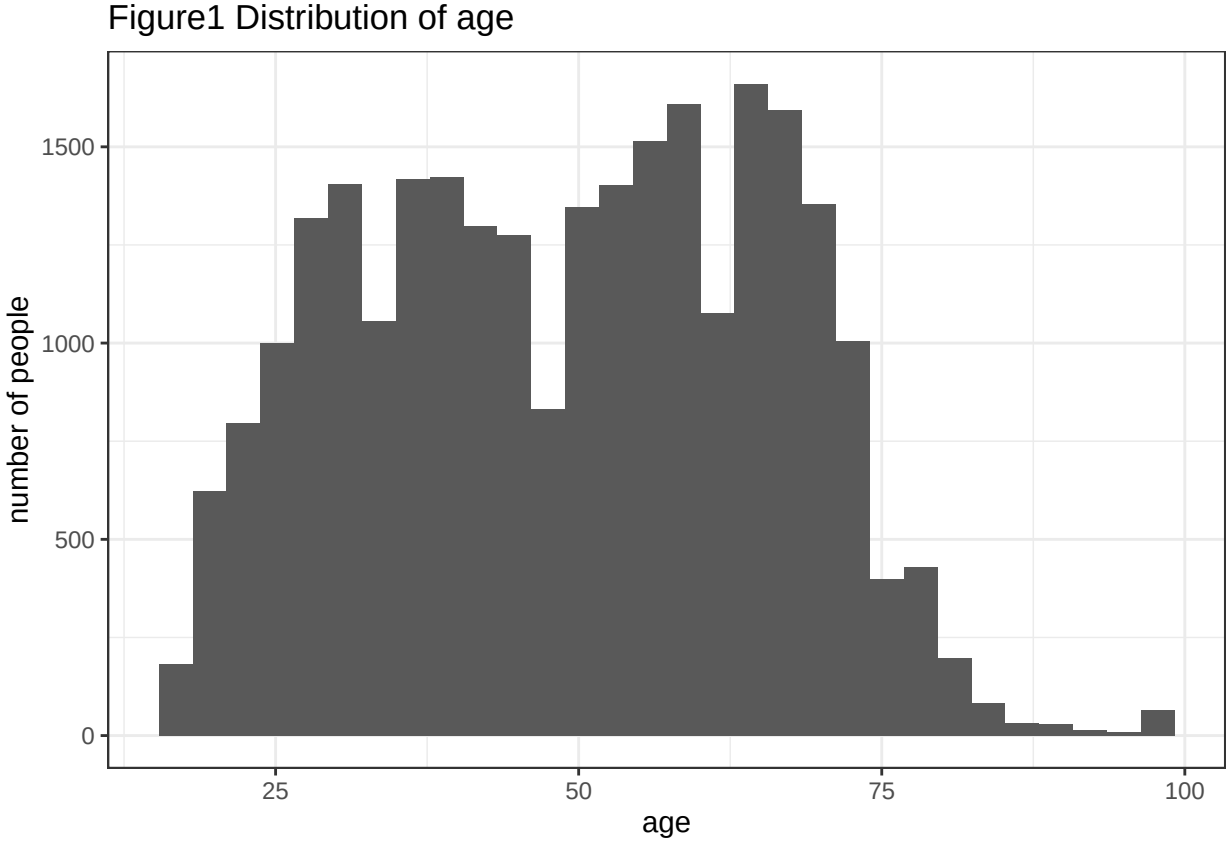
Introduction

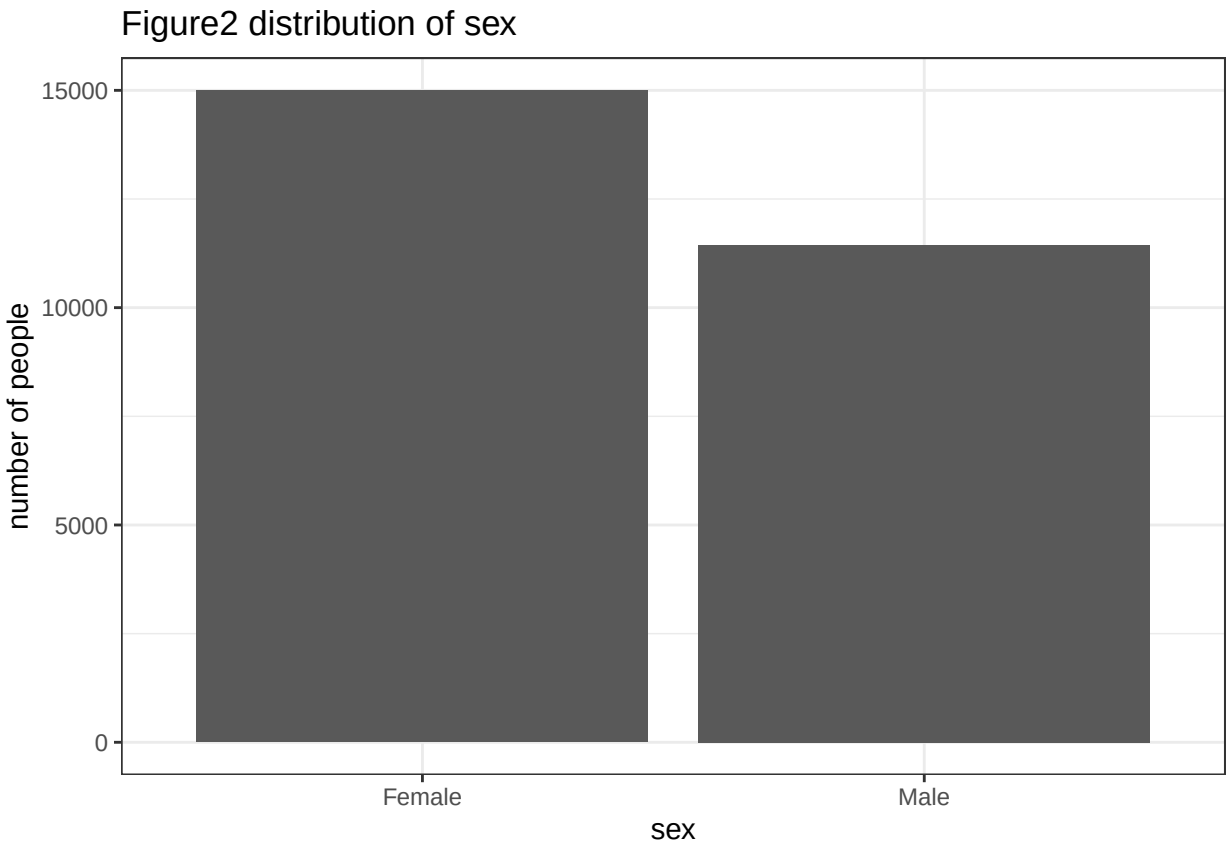
Federal election is a essential process for Canadian to decide the future Prime Minister.This is a important decision to make as it will influence the country for years. The latest federal election in Canada was held on October 21, 2019. In Canada, all Canadian citizens who are older than 18 are allowed to vote. However, according to the official voting results, the turnout rate is only 67% in the 2019 election. This means about 33% of the population in Canada did not participate in the decision. Therefore, this report is interested at whether the results would be different if all citizens who are over 18 had voted in the 2019 election. The two major parties in the 2019 Canadian are Liberal Party and Conservative Party which got 157 seats and 121 seats respectively. Since the third political party Bloc Québécois only won 32 seats from the 2019 election, the results of the election will not change due to the turnout. Hence, this report will be only focused on Liberal Party and Conservative party. The two data sets that the report is based on are Canada Election study 2019 and the 2017 General Social Survey. The specific discussion of the data sets will be introduced in the Data section below. A multilevel regression with poststratification model(MRP model) will be used in this report.Hence the Canada Election study will be first used to fit a logistic model. The multilevel logitic regression model will allow us to identify how different factors are affecting the decision of voting. Then, the General Social Survey will be used to conduct post-stratification technique.Both of the model and post-stratification will be further discussed in the Model Section of the report. In the end, this report will analyze the results of the multilevel regression model with poststratification model to drive to a conclusion whether the turnout will affect the election or not.

Methodology

Data

The data that I used for multilevel regression model in this report is the compaign period survey from Canada Election study 2019. The campaign period survey is conducted in the form of online survey through September 13 to October 21 in 2019. This pre-election survey contains total 37,822 participants. All of these participants are either Canadian citizens or permanent residents who are over 18.The sample were collected through Qualtrics(A survey-based tool), with target stratified by age, gender and region. As figure 1 showing below, the sample data were collected with the aim of 28% of respondents aged 18-34, 33% aged 35-54 and 39% aged 55 and higher. The survey was trying to be more balance.Yet, according to statistics Canada, the number between aged 45-49 and 50-54 should be around the same, there should not exist a gap as the figure1 shown. Moreover, we should notice that the distribution of sex is not balanced as figure2 shown.Even though the study was aimed to collect data from 50% male and 50% female, the actual data has more female respondents. This the reason why we need to use the post-stratification technique. The reason why I did not use the data from the post election survey of the same study is because the post election survey only contains 10337 cases. Even tough the post election survey is more reliable than the campaign period survey as it records the actual vote decision, the difference of the sample size make it hard to represent the whole population.





The census data that I will use for post-stratification in this report is the 2017 General Social Survey(GSS) on the family. According to the user guide of 2017 General Social Survey, data collected by General social Survey comprised two components.The core content is designed to measure changes in society related to well beings. The classification variables such as age and gender which I used in this report wil be used to create strata for post-stratification. The target population of GSS includes all persons who are 15 or older in Canada.Yet, it excludes full-time residents of institution and residents in Yukon,Northwest Territories and Nunavut.The frame population consists of lists of telephone number in use available to Statistics Canada and lists of address registered within the ten provinces. Each of the ten provinces were divided into strata and then a simple random sample without replacement was performed.Respondents who are 15 years old or older are eligible and were interviewed through telephone. If the phone call was not reposed, interviewers were made numerous call backs. The strength of the data is that it used stratified random sampling, so it has a acceptable sampling variability. Yet, one significant drawbacks of GSS is that it is hard to generalize the results to the whole country. The sample size of GSS is 20,602 which is relatively large compared to other survey. However, the estimate population of Canada is 38,005,238 which is much larger. As the report is focused on the citizens who are legal to vote, all participants who are under 16 in GSS was removed from the data.This is because people who are under 18 are not allowed to vote in the election, and the remaining participants in the data should all be over 18 years old in 2019.

Model

This report is focusing on how the turnout will affect the results of 2019 Canadian election and a logistic regression model will be used for analysis. In the logistic regression model, the two independent variables are age and sex, age is a numerical variable and sex is a categorical variable. The dependent variable is the proportion of votes that liberal party will get.The reason why I choose these two independent variables is because I believe people with different age and gender will encounter different situations in society which will affect their political party preference.Also, as the census data is from 2017,other variables like income

and marriage status have higher possibility to change as the time pass.Hence choosing age and gender as the independent varaibles will ensure the accuracy of the analysis.The usage of a logistic model will allow me to predict the proportion of votes that liberal party will get. If we use a linear model, we will not be able to obtain a probability as output. Moreover, the results of a election is binary,the results is either the potical party will win or not. Therefore, logistic model is suitable to use in this report.The logistic regression model I am using is expressed as:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{SEX} + \epsilon$$

The $log(\frac{p}{1-p})$ on the right-hand side represents the proportion of voters who will vote for the liberal party and it could also be called as log odds.The numerator p is the probability liberal party win the election and the denominator $1-p)$ is the probability liberal party lose the election. Since the model fitting was done in R using the glm function, R will automatically read "Male" as 1 and "Female" as 0 for the variable SEX.On the left hand side, β_0 represents the intercept of the model, and it is the probability of a female voting for the liberal party at age 0. Then β_1 and β_2 in the model represent the change in log odds for every one unit increase in x_{age} and x_{gender} when other variable hold constant.

Post-Stratification

A post-stratification technique is performed in this report to enhance the estimation of the proportion that liberal party will get.Post-stratification is a technique that I divided sample data into strata after we randomly selected data from the population. We often use this technique when it is hard place the data into their correct strata until I finish sampling. Also, this technique allows us to minimize bias caused by underrepresented group in the population as it adjusting sampling weights. In this report, I create cells based on different ages and genders. For example, there will be 2 cells for age 18, one is for female who is 18 and the other one is for male who is 18. In this analysis, there are total 130 cells.Then, I will use the logistic model described above to estimate the proportion of voters in every cells. After that, I will use the formula $\hat{y}^{ps} = \frac{\sum N_j \hat{y}_j}{\sum N_j}$ to estimate the proportion of voters who will vote for liberal party. The numerator $\sum N_j \hat{y}_j$ is the sum of estimates I got from the logistic model for each cell times the population of each cell. The denominator $\sum N_j$ is the population of all cells.

Results

liberal party table	Estimate	p-value
$\hat{\beta_0}$	-0.9146	p<0.0001
$\hat{\beta_1}$	0.0055	p<0.0001
$\hat{\beta_2}$	-0.0840	0.0017

conservative party table	Estimate	p-value
$\hat{\beta_0}$	-1.4295	p<0.0001
$\hat{\beta_1}$	0.0115	p<0.0001
$\hat{\beta_2}$	0.3042	p<0.0001

The liberal party table above summarize the logistic model for estimating proportion of people voting for the liberal party. From this table, I could obtain the model representation: $\hat{y} = -0.9146 + 0.0055 x_{age} - 0.084 x_{SEX}$.Based on this logistic model which accounted for age and sex, I have performed a post stratifica-

tion analysis. The result of the analysis showed that the proportion of votes which will go to the liberal party is 0.3395. The conservative party table summarize the logistic model for estimating proportion of people voting for the conservative party.The model is represented as: $\hat{y} = -1.4295 + 0.0115x_{age} + 0.3042x_{SEX}$. The result of the analysis showed that the proportion of votes which will go to the conservative party is 0.3362.

Discussion

Summary and Conclusion

In the report, I have first built two logistic models to predict the proportion of people voting for the liberal party and the conservative party using two variables age and gender.According to the results section, the model for liberal party is $\hat{y} = -0.9146 + 0.0055x_{age} - 0.084x_{SEX}$ and the model for conservative party is $\hat{y} = -1.4295 + 0.0115x_{age} + 0.3042x_{SEX}$.

This means for liberal party when a person's age increased by 1 year with other predictors hold constant,the probability of this person to vote liberal party increase by 0.0055. Also, females are more willing to vote for liberal party as the estimate for sex is -0.084 which is negative.Moreover, as all p values for the coefficients are smaller than 0.05,I could conclude that both age and gender have a significant impact to the proportion that the liberal party will get. For conservative party when a person's age increased by 1 year with other predictors hold constant,the probability of this person to vote liberal party increase by 0.0115. Also, males are more willing to vote for conservative party as the estimate equals to 0.3042 which is positive.Same as the model for liberal party, all p values are smaller than 0.005 which suggests both age and sex are significant.

Then I have used this model to do a post stratification analysis by diving sample data into 130 cells based on age and gender. I have adjusted the weights of each cell by the cell's population and then get $\hat{y}^{ps}$. The results showed that the estimated proportion of voters will vote liberal party is 0.3395 and the estimated proportion of voters will vote for conservative party is 0.3362. Hence, the conclusion is that the turnout will not change the result of the election.

Weaknesses and nex steps

The results of the report, the difference between proportion of votes will go to liberal party and conservative party are relatively close. Hence, the results of the analysis might be different if method improve in the future.

The survey data that I used in the report is collected during the campaign period but not after the election. This means participants of the survey might not actually vote the party they answered here. Also, the census data that I used for post-stratification is not large enough to represent all Canadian citizens who are over 18. This means the balance of our data is not sufficient. In the future, a survey that contains more respondents and asks who did the voters vote in 2019 should be conducted.Also, the post-stratification technique with a larger census data will also improve the statistic power of the model.

Moreover, the logistic models in the report are naive. The factors that affects the results of Canadian election are complicated. In order to improve the models, more independent variables should be added.

Reference

1.The Pennsylvania State University. (2020). 6.3 - Poststratification and further topics on stratification: STAT 506. Retrieved November 01, 2020, from https://online.stat.psu.edu/stat506/lesson/6/6.3

2.Alexander, R. (2020, May 17). Telling Stories With Data. Retrieved December 16, 2020, from https://www.tellingstorieswithdata.com/

3.Government of Canada, S. (2020, December 16). Population estimates. Retrieved October 16, 2020, from https://www150.statcan.gc.ca/n1/en/subjects/population_and_demography/population_estimates

4.General Social Survey Cycle 31:Families Public Use Microdata File Documentation and User's Guide. (2020). Ottawa: Statistics Canada.

5.2017 General Social Survey: Families Cycle 31 Public Use Microdata File. (2020). Ottawa: Statistics Canada.

6.Alexander, R., & Caetano, S. (2020, October 22). gss_cleaning.R.

7.Clarke, S., & Levett, C. (2019, October 23). Canada election 2019: Full results. Retrieved December 19, 2020, from https://www.theguardian.com/world/2019/oct/22/canada-election-2019-full-results

8.Duffin, P., & 5, O. (2020, October 05). Canada - population, by gender and age 2020. Retrieved December 20, 2020, from https://www.statista.com/statistics/444858/canada-resident-population-by-gender-and-age-group/

9.:Stephenson, Laura B., Allison Harell, Daniel Rubenson and Peter John Loewen. The 2019 Canadian Election Study – Phone Survey. [dataset].

Appendix

1.Code and data supporting this analysis is available at:"https://github.com/LiviaSta/Effect-of-turnout-for-2019-canadian-election".