# The Results of 2019 Canadian Federal Election if 'Everyone' had voted

Huiyi Lu(1004754615)
Code and data supporting this analysis is available at https://github.com/huiyilu99/304finalproject

12/22/2020

## Contents

# 1 Abstract

In the 2019 Canadian Federal Election, the Liberals won by a very small margin. In this study, we generated the logistic regression model rom the CES2019_web data set, and then applied to the 2013 General Social Survey (GSS) on Social Identity using post-stratification to investigate the results if turnout ratio was 100%. In particular, we analyzed four independent variables (age, sex, education, and province) that would potentially affect people's voting decisions. From the model, we can predict that the Liberals still won if everyone had voted.

# 2 Keywords

Canadian Federal Election, Turnout, Vote, Observational Study

# 3 Introduction

The Liberal Party, led by Prime Minister Justin Trudeau won the 2019 Canadian Federal Election by receiving 33.12% of the popular vote and 157 seats. The Conservatives won 34.34% of the popular vote and 121 seats. The gap between the Conservatives and the Liberals was quite small.The Liberals still gained enough seats to form a minority party. However, it was the second time in Canadian history that a governing party formed a government with less than 35% of the popular vote. Most of the votes that the Liberals lost lean towards the Conservatives. Liberal party lost 20 seats, while its main opposition rivals the Conservatives gained 26(Canada election 2019: full results).

Only 67% of the voters turned out on the date of the election (Voter Turnout at Federal Elections and Referendums). The voter turnout ratio in 2019 was consistent with the past data. The question arises is would the 2019 Canadian Federal Election be different if all the voters had turned out on the election day. This report aims to generate a MRP model to figure out the importance of turnout.

Two data sets will be used to investigate how the 2019 Canadian Federal Election would be different if all the voters had turned out. The Methodology section(Section 4) describes the data and the model that were used to perform the voting analysis. Results of the voting are provided in the Results section(Section 5). Weakness of the analysis, conclusion and improvements could be done are presented in the Conclusion section(Section 6).

# 4 Methodology

## 4.1 Data

Two datasets, CES2019_web and 2013 GSS were used in this analysis. CES2019_web was obtained from the R package cesR, downloading from github. The R package cesR contains multiple Canadian Election Study datasets. CES2019_web was chosen since the analysis aims to investigate the 2019 Canadian Federal Election. CES2019_web was an online survey. Its target population was all Canadian citizens and permanent residents. Its frame population was all Canadian citizens and permanent residents who had access to the survey. Its sample population was Canadian citizens and permanent residents who had a right to vote, and saw and answered the survey. The 2013 GSS dataset was obtained from the U of T library. Data for 2013 General Social Survey (GSS) on Social Identity (SI) was collected from June 2013 to March 2014. The target population for the survey included all persons 15 years of age and older in Canada, excluding residents of the Yukon, Northwest Territories, and Nunavut, and. full-time residents of institutions. The frame population of the survey included all persons whose telephone numbers in use (both landline and cellular) were available to Statistics Canada from various sources (Telephone companies, Census of population, etc.); and who had a

record in the Address Register (AR) within the ten provinces. 2013 GSS was cleaned using the code provided in the previous assignment.

To correspond CES2019_web variables to the 2013 GSS variables for modeling and predictions, I reconstructed 4 variables, *province*, *age_group*, *edu*, and *sex* in both data sets. The variables *vote_liberal* and *vote_conserv* are exclusively created in the survey data set.

-age

Initially, CES2019_web dataset only included year of birth of the respondents, and 2013 GSS dataset included the age variable. I classify them into 5 groups that would contain all the eligible ages. Specifically, we identify the young (18 - 29), and the old (66 - 99[1]), and the retirement (50 - 65). The remaining ages are grouped with a 10-year difference.

<age30-39, age18-29, age50-65, age66-99, age40-49>.

-province

The categories in CES2019_web included Yukon and Nunavut, but these two provinces were not included in the census data. Only 0.2% of the respondents were from these two provinces, therefore, cases from Yukon and Nunavut were removed.

<Quebec, Ontario, British Columbia, Alberta, Saskatchewan, Manitoba, Newfoundland and Labrador, Nova Scotia, Prince Edward Island, New Brunswick>.

-education

I reconstructed *edu* by grouping the choices under *education* in 2013 GSS and *cps19_education* in CES2019_web. In 2013 GSS, Less than high school diploma or its equivalent is classified into Primary Education; High school diploma or a high school equivalency certificate is classified into Secondary Education; Trade certificate or diploma, and College, CEGEP or other non-university certificate or diploma, and University certificate or diploma below the bachelor's level, and Bachelor's degree(e.g. B.A., B.Sc., LL.B.) are classified into Post-secondary Education; University certificate, diploma or degree above the bachelor's level is classified into Post-graduation Education. in CES2019_web, Some secondary/ high school, and Completed elementary school, and No schooling were classified intoPrimary Education; Some university, Completed secondary/ high school, Some technical, community college, CEGEP, College Classique were classified into Secondary Education; Bachelor's degree, Completed technical, community college, CEGEP, College Classique were classified into Post-secondary Education; Professional degree or doctorate, and Master's degree were classified into Post-graduation Education; Don't know/ Prefer not to answer is classified into NA.

<Post-graduation Education, Secondary Education, Primary Education, Post-secondary Education>.

-sex

Initially, in the 2013 GSS data, the variable has the following 2 categories: <Female, Male>. In the CES2019_web data, the variable contains the following 3 categories: <A woman, A man, Other (e.g. Trans, non-binary, two-spirit, gender-queer)>. To match the categories, I reclassify the observations into 2 categories.

<Female, Male>

-vote_liberal

In the CES2019_web data set, variable *cps19_votechoice* showed the respondents' preferences towards parties. In particular, the Liberal Party and the Conservative Party are two major parties in Canada. For simplicity, I reclassify the observations into 2 categories, vote for the Liberal Party(1) or not(0). Similar process applied to *vote_conserv*.

<0, 1>.

[1]The maximum age in both CES2019_web and 2013 GSS

I chose to remove all the observations with NA in all the variables selected. I realized that the model will remove those NA observations automatically and there is no point to keep those observations if the model does not need them. Also, in further analysis, the model diagnostics require the same amount of observations for the data as well as the model used.

## 4.2 Model

I am interested in predicting the vote results of the 2019 Canadian Federal Election if 'everyone' had voted. The coding is done in R. In order to make the prediction, I employed the post-stratification technique on the census data. Before predictions, we estimate the probability of each post-stratification cell voting for the Liberals and the Conservatives separately with the MRP model. The MRP model was generated from the survey data with selected variables.

A logistic regression model was used in R to calculate the probabilities, which is specific for binary response modeling. A frequentist method was applied since there is no such precise prior information available to use for the Bayesian inference. Wrong prior information may lead to an opposite result.

The following mathematical notation is the logistic regression model we create in R with four predictors, *edu*, *age_group*, *province*, and *sex*. *age_group*, a categoricla variable, is used in the model other than age, a discrete variable, since age and log odds of voting for the Liberals are not linearly related, but respondents within the same age group may have similar opinions towards the political parties. All the variables except sex have small p-values in the model for voting for the Liberals. The variable sex has a small p-value in the model for voting for the Conservatives. Therefore, all the variables selected are statistically significant in this model.

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_{edu} X_{edu} + \beta_{age_group} X_{age_group} + \beta_{province} X_{province} + \beta_{sex} X_{sex}$$

$p_i$: the probability of the respondent voting for the Liberals in the 2019 Canadian Federal Election

$\beta_0$: the intercept value; represents the log odds of an 18-29 ages old female being in Alberta and receives post-graduation education. These information in each variable are treated as the reference groups.

$\beta_{edu}$: the slope values for education; represent the difference of preference between people receives post-graduation education and other levels of education

$X_{edu}$: = 1 if the respondent's education level is not post-graduation

$\beta_{age_group}$: the slope values for age; represent the difference of preference between people aged 18-29 and the other 4 age groups

$X_{age_group}$: = 1 if the respondent's age is in range other than 18 - 29

$\beta_{province}$: the slope values for provinces; represent the difference of preference between people live in Alberta and other provinces

$X_{province}$: = 1 if the respondent's age is in provinces other than Alberta

$\beta_{sex}$: the slope values for gender; represent the difference of preference between females and males

$X_{sex}$: = 1 if the respondent is identified as a male

Based on the above variable information, we can effectively predict a person's probability of voting for the Liberals. The variables separate the information into different categories that allow each post-stratification cell to be classified precisely and give back a relatively correct probability. In this model, there were 400 post-stratification cells. An alternative model may be Bayesian regression model, but it requires prior information.

### 4.3 Post-Stratification

The post-stratification method partitions the census data into demographic cells according to the chosen variables. Usually the number of cells should equal all possible combinations of the variables. In this model, the census data set was separated into 400 different demographic cells based on the variables *education*, *age_group*, *province*, and *sex*. The response variables, *vote_liberal* and *vote_conserv*, are estimated within each demographic cell. The vote for each party was then calculated by weighting each cell based on its relative proportion in the population.

## 5 Results

Table 1: Predicted Vote for Liberasl v.s. Conservatives

| liberals | conservatives |
|---|---|
| 0.2876 | 0.2748 |

Referring to **Table 1(Predicted Vote for Liberals v.s. Conservatives)**, I estimate that the Liberals will likely receive 28.8% of the total number of votes, and the Conservatives will likely receive 27.4% of the total number of votes. This is based off our post-stratification analysis of the proportion of voters in favour of the Liberals and the Conservatives modelled by logistic models, which accounted for *edu*, *age_group*, *province*, and *sex*. *age_group*.

Referring to Appendix 7.1(**Table 2(Summary of Logistic Regression for Liberals)**) and 7.2(**Table 3(Summary of Logistic Regression for Conservatives)**), most variables have small p-values in both the Liberals and the Conservatives models, which means these variables are statistically significant. However, sex in the Liberals model has a large p-value and in the Conservatives model has a small p-value. The variables sex is still included since I belive sex is an important factor when making decisions. Overall, according to the prediction of the vote, I believe that the Liberals would still have a higher chance of winning if all the voters had turned out. The results of the election would not be different.

## 6 Discussion

### 6.1 Summary

Two data sets, CES2019_web and 2013 GSS were used in this analysis. CES2019_web was obtained from the R package cesR, downloading from github. The R package cesR contains multiple Canadian Election Study data sets. The 2013 GSS data set was obtained from the U of T library. Data for 2013 General Social Survey (GSS) on Social Identity (SI) was collected from June 2013 to March 2014. The CES2019_web survey data consists of some personal information and the interviewees' political opinions, while the 2013 GSS consist of some personal information and some variables on voter participation. Four variables were cleaned and created (province, age_group, edu, and sex) in both data sets to assure the consistency of categories under each categorical variable. Cases with NA's were removed from the data sets. Logistic regression was used on survey data to assess the impact of each independent variable I chose. A post-stratification method was adopted to partition the census data into different cells according to the selected variables. Thus the census data is split into 400 different cells. I estimate the proportion of voters in each province, then weigh each proportion estimate by the respective population size of that province and sum those values and divide that by the entire population size in the census data to calculate the vote. The results are that the Liberals will likely receive 28.8% of the total number of votes, and the Conservatives will likely receive 27.4% of the total number of votes, if all the voters had turned out. This model might be biased since the census data was

only split into 400 different cells. More cells can reduce the bias. Computation errors might occur during the record of the survey. Interviewees might answer the survey or the census without thinking thoroughly.

## 6.2 Conclusion

Based on the statistical analysis conducted in this report, it is estimated that the Liberals will likely receive 28.8% of the total number of votes, and the Conservatives will likely receive 27.4% of the total number of votes, if all the voters had turned out, referring to **Table 1 (Predicted Vote for Liberals v.s. Conservatives)**. The results are consistent with what we see today, Justin Trudeau is still the Prime Minister. It comes to the conclusion that a higher voter turnout ratio does not change the results when it is high already. The gap between the Liberals and the Conservatives is only about 1% in this model. Amid the COVID-19 pandemic, the Conservatives should fully utilize this chance to earn an advantage in next election.

Voter turnout ratio has been declining in Canadian federal elections since 1988(The Democracy Defibrillator: The Decline of Canadian Voter Turnout in Federal Elections, and Suggestions for Revitalisation). Voter turnout in Canada is lowest for young voters(The Roots of Social Capital: Attitudinal and Network Mechanisms in the Relation between Youth and Adult Indicators of Social Capital). A general decline in electoral participation among the under-35 population has been observed in many democratic countries around the world, especially in Canada. Decline in turnout ratio represents low engagement in politics.

## 6.3 Weakness

In this report, I estimate the proportion of votes in each province, then weigh each proportion estimated by the respective population size of that province, and sum those values and divide that by the entire population size in the census data to calculate the vote. However, Canada's electoral system is referred to as a "first past the post" system. The candidate with the most votes in a riding wins a seat in the House of Commons and represents that riding as its Member of Parliament (MP). Our model may not be consistent with the electoral system. The census data is split into 400 different cells, however, bias may still exist since 400 cells are usually not enough. In terms of the data sets used in the analysis, computation errors might occur during the record of the survey. Interviewees might answer the survey or the census without thinking thoroughly.

## 6.4 Next Step

To enhance the accuracy of the predictions using a logistic model, we can gather more information and find more significant variables between the two data sets for further analysis. Also, We can also evaluate using other measurements such as Bayesian Information Criterion (BIC) and Residual Sum of Squares (RSS) for model selection.

# 7 Appendix

## 7.1 Model for Liberals

Table 2: Summary of Logistic Regression for Liberals

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -1.42 | 0.07 | -21.38 | 0.00 |
| age_groupage30-39 | -0.01 | 0.05 | -0.14 | 0.89 |
| age_groupage40-49 | -0.05 | 0.05 | -1.06 | 0.29 |
| age_groupage50-65 | -0.01 | 0.04 | -0.28 | 0.78 |
| age_groupage66-99 | 0.15 | 0.05 | 3.43 | 0.00 |
| sexMale | 0.01 | 0.03 | 0.35 | 0.73 |
| provinceBritish Columbia | 0.67 | 0.06 | 11.18 | 0.00 |
| provinceManitoba | 0.62 | 0.08 | 7.98 | 0.00 |
| provinceNew Brunswick | 0.93 | 0.09 | 9.90 | 0.00 |
| provinceNewfoundland and Labrador | 1.29 | 0.10 | 12.85 | 0.00 |
| provinceNova Scotia | 1.16 | 0.08 | 13.84 | 0.00 |
| provinceOntario | 1.01 | 0.05 | 20.16 | 0.00 |
| provincePrince Edward Island | 0.92 | 0.19 | 4.75 | 0.00 |
| provinceQuebec | 0.84 | 0.05 | 15.72 | 0.00 |
| provinceSaskatchewan | -0.26 | 0.10 | -2.53 | 0.01 |
| eduPost-secondary Education | -0.26 | 0.04 | -6.69 | 0.00 |
| eduPrimary Education | -0.64 | 0.07 | -8.81 | 0.00 |
| eduSecondary Education | -0.46 | 0.04 | -11.21 | 0.00 |

## 7.2 Model for Conservatives

Table 3: Summary of Logistic Regression for Conservatives

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -0.73 | 0.06 | -11.27 | 0 |
| age_groupage30-39 | 0.32 | 0.05 | 6.20 | 0 |
| age_groupage40-49 | 0.47 | 0.05 | 9.23 | 0 |
| age_groupage50-65 | 0.57 | 0.05 | 12.47 | 0 |
| age_groupage66-99 | 0.63 | 0.05 | 12.72 | 0 |
| sexMale | 0.48 | 0.03 | 17.75 | 0 |
| provinceBritish Columbia | -1.26 | 0.05 | -24.50 | 0 |
| provinceManitoba | -0.83 | 0.07 | -12.46 | 0 |
| provinceNew Brunswick | -1.42 | 0.10 | -14.75 | 0 |
| provinceNewfoundland and Labrador | -1.65 | 0.12 | -14.25 | 0 |
| provinceNova Scotia | -1.79 | 0.10 | -18.58 | 0 |
| provinceOntario | -1.22 | 0.04 | -30.77 | 0 |
| provincePrince Edward Island | -1.88 | 0.24 | -7.75 | 0 |
| provinceQuebec | -2.01 | 0.05 | -41.80 | 0 |
| provinceSaskatchewan | -0.34 | 0.07 | -4.89 | 0 |
| eduPost-secondary Education | 0.27 | 0.04 | 6.10 | 0 |
| eduPrimary Education | 0.34 | 0.07 | 4.66 | 0 |
| eduSecondary Education | 0.35 | 0.05 | 7.71 | 0 |

# 8 References


1. Clarke, Seán, and Cath Levett. "Canada Election 2019: Full Results." The Guardian, Guardian News and Media, 23 Oct. 2019, www.theguardian.com/world/2019/oct/22/canada-election-2019-full-results.

2. "Voter Turnout at Federal Elections and Referendums." Elections Canada, www.elections.ca/content.aspx?section=ele&

3. Ouellet, Andre Real, "The Democracy Defibrillator: The Decline of Canadian Voter Turnout in Federal Elections, and Suggestions for Revitalisation" (2019). Major Papers. 77. https://scholar.uwindsor.ca/major-papers/77

4. &, Dietlind Stolle1, and Marc Hooghe2. "The Roots of Social Capital: Attitudinal and Network Mechanisms in the Relation between Youth and Adult Indicators of Social Capital." Acta Politica, Palgrave Macmillan UK, 13 Dec. 2004, link.springer.com/article/10.1057/palgrave.ap.5500081.

5. Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.

6. Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

7. Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

8. Stephenson, Laura, Allison Harrel, Daniel Rubenson and Peter Loewen. Forthcoming. 'Measuring Preferences and Behaviour in the 2019 Canadian Election Study,' Canadian Journal of Political Science.

9. Kassambara, and Michael U. "Logistic Regression Assumptions and Diagnostics in R." STHDA, 11 Mar. 2018, www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/.

10. Paul A. Hodgetts and Rohan Alexander (2020). cesR: Access the CES Datasets a Little Easier.. R package version 0.1.0.

11. Technology, Advancing Knowledge through. Computing in the Humanities and Social Sciences. www.chass.utoronto.ca/.