Liberals estimated to have won 34.0% of the popular vote in 2019 Canadian Federal Election if all eligible voters voted

Amy Chen

22 December 2020

Abstract

In this paper, a multilevel logistic regression model was built using the 2019 Canadian Election Study (CES) dataset (Stephenson et al.) followed by post-stratification analysis using the Census of Canada 2016 dataset (Census of Canada, 2016) to determine the popular vote for both the Liberals and Conservatives in the 2019 Canadian Federal Election if all eligible voters voted. It was found that at 100% voter turnout the Liberals would have won 34.0% of the popular vote and the Conservatives with 33.7% of the popular vote. This differs from the actual result where the Liberals had 33.12% of the popular vote and the Conservatives had 34.34% of the popular vote, highlighting the imporatance of voter turnout.

Keywords: 2019 Canadian Federal Election, Voter Turnout, Popular Vote, Multilevel Regression and Post-stratification, MRP

*Code and data supporting this analysis is available at: https://github.com/Amy0623/Final_project

Introduction

Historically, voter turnout in the Canadian Federal Elections was never 100%, with the highest voter turnout to be 79.4% in the 1958 election (Voter Turnout at Federal Elections and Referendums). In recent years, voter turnout is only in the 60s (Voter Turnout at Federal Elections and Referendums). However, this issue is not unique to only Canada. The recent US Federal election cited the highest voter turnout since 1990, however the percentage was still only 66.20% (2020 United States presidential election). This means about a third of all eligible voters are not voting in these elections. Thus, by analysing what would happen if these voters also voted in elections, it will give insight into the importance of voter turnout.

The 2019 Canadian Federal Election will be used to analyse the importance of voter turnout. In 2019, the Liberals won the Canadian Federal Election with 157 seats to form a minority government, while the Conservatives won 121 seats (2019 Canadian federal election). However, the Liberals won only 33.12% of the popular vote, while Conservatives won 34.34% (2019 Canadian federal election). Voter turnout was at 67.0% (Voter Turnout at Federal Elections and Referendums). In this report, a multilevel regression model will be built using the 2019 Canadian Election Study (CES) dataset (Stephenson et al.) followed by poststratification analysis with the Census of Canada 2016 dataset (Census of Canada, 2016) to determine the popular vote outcome if all eligible voters voted in the 2019 Canadian Election.

In the Methodology section, I will describe the data, the model and the poststratification analysis that was used. Results of the analysis on how the outcome of the 2019 Canadian Federal Election will change given that all eligible voters voted will be presented in the Results section and conclusions, weaknesses and next steps will be made in the Discussion section.

Methodology

I will be predicting the popular vote outcome of 2019 Canadian Federal Election by building a multilevel logistic regression model using the 2019 Canadian Election Survey (CES) dataset (Stephenson et al.) then employing post-stratification based on data retrived from the Census of Canada 2016 dataset (Census of Canada, 2016). I will descibe the data, model specifics and the post-stratification technique in the following subsections.

Data

The data used to build the multilevel logistic regression model is the 2019 Canadian Election Survey data (web survey) (Stephenson et al.) downloaded from http://www.ces-eec.ca/. The CES (web survey) was an online survey targeting Canadian citizens and permanent residents aged 18 or over. Data was collected through an initial survey that was carried out during the Campaign period (sep 13 - Oct 21 2019) with 37822 responses collected and post-election (Oct 24 - Nov 11 2019) 10,340 respondents from the Campaign period survey were recontacted for a follow-up survey (Stephenson et al.). I then filtered the data by removing the respondents who were not eligible to vote, do not intend to vote or did not provide a clear vote choice. Then new categorical variables were made to determine if the respondent will vote/voted Conservative and if they will vote/voted Liberal, prioritizing the post-election survey vote choice if it exists. Then age groups (18-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75+) and education levels (No certificate, diploma or degree, HS Diploma or equivalent, Non-university degree, Bachelor's degree, and Degree above Bachelor's) where made and province was cleaned to have Nunavut, Northwest Territories and Yukon considered as Northern Canada. Then the variables vote_conservative, vote_liberal, province, age_group, sex, and education were selected based on what is commonly considered when analysing election results and what is avaliable in the poststratification dataset.

The data used to in poststratification analysis is the Census of Canada 2016 dataset (Census of Canada, 2016) retrieved from the CHASS website. This dataset, which represents about 2.7% of the Canadian population, contains information from a sample of the 2016 Canadian Census questionnaire 2A-L (Census of Canada, 2016). The Canadian Census is a census carried out every 5 years given to the entire Canadian population. However, it does not include Canadian citizens living outside of Canada, full-time Canadian Forces members outside of Canada and people living in institutional/non-institutional collective dwellings (Census of Canada, 2016). Using this dataset, I filtered out the people who are not eligible to vote (underage or not citizens/permanent residents). Then I selected and cleaned for age groups, sex, education and province to match the data used to build the multilevel logistic regression model.

However, one issue with these datasets is that the CES dataset contains data from 2019 and the Census dataset contains data from 2016. Going forward it is assumed that the Canadian population did not drastically change from 2016 and 2019, and thus the 2016 dataset gives a representative sample of the 2019 dataset.

Model Specifics

First, I will be using two multilevel logistic regression models, one to model the proportion of voters who will vote Conservative and another to model the proportion of voters who will vote Liberal. Logistic regression is used because voting Conservative and voting Liberal are both defined as categorical variables. I will be using age group (18-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75+), sex (male or female) and education (No certificate, diploma or degree, HS Diploma or equivalent, Non-university degree, Bachelor's degree, and Degree above Bachelor's) as individual level categorical variables and province (Nunavut, Northwest Territories and Yukon are considered Northern Canada) as a group level categorical variable to build both models.

The multilevel logistic regression model I am using is a random intercept model:

$$log(\frac{p}{1-p}) = \beta_{0j} + \beta_{age\_group} x_{ij} + \beta_{sex} x_{ij} + \beta_{education} x_{ij} + \epsilon_{ij} \qquad (eq.1)$$

where p represents the proportion of voters who will vote Conservative or Liberal depending on the model. β_{0j} represents the intercept of the model dependent on province, and is the logistic probability of voting for the candidate of someone who is 18-24 years old, female and has a Bachelor's degree living in the province j. Additionally, β_{age_group}, β_{sex}, $\beta_{education}$ represents the slopes of the model, which are the same for all provinces. So for a person of a certain age group, sex, and education level, we can expect the $log(\frac{p}{1-p})$ value to increase by $\beta_{age\_group} + \beta_{sex} + \beta_{education}$ where the betas are the values corresponding to each category.

Furthermore, we have:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \qquad (eq.2)$$

where β_{0j} represents the intercept of the model dependent on province. γ_{00} represents the overall intercept, which is value when all predictors are equal to 0 and γ_{01} is the slope between β_{0j} and the province j and u_{0j} is the random error component for the deviation of the province intercept from the overall intercept.

The model was run using R (R Code Team (2020)) and the tidyverse package (Wickham et al.).

Post-Stratification

To esimate the proportion of voters who will vote Conservative and the proportion of voters who will vote Liberal I will use a post-stratification analysis. Post-stratification is used to correct for differences between the sample and target population. It is done by partitioning the population into various demographic cells, estimating the response variable within each cell based off the sample, then combining the cell-level estimates to a population-level estimate by weighing each cell by its relative proportion in the population. I will create cells based off age group, sex, education level and province. So an example cell would be 18-24 male who has a Bachelor's living in Ontario. Then by using the model in the previous section, I will estimate the proportion of voters who will vote for the party in each cell. Next, I will weight the proportion estimate of each cell by the population size of that cell, sum all of the proportion values together and then divide by the entire population size.

Results

Table 1 gives the logistic multilevel regression fixed effect results of both the model that predicts the proportion of voters that will vote Conservative and the proportion that will vote Liberal. The β values (except intercept) are given as the non-bracketed numbers, and the 95% confidence interval is given in the brackets. γ_{00} is given by the non-bracketed value for intercept.

Table 1: Regression Results

	Dependent variable:	
	Vote conservative	Vote liberal
	(1)	(2)
25-34	0.478*** (0.343, 0.614)	−0.061 (−0.179, 0.057)
35-44	0.782*** (0.648, 0.916)	0.055 (−0.062, 0.172)
45-54	0.904*** (0.770, 1.038)	0.076 (−0.041, 0.194)
55-64	0.948*** (0.817, 1.078)	0.216*** (0.102, 0.330)
65-74	0.897*** (0.764, 1.030)	0.291*** (0.175, 0.408)
75+	1.135*** (0.971, 1.299)	0.325*** (0.173, 0.476)
Male	0.404*** (0.350, 0.459)	−0.119*** (−0.172, −0.066)
Degree above Bachelor's	−0.230*** (−0.325, −0.135)	0.062 (−0.022, 0.147)
HS Diploma or equivalent	0.254*** (0.184, 0.324)	−0.403*** (−0.470, −0.337)
No certificate, diploma or degree	0.242*** (0.110, 0.374)	−0.552*** (−0.687, −0.418)
Non-university degree	0.296*** (0.217, 0.374)	−0.448*** (−0.524, −0.372)
Intecept	−1.879*** (−2.269, −1.488)	−0.478*** (−0.818, −0.138)
Observations	27,143	27,143
Akaike Inf. Crit.	32,020.860	33,828.460
Bayesian Inf. Crit.	32,127.580	33,935.180
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 2 gives the values for u_{0j}, the random error component for the deviation of the province intercept from the overall intercept.

Table 2: Random Effect Intercepts

Province	Conservative Intercept	Cons. Int. SD	Liberal Intercept	Libs. Int. SD
Alberta	1.292	0.036	-0.877	0.045
British Columbia	-0.059	0.038	-0.162	0.038
Manitoba	0.394	0.058	-0.224	0.062
New Brunswick	-0.179	0.088	0.212	0.081
Newfoundland and Labrador	-0.398	0.107	0.558	0.092
Northern Canada	-0.208	0.262	0.112	0.238
Nova Scotia	-0.549	0.088	0.467	0.072
Ontario	-0.035	0.021	0.207	0.020
Prince Edward Island	-0.597	0.223	0.387	0.183
Quebec	-0.576	0.035	0.432	0.029
Saskatchewan	0.947	0.064	-1.098	0.088

Table 3 gives the post-stratified estimate of the proportion of voters that will vote Conservative and the proportion of voters that will vote Liberal.

Table 3: Post-Stratification Results

Vote conservative	Vote liberal
0.337	0.34

I estimate that the Conservatives will win 33.7% of the popular vote, while the Liberals will win 34.0% of the popular vote. This is based of post-stratification analysis of two multilevel logistic regression models that estimated the proportion of voters that will vote in favour of the Conservatives and the proportion of voters that will vote in favour of the Liberals, which considered the age group, sex and education level as individual level variables, and province as a group level variable.

Discussion

Summary

The data used to build the multilevel logistic regression model is the 2019 Canadian Election Survey data (web survey) (Stephenson et al.) retrieved from http://www.ces-eec.ca/. It was cleaned by filtering out the respondants who are not eligible/not planning to vote or did not provide a clear vote choice, then selecting for a voters age group, sex, education level, province, and which party they are going to vote for/voted for. New variables were added to indicate if a voter will vote Conservative or Liberal based off of their vote intention/vote choice. Furthermore, age groups were built, education level was reduced into fewer groups, and the Northern provinces were condensed into Northern Canada. The data used in post-stratification is the Census of Canada 2016 dataset (Census of Canada, 2016) retrieved from the CHASS website. A person's age group, sex, education level, and province was selected, and data was cleaned to match the data used in the multilevel logistic regression model. Demographic cells were then made. The model and the post-stratification analysis were both run in R (R Code Team (2020)) using the tidyverse (Wickham et al.) and lme4 packages (Douglas et al.).

Conclusions

From the results we see that the Liberals are expected to win 34.0% of the popular vote and the Conservatives are expected to get 33.7% of the popular vote if all eligible voters voted in the 2019 Canadian Federal Election. This is contrary to what we actually saw in the 2019 Canadian Federal Election, where the Conservatives won the popular vote at 34.34% while the Liberals had 33.13% of the popular vote. This means that if the Canadian election system was based off the popular vote instead of First Past the Post, the Conservatives are the actual winners of the 2019 Canadian Federal Election at 67% voter turnout. However, at 100% voter turnout, the Liberals are predicted to win the election. This emphasizes the importance of going out and actually voting.

However, because Canada uses a First Past the Post system popular vote does not actually determine who wins the election. This can be seen in the actual results of the 2019 Canadian Federal Election as the Conservatives won the popular vote, but the Liberals won the most seats, thus winning the election. This means that there is a chance that the Conservatives will win the election even at 100% voter turnout based on how the federal electoral districts vote. However, this seems unlikey as this analysis suggests that the Liberals would gain 0.88% in the popular vote while Conservatives would lose 0.64% of the popular vote. Therefore, there does not seem to be a change in which party will win the election even if all eligible voters voted, however the number of seats won could change with the Liberals predicted to win more seats.

Weaknesses

Because this analysis focuses on the popular vote outcome and not the how many seats are won be each party, it is harder to make conclusions on the winner of the election as winning the popular vote does not imply winning more seats. Furthermore, there is a chance that religion may influence a voter's vote choice. Based off exit polls of the 2016 US election, Christians were more likely to vote Trump than Jewish, other religion, and non-religious voters (2016 United States presidential election), which means the same effect may be seen in Canadian voters.

Next Steps

One next step is to do another analysis accounting for religion to see if that variable has an effect on popular vote outcome after the 2021 Canadian Census data is collected and released. This is because the 2021 Census will include religion as religion is question every 10 years. Another next step is could be to do another analysis of what the seat result of the election will be at 100% voter turnout. However, new data will need to be collected to contain the electoral district of each respondent to do the analysis. Thus, this analysis could be done for the next Canadian Federal Election instead.

References

"2019 Canadian federal election." Wikipedia, https://en.wikipedia.org/wiki/2019_Canadian_federal_election.

"2020 United States Presidential Election." Wikipedia, https://en.wikipedia.org/wiki/2020_United_States_presidential_election.

"Census of Canada, 2016." Statistics Candada, 2016. Accessed 19 December 2020.

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.1. https://CRAN.R-project.org/package=stargazer

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Stephenson, Laura B., Allison Harell, Daniel Rubenson and Peter John Loewen. The 2019 Canadian Election Study –Online Collection. [dataset]

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

"Voter Turnout at Federal Elections and Referendums." Elections Canada, 20 Dec. 2020, www.elections.ca/

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686