Predict 2019 Canadian Federal Election with Full Turnout using MRP

Jo-Yen Lin

December 22nd, 2020

Code and data supporting this analysis is available at: https://github.com/jlin213/Predict-2019-Canadian-Federal-Election-with-Full-Turnout-using-MRP

Abstract

Multilevel regression with post-stratification (MRP) model is used for correcting model estimates for known differences between a sample population, and a target population. In this study, we are interested in what the 2019 Canadian Federal Election would be with full turnout. We will be using 2019 Canadian Election Survey as our sample survey data, and 2017 General Survey Study as census data. While the Liberal Party won the election with 39.47% of votes, the results from our MRP predicted that the Conservatives will have 28% of votes the Liberals will have 22% of votes and the NDP will have 9% of votes.

Keywords

Multilevel Regression, Post-stratification, Election, Prediction, MRP

Introduction

Statistical analysis can be applicable to not only scientific community, but also political perspectives. In the 2019 Canadian Federal Election, Justin Trudeau continues his position as Prime Minister with a Liberal Minority Government, and the Conservative Party won the popular vote. While the election was close between the opposition parties, there was only 66% of the eligible voters voted in 2019. Going back to previous elections, the last time the percentage of voter turnout is above 70% was in 1988 (Election Canada, n.d.). This has been a concerning issue every election not only in Canada, but other democratic countries as well. The study is interested in predicting how the election would have turned out if everyone who are eligible voted.

In this study, a multilevel regression with post-stratification (MRP) model based on the 2019 Canadian Election Survey as survey data, and uses post-stratification technique with the 2017 General Survey Study as census data to see how the 2019 Canadian Federal Election will turn out if the whole Canadian population voted. With the close race between the Liberal Party and the Conservative Party, it could possibly be that the situation would have changed if everyone voted. With post-stratification analysis, we can estimate the different proportions of voters who will vote for the Liberal Party, the Conservative Party, or the NDP Party through with the GSS data. The cells are divided based on race, education, and regions. Then, with multilevel regression, we can smooth the non-representative cells with overall averages.

Two data sets will be used to for this study. The first data set is the 2019 Canadian Election Survey by phone survey. The second data set is the 2017 General Survey Study that will be used as census data. The Methodology section will include description and characteristics of the data sets as well as the model that was chosen with variables of interest. In the Results section, the study will analyze the results from post-stratification analysis and determine which party will most likely to win.

Methodology

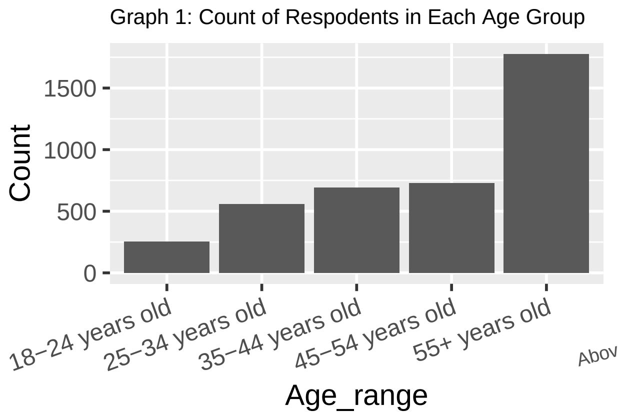
Data

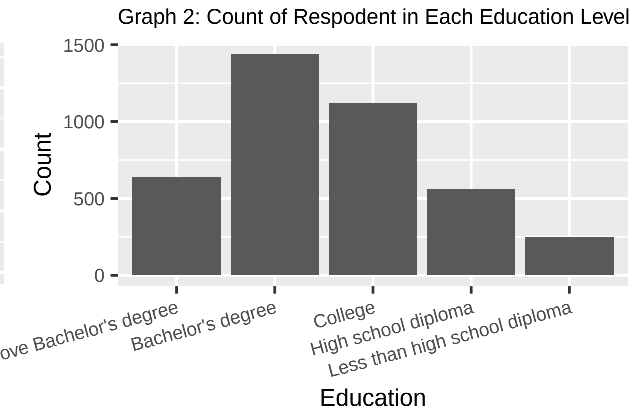
The 2019 CES data was conducted by Laura Stephenson, Allison Harell, Daniel Rubenson and Peter Loewen and later published in April 2020. In particular, this study selected 2019 CES phone survey data. The participation was voluntary and targeted at eligible voters in all 13 provinces. The key features of the survey include large question banks (variables) and political interests were indicated clearly due to nature of the survey. The 2017 GSS data was collected by Statistics Canada through General Social Survey (GSS) from February 1st to November 30, 2017 through telephone interviews and participation was voluntary. The population is targeted at all Canadians who are 15 years old or older and living in the 10 provinces. From Statistics Canada, "the GSS uses a frame that combines landline and cellular telephone numbers from the Census and various administrative sources with Statistics Canada's dwelling frame" (Government of Canada, 2020). Additionally, only one person from each household is selected. The key features of the survey include large sample size (43,000 people for year 2017), with diverse characteristics of childhood, immigration status, marriage, education background. GSS allows research in marriage and family to have extensive data.

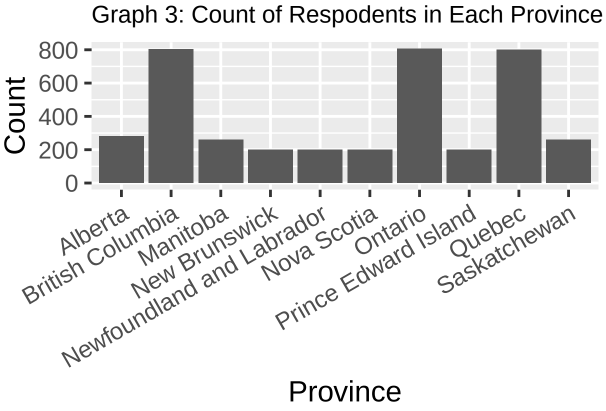
Table 1: Levels of Variables of Interest

Age_range	Education	Province
18-24 years old	Bachelor's degree (e.g. B.A., B.Sc., LL.B.)	Alberta
25-34 years old	College, CEGEP or other non-university certificate or di…	British Columbia
35-44 years old	High school diploma or a high school equivalency certificate	Manitoba
45-54 years old	Less than high school diploma or its equivalent	New Brunswick
55+ years old	University certificate, diploma or degree above the bach…	Newfoundland and Labrador
		Nova Scotia
		Ontario
		Prince Edward Island
		Quebec
		Saskatchewan

We will select several variables from the CES data set as our variable of interests in estimating the proportion of voters in each party. In this study, we will predict with the MRP model with age range, education level, province, and religion affiliation. In Table 1, we can see the categories of age range, education level, and province for CES respondents. The CES data set has been cleaned with R script provided in the associated Github link. Particularly, many categories has been renamed in order to synchronize with our GSS data in the analysis later. For both 2019 CES data and 2017 GSS data, many of the Non-Application(NA) data points are removed. Especially for 2019 CES data, respondents who chose to not respond, forgot, or skipped the questions are considered as NA data points, since we cannot make any inference on their answers. It is also worth noting that during data cleaning, for the 2017 GSS data, education level "Trade certificate or diploma" was considered as "College, CEGEP or other non-university certificate or di…", and "University certificate or diploma below the bachelor's level" was included in "Bachelor's degree (e.g. B.A., B.Sc., LL.B.)" category.







Graph 1, Graph 2, and Graph 3 shows the demographic count of each categories in the CES respondents. We can see that large portion of respondents are 55+ years old and there is a trend of increase in age with increase of respondents. We can also see that in Graph 2, large portion of respondents have a Bachelor's degree and the education level is right skewed. In terms of location, respondents living in British Columbia, Ontario, and Quebec the most.

Model

The models that will be used are three multilevel logistic linear regression (MLR) models, where one will be modeling proportion of voters who will vote for the Liberal Party and another will be modeling the proportion of voters who will vote for the Conservative Party and the last one will be modeling the proportion of voters who will vote for the NDP Party. The predictors are age range, education level, and religion affiliation, with province to model the intercept at group level. Mathematically, the model for proportion of voters who will vote for the Liberal Party is:

$$logit(y_{Liberal}) = \beta_{0j} + \beta_1 X_{education} + \beta_2 X_{age\_range} + \beta_3 X_{Religion} + u_{ij}$$

$$\beta_{0j} = r_{00} + r_{01} X_{Province}$$

The model for proportion of voters who will vote for the Conservative Party is:

$$logit(y_{Conservative}) = \beta_{0j} + \beta_1 X_{education} + \beta_2 X_{age\_range} + \beta_3 X_{Religion} + u_{ij}$$

$$\beta_{0j} = r_{00} + r_{01} X_{Province}$$

The model for proportion of voters who will vote for the NDP Party is:

$$logit(y_{NDP}) = \beta_{0j} + \beta_1 X_{education} + \beta_2 X_{age\_range} + \beta_3 X_{Religion} + u_{ij}$$

$$\beta_{0j} = r_{00} + r_{01}X_{Province}$$

In our main model equation, the X variables are categorical and specific levels can be referenced in Table 1. Since we are using using logistic regression,we will have log odds (logit) function on the left side of our model equations. $logit(y_{Liberals})$ represents the proportion of voters who will vote for the Liberal Party, $logit(y_{Conservative})$ represents the proportion of voters who will vote for the Conservative Party, and $logit(y_{NDP})$ represents the proportion of voters who will vote for the NDP Party. For β_{0j}, r_{00} and r_{00} are the intercept and slope of the random effects term, in which we use different provinces to model the intercept. u_{0j} is the random error component for the deviation of the intercept of different provinces from the overall intercept.

With post-stratification analysis, we can estimate the different proportion of voters who will vote for the Liberal Party, the Conservative Party, and the NDP Party with the 2017 GSS data.The cells are divided based on age_range, religion, education, and province. Therefore, each bin is different levels of age_range, religion, education, and province. Then, by estimating the proportion of voters in each bin, we will weight each proportion estimate based on the population size given by the GSS data. Finally, we will sum all these values and divide by the entire population size.

Results

Table 2: Prediction of Proportion of Voters in Each Party

	Proportion of voter
Liberals	0.2221381
Conservative	0.2841272
NDP	0.0878179

After the post-stratification analysis, we obtained the following results in Table 2. We predict that the proportion of voters who will vote for the Liberals Party is 0.222. That is, 22% of eligible voters with known education level, age range, located province and religion affiliation will vote for the Liberal Party according to the MRP model mentioned in the Model Section above. We also predict the proportion of voters who will vote for the Conservative Party is 0.284 and the proportion of voters who will vote for the NDP party is 0.088.

Discussion

Summary and Conclusions

From the Results Section, we found that the model predicts the Conservative Party will likely to win given by the preferences of respondents in 2019 Canadian Election Survey and weighted average with 2017 General Survey Study data. We can see that all the predicted percentage of voters are lower than actual election results. This suggests that it is possible that proportion can spread out to other potential parties that this study did not account for, such as the Bloc Québécois Party and the Green Party. It is also possible that during 2019 Canadian Federal Election, these voters who turn out for the Liberal Party were ones responded unsure of their votes.

Weaknesses and Next Steps

One big weakness that this study fails to account is the small sample size of 4000 in 2019 Canadian Election Study by phone survey. Due to access of data set, the 2019 CES phone survey was easier to obtain and clean than the web version. In comparison,the 2017 GSS data has over 43,000 data points. Another weakness is that as mentioned in the Data Section and Discussion section, many of the data points are considered as NA because respondents didn't indicate their political party of interest or refused to provide any. In this case, this reflects to the 66% voter turnout in 2019 Election, where rest of 33% of eligible voters didn't vote. The survey was conducted during the advanced voting period; therefore, many respondents who didn't respond to the political party of interest question may be still in process of deciding which party to vote as well.

There are many areas that can be improved on the study, data, and the model. As mentioned previously, a larger sample size of the survey data would help predicting the voter turnout better. If possible, a more updated census data would reflect the current target population better than a data set from three years ago. One area of improvement is CES will be including a democracy checkups, which will collect data during non-election years. Further studies can investigate more whether voters are likely to change their political party of interest during the election period or during the non-election years. For the model itself, it is possible to use AIC and BIC to evaluate the variables selected, and see if any variables should be added or dropped according to the AIC value.

References

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

Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

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.

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

Faculty of Arts & Science, University of Toronto. (26, January 2017). Computing in the Humanities and Social Sciences. Retrieved from http://www.chass.utoronto.ca/

Government of Canada, S. C. (2020, April 30). General Social Survey – Family (GSS). Retrieved from https://www.statcan.gc.ca/eng/survey/household/4501

Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, "2019 Canadian Election Study - Phone Survey", https://doi.org/10.7910/DVN/8RHLG1, Harvard Dataverse, V1

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

Wickham, H., & Grolemund, G. (2016). R for data science: import, tidy, transform, visualize, and model data. " O'Reilly Media, Inc.".

Xie, Y., Dervieux, C., & Riederer, E. (2020). R Markdown Cookbook. CRC Press.

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.