A Relook at the 2019 Canadian Federal Election Using Multilevel regression with post-stratification

Wenzhuo Zeng

12/20/2020

Code and data supporting this analysis is available at: https://github.com/wendy0412/A-Relook-at-the-2019-Canadian-Federal-Election-Using-Multilevel-regression-with-post-stratification.git

Abstract

The paper is to identify how the 2019 Canadian Federal Election would have been different if the voter turnout was 100%. Two multilevel regression model models was built to study and predict the voters' vote intention, then post-stratification analysis was performed to estimate the vote intention for the whole Canadian population. A 100% turnout was examined to benefit Liberal Party.

Keywords

2019 Canadian federal election, Voter turnout, Liberal Party, Conservative Party, Multilevel regression, Generalized linear mixed effect model, Post-stratification

Introduction

Voter turnout in Canadian federal election has been quite low since 2000, staying at around 60% (Elections Canada, 2020). In 2015 when Justin Trudeau first led the Liberal Party to victory, the voter turnout reached its highest in the decade, then decreased to 66% for the 2019 Canadian federal election. This means that one third of the Canadian population was not represented in the election results. The most common reason for not voting was "not interested in politics" (Statistics Canada, 2020), which brings in an interesting topic: how would the voice of the one third population changes the election result? A study by Rubenson et at. (2007) examined that higher turnout slightly benefits certain parties.

The goal of this paper is to take a relook at the 2019 Canadian Federal Election, to find out how the results of the election would have been different if the voter turnout was 100%. One way of doing it is thorough building multilevel regression models with post-stratification (MRP). Adopting MRP adjusts sampling weights and the difference between sample and population, and balances the underestimation caused by the underrepresentation of a group, thus provides a fair estimate for the whole population (Wang et al., 2015).

Two data sets will be used to examine the outcome, which will be presented in the methodology section along with two multilevel regression models. In the result section, the calculated proportion of voters in favour in voting for Liberal Party and Conservative Party will both be presented. Conclusions and any weakness of this study will be discussed in the Discussion section.

Methodology

Data

Two sets of data are used for this study. The first set of raw data used is 2019 Canadian Election Study - Online Survey (2020), which is a survey contains information of Canadians' demographic, political views and behaviour and more. The data set contains 37,822 observations and 620 variables, and helps to capture an aspect of Canadians' political life. The second data set is from the 2016 Census of Population provided by Statistics Canada (2017). The two data sets are then cleaned, the cleaned data sets are both left with four common variables: sex, age, province, and highest education attainment. One more variable, cps19_votechoice, is kept in survey data indicates which party the person would vote for. A new indicator variable, vote_liberal is created using cps19_votechoice, where 1 stands for intention voting for Liberal Party and 0 otherwise. One other indicator variable, vote_conservative, is also created with 1 stands for intention voting for Conservative Party and 0 otherwise. Table 1 shows 6 observations of the cleaned survey data set, Table 2 presents 6 observations from the census data. 32 cells are also created ready to be used for building models and post-stratification, created cells will be discussed in the Post-stratification subsection. Weaknesses of using the two data sets will be further discussed in the Discussion section.

Table 1: Survey Data at A Quick Look

age	sex	province	education	cps19_votechoice	vote_liberal	vote_conservati
25 to 34	Female	Quebec	university	Green Party	0	0
55 to 64	Female	Quebec	technical,community college	Liberal Party	1	0
25 to 34	Female	Ontario	university	ndp	0	0
55 to 64	Male	British Columbia	university	Conservative Party	0	1
55 to 64	Female	Ontario	technical,community college	Liberal Party	1	0
55 to 64	Male	Ontario	university	ndp	0	0

Table 2: Census Data at A Quick Look

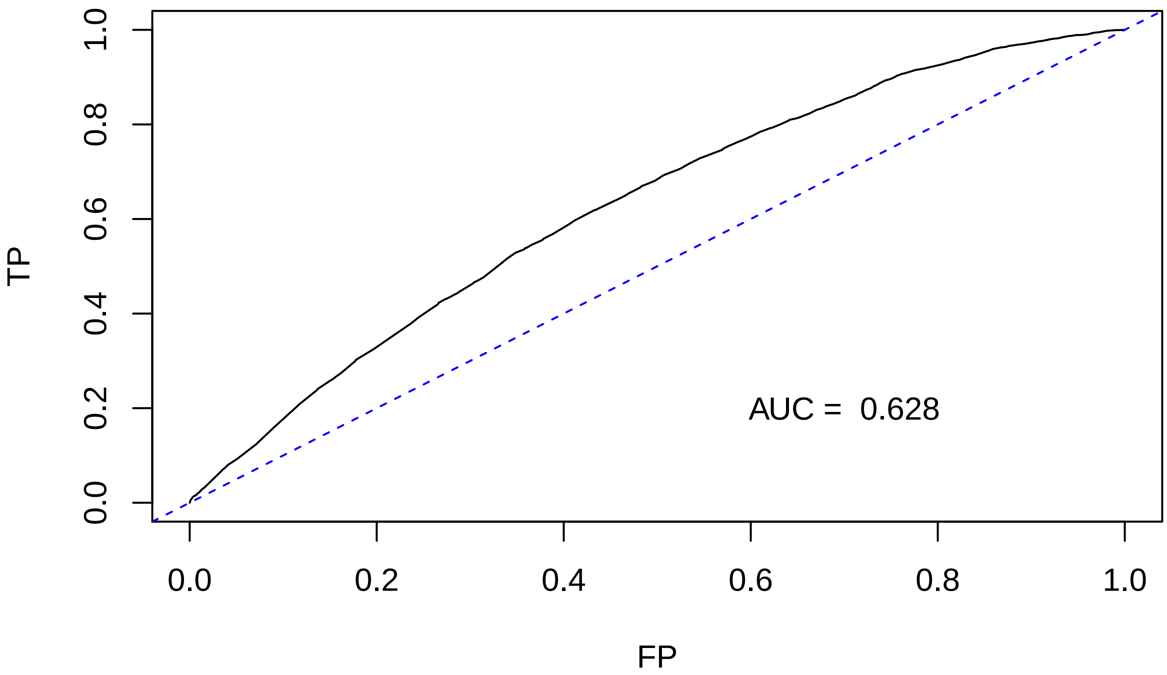
age	sex	province	education	total_count
25 to 34	Male	Newfoundland and Labrador	no degree	2615
25 to 34	Male	Newfoundland and Labrador	high school	6440
25 to 34	Male	Newfoundland and Labrador	technical,community college	5420
25 to 34	Male	Newfoundland and Labrador	technical,community college	7030
25 to 34	Male	Newfoundland and Labrador	university	645
25 to 34	Female	Newfoundland and Labrador	no degree	1970

Model

The first model is a generalized linear mixed effect model which used sex, age, province, and highest education attainment as predictors, and vote_liberal as response variable to indicate a voter's voting result. This model predicts how likely a voter is to vote for Liberal Party, and is presented as follows:
$\log(\frac{P(Y_{Li}|X_{Li})}{1-P(Y_{Li}|X_{Li})}) = X_{Li}\beta_{Li} + \mu_L$

Y_{Li} stands for whether a voter votes for the Liberal Party. If the ith voter votes for the Liberal Party, then $Y_{Li} = 1$, $Y_{Li} = 0$ if otherwise.$P(Y_{Li}|X_{Li})$ would be the probability of the ith person voting for Liberal Party, given X_{Li}. μ_L is the intercept. In this model the coefficient for intercept is -2.04, which means that a 25 to 34 years old lady lived in Alberta whose highest education attainment is high school is less likely to vote for the Liberal Party. Each β_L represents the coefficient of each predictor. Area under the ROC curve is also calculated and plotted in Figure 1, the AUC is 0.628, indicates that this model can predict voters vote for Liberal Party and voters do not 63 out of 100 times correct.

Figure 1: AUC−ROC curve for model 1

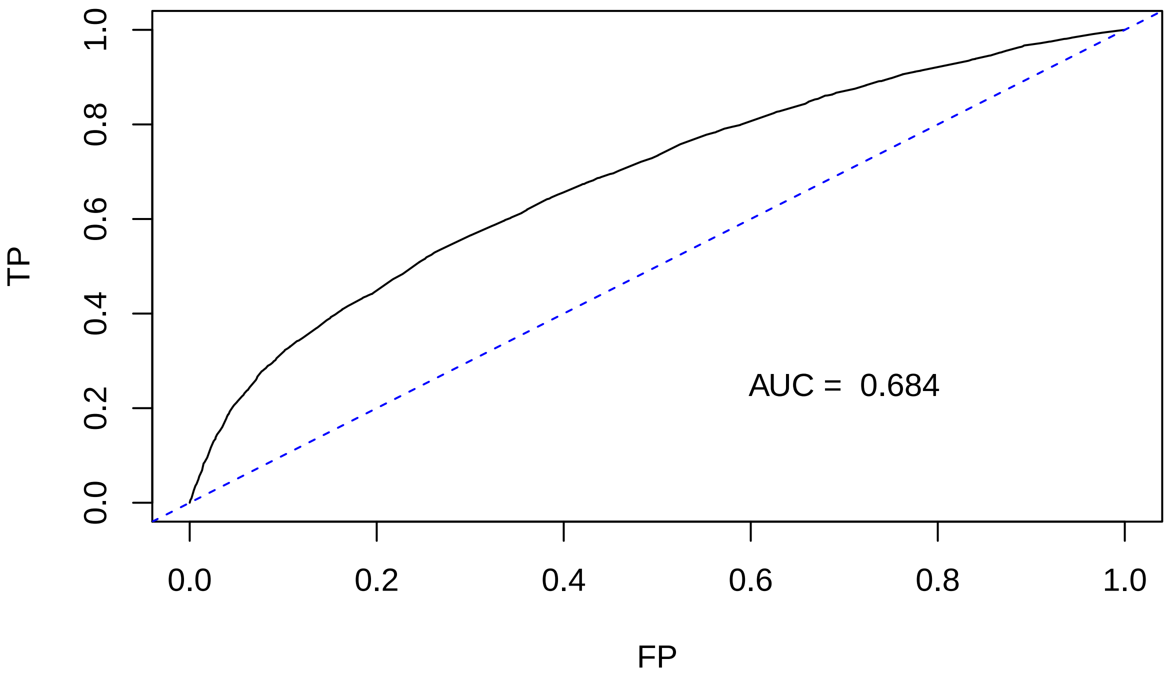


The second model is also a generalized linear mixed effect model and has the exact predictors as the first model, with a different response variable, vote_conservative. This model predicts how likely a voter is to vote for Conservative Party:

$\log(\frac{P(Y_{Ci}|X_{Ci})}{1-P(Y_{Ci}|X_{Ci})}) = X_{Ci}\beta_{Ci} + \mu_C$

The coefficient for μ_C is 0.24642, means that a 25 to 34 years old lady lived in Alberta whose highest education attainment is high school is quite likely to vote for the Conservative Party. AUC for model 2 is 0.684, ROC curve plotted in Figure 2.

Figure 2: AUC−ROC curve for model 2



Weaknesses of two models will be further discussed in Discussion section.

Post-stratification

| Post-stratification is a method to adjust sampling weights and the difference between sample and population. It balances the underestimation caused by the underrepresentation of a group. To do so, the population are divided into cells. The built model by survey data is applied to census data and therefore estimate the interest within each cell. The weight in each cell is calculated by its proportion in the population. 32 cells are created using sex, age and highest education attainment, because they might have an influence on popular vote. People with age difference might hold different political views. People with the dissimilar educational background may also resonate with the views of different parties.

Results

Based on the two generalized linear mixed effect models and post-stratification analysis and calculation of the proportion of voters in favour of Liberal Party. The estimated proportion of voters in favour of voting for Liberal Party is 40.8%. The estimated proportion of voters in favour of voting for Conservative Party is 21.2%.

Discussion

Summary

So far, two models has been built to learn and predict voters preferences based on the information provided by survey data. Then the two models were applied on to census data by post-stratification analysis to give a larger prediction for the general Canadians' preference in voting. Thus resemble the voting process with a 100% voter turnout. But what does the result mean?

Conclusion

| The estimated proportion of voters in favour of voting for Liberal Party is calculated to be 40.8%. there is a 7% percent increase compared to 33.12% which is the real proportion of voters voted for Liberal Party in 2019 Canadian federal election (Elections Canada, 2020). However the estimated proportion of voters in favour of voting for Conservative Party, 21.2% is 13% less than the real proportion of voters voted for Conservative in 2019 Canadian federal election. This means that a 100% voter turnout may favour Liberal Party, and disfavour Conservative Party.

Weakness and Next step

However, the higher estimated popularity for Liberal Party does not guarantee the victory for Liberal Party, because of the electoral system of Canada. Each member of the House of Commons represents a single electoral district based on geographical divisions. The candidate with the most number of votes in every electoral district wins one seat in the House of Commons and will represent that electoral district as its member of Parliament. The party wins the most seats wins the election. A high proportion of votes does not always means a high number of seats. Therefore, the analysis of this paper can not predict which party would win. A reasonable further step would be analyze and study the vote intension in each electoral district and to estimate the number of seats each party might win.

The census data used in this study has limited variables provided, therefore a lot of variables in the survey data that might be helpful in construct models could not be used, such as income, race, and more. The census data also only provide information on ones who aged above 25 years old. Observations aged between 18 to 25 are therefore deleted just to be consistent with census data, which definitely would effect the model, as a part of the population is not being studied. Another issue with the two data sets is that, while survey data records participants' gender which is a socially constructed concept, census data measures people's sex which is defined by biological features and genes, and is binary. This study simply took one of a few methods mentioned in a paper by Kennedy et al. (2020), which is to remove participants with a non-binary gender, and to categorize anyone identify as a woman (gender) as female (sex), anyone identify as a man as male. The ethicality of this procedure can be questioned, and this can effect the model, however, a procedure like this has to be done. One last weakness of the census data set that also needs attention is that the data set is a census for 2016, as it is the latest census data by Statistics Canada. Applying a model built on survey data from 2019 to a census data from 2016 may also effect the analysis and results.

As mention in the Model section, area under the ROC curve for both models is between 60% and 70%, which means that the two models demonstrate some ability in predicting the popular votes. However these two numbers can still be improved by adding more variables to the models, and by applying post-stratification analysis on a different census data set.

References

David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.2. https://CRAN.R-project.org/package=broom

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.

Elections Canada. (2020). Voter Turnout at Federal Elections and Referendums. Retrieved December 21, 2020, from https://www.elections.ca/content.aspx?section=ele

Elections Canada. (2020). Official Voting Results FORTY-THIRD GENERAL ELECTION. Retrieved December 21, 2020, from https://www.elections.ca/res/rep/off/ovr2019app/home.html

Electiona Canada. (2020). The Electoral System of Canada. Retrieved December 21, 2020, from https://www.elections.ca/content.aspx?section=res

Kay M (2020). *tidybayes: Tidy Data and Geoms for Bayesian Models*. doi: 10.5281/zenodo.1308151 (URL:https://doi.org/10.5281/zenodo.1308151), R package version 2.1.1, <URL:http://mjskay.github.io/

tidybayes/>. Kennedy, L., Khanna, K., Simpson, D., & Gelman, A. (2020). Using sex and gender in survey adjustment. arXiv preprint arXiv:2009.14401.

Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret

Rubenson, D., Blais, A., Fournier, P., Gidengil, E., & Nevitte, N. (2007). Does low turnout matter? Evidence from the 2000 Canadian federal election. Electoral Studies, 26(3), 589-597. doi:10.1016/j.electstud.2006.10.005

Sing T, Sander O, Beerenwinkel N, Lengauer T (2005). "ROCR: visualizing classifier performance in R."*Bioinformatics*, *21*(20), 7881. <URL: http://rocr.bioinf.mpi-sb.mpg.de>.

Statistics Canada. 2017. Education Highlight Tables, 2016 Census. Statistics Canada Catalogue no. 98-402-X2016010 Ottawa. Released November 29, 2017. http://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/edu-sco/index-eng.cfm

Statistics Canada. (2020, February 26). Reasons for not voting in the federal election, October 21, 2019. Retrieved December 21, 2020, from https://www150.statcan.gc.ca/n1/daily-quotidien/200226/dq200226b-eng.htm

Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, "2019 Canadian Election Study - Online Survey", https://doi.org/10.7910/DVN/DUS88V, Harvard Dataverse, V1

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. International Journal of Forecasting, 31(3), 980-991. doi:10.1016/j.ijforecast.2014.06.001

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

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77 http://www.biomedcentral.com/1471-2105/12/77/