Predicting Canadian Election 2019 differently

Canadian Federal Elections results if everyone had voted

AR

22 December 2020

Abstract

This project is based on Multilevel regression with poststratification (MRP) to predict the 2019 Canadian Federal Election. Data set is based on combination of the 2019 online survey (CES) for multiple regression analysis, and 2017 General Social Survey on family (GSS) for poststratification. Based on end result, the Liberal party won with 37%, which a deviation of 2% from actual results.

Keywords: Predictions, MRP, Canadian Federal Elections, CES, GSS, Liberal

Code and data supporting this analysis is available at:

https://github.com/Ar4yk/Final_Project

Introduction

Canadian House of Commons has 338 seats, held by members elected by citizens who voted in the general election in 338 electoral districts. In each electoral district, the candidate with the most votes will win a seat in the House of Commons and represent the electoral district as its member of parliament. The 2015 federal election resulted in a Liberal majority government headed by Justin Trudeau with 184 seats in House of Commons. In addition, Liberal party was able to get 39.47% of 17,559,353 national votes. Voter turnout was 68.5%, the highest turnout since 1993.

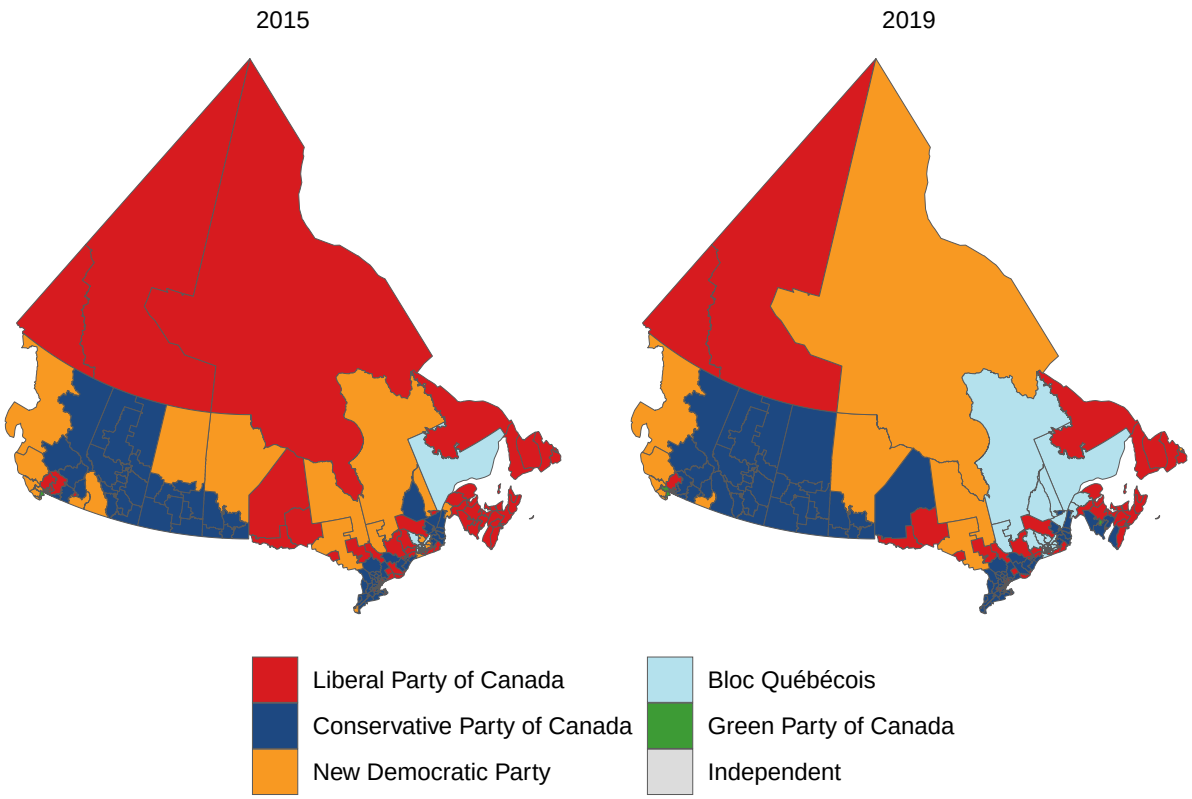
Table 1: 2015 results

Parties	Votes	Leaders	Seats
Liberal	39.47%	Justin Trudeau	184 / 338
Conservative	31.89%	Andrew Scheer	99 / 338
New Democratic	19.71%	Jagmeet Singh	44 / 338
Bloc Québécois	4.66%	Yves-François Blanchet	10 / 338
Green	3.45%	Elizabeth May	1 / 338

However, in the 2019 federal election, Trudeau won only 157 seats and lost majority in the overall vote by getting 33.1% of 17.9 million national votes. In addition, third of Canadians did not participate in the election vote, which results in voter turnout less than 67%. Despite loss in both seats and overall vote, Liberal party was able to form minority government. The overall 2019 Canadian elections results show, that current system of electoral college is not efficient. Therefore, the main objective of this report is to identify the outcome of the 2019 federal election if everyone had voted.

Table 2: 2019 results

Title	Votes	Leaders	Seats
Liberal	33.1%	Justin Trudeau	157 / 338
Conservative	34.3%	Andrew Scheer	121 / 338
Bloc Québécois	7.6%	Yves-François Blanchet	32 / 338
New Democratic	16.0%	Jagmeet Singh	24 / 338
Green	6.5%	Elizabeth May	3 / 338



Multilevel regression with post-stratification (MRP) was used to identify who would have won the 2019 Canadian federal election if everyone voted. The report used the 2019 CES online survey and the 2017 GSS data data sets. The 2019 CES online survey data was used to create a logistic regression model and 2017 GSS data was used for poststratification analysis. More information about data sets will be in **Data** section. Further topics involve sections: **Model/Methodology**, **Results** , **Discussion and Limitations** and **Next steps**.

Data

Multilevel regression with post-stratification (MRP) is a statistical technique used for correcting model estimates for known differences between a sample population and a target population. Thus, two data sets are used for MRP. Survey data will be based on 2019 Canadian Election Study (CES) and census/target data on 2017 Canada's General Social Survey (GSS). The 2019 CES data set had gathered attitudes and opinions from Canadians during the 2019 federal election. Since survey was conducted through a non-probability online survey, people without computer skills or internet access are not represented. Census data set is based on 2017 Canada's General Social Survey. Established in 1985, Canada's General Social Survey (GSS) main objective was to gather data on social trends in order to monitor changes in the living conditions and well being of Canadians.The data set contains responses of the 2017 General Social Survey. The contents of the survey are focused on family characteristics in Canada. For example, their health conditions, life satisfaction,

education levels, economic status and many others. CES data set includes The online survey data 37,822 observations and 620 variables , while GSS cycle 2017 includes 461 variables for 20602 observations in subset.

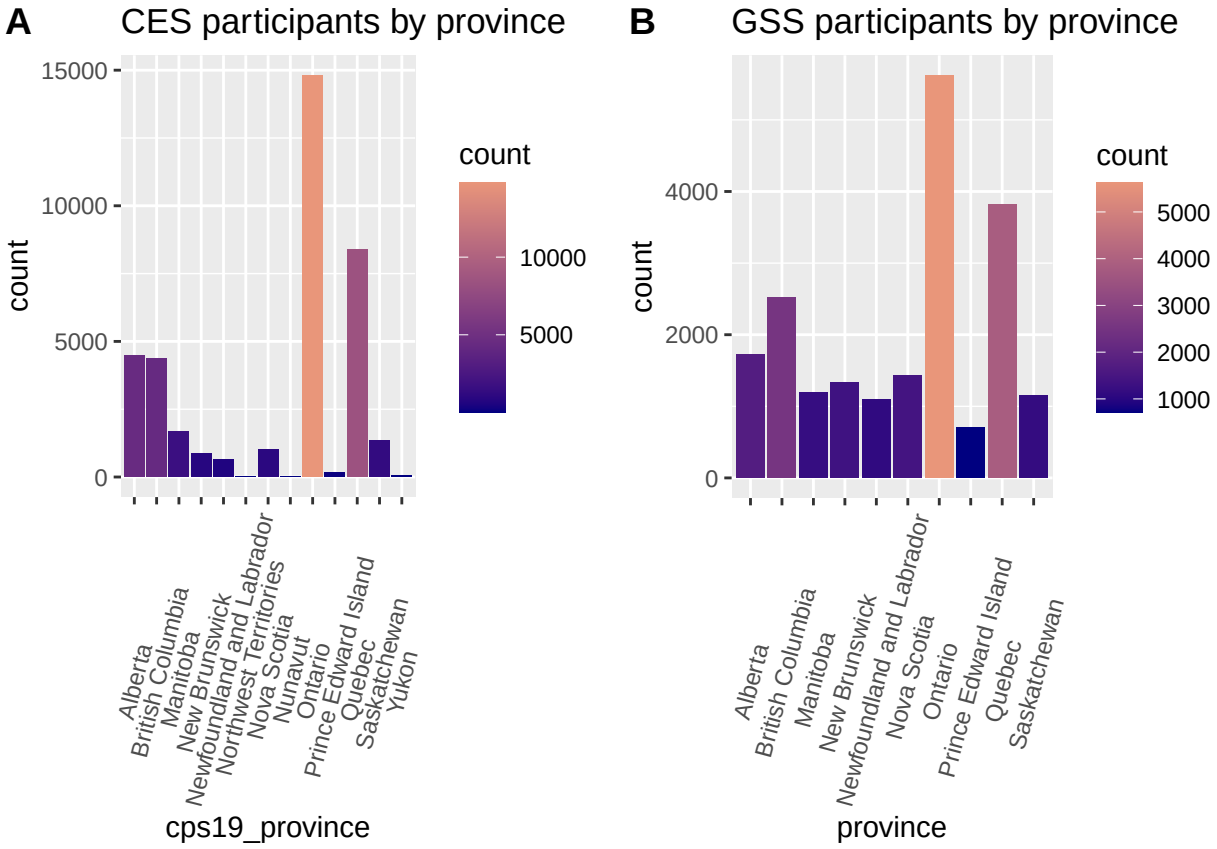
For multilevel logistic regression, CES variables selected:

1. Vote choice for specific party. We will focus on Liberal Party, since its won 2015 elections.
2. Gender, Male and Female
3. Age of the individuals.
4. Province of the residence
5. Educational level.

For poststratification analysis, the following variables from GSS are selected:

1. Sex, gender in two categories: Male and Female.
2. Provinces, where survey was conducted.
3. Age of the individuals minus younger then 18 years.
4. Household income bracket.
5. Educational levels

Without initial data cleaning, by the plots of participants in each study, two Provinces represent the majority: Ontario and Quebec. Two plots are include all variables and observations.



Model and Methodology

Based on previous elections surveys, such as Canadian Election Study 2015 and 2019, the fitted model will be based on multilevel logistic regression:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{gender} + \beta_{3i} x_{province} + \beta_{4j} x_{education}$$

1. $\hat{p}$ Percentage to vote for Liberal party
2. Variables marked as β_i, where variables from $1 \leq i \leq 4$ for the slopes.
3. β_0 is the measure of the model intercept and describes the probability of voting for the Liberal Party when the voter is male, from province Ontario with University degree.
4. With one unit increase in, voting for the Liberal Party measured in log odds is increased by β_1.
5. The β_{3i}'s represent the log odds voting for the Liberal Party in different provinces.
6. β_{4i}'s represent the log odds voting for the Liberal Party based on education.

Results

I was unable to make poststratification analysis.

Discussion and Limitations

1. Due to initial data cleaning, our results will be biased, since we remove many NA's values. Without thoughtful check for predictor variables we may miss multiple cases of multicollinearity.
2. Both data sets are not representative due to below average survey response rate. Response rate in Canadian Election Study (CES) is 5.6%. Response rate for General Social Survey (GSS) is 52.4%. Thus we may experience sampling biases.
3. Following our point 2, majority of the respondents in both data sets are in 45-60 age range. Therefore, both data sets are missing younger voters population.

Next Steps

1. CES data is bigger then GSS, thus for the next project we could use a bigger data set for post-stratification analysis. Smaller data set will cause a bias.
2. Adding or removing predictors by using Stepwise Regression.
3. Improving data claening techniques. The original data cleaning is too simple for such a complex data sets with numerious variables and observations.
4. Model solely built to calculate possible election outcome for Liberal Party. For the "full picture", next model should include regression estimates for all major parties, like Conservaties and Bloc Québécois.

References

1. 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
2. "General Social Survey, Cycle 31 : Families." Statistics Canada, Minister Responsible for Statistics Canada.
3. 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. LINK: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DUS88V

4. "2015 Canadian federal election." Wikipedia, https://en.wikipedia.org/wiki/2015_Canadian_federal_election. Used 19th December 2020.

5. "2019 Canadian federal election." Wikipedia, https://en.wikipedia.org/wiki/2019_Canadian_federal_election. Used 19th December 2020.

6. Dunnington, D., electionca package. 22 January 2020. Retrived from https://github.com/paleolimbot/electionca.

7. R Core Team. (2020). The R project for statistical computing. Retrieved from https://www.rproject.org/

8. Wickham, H., Averick M., Bryan J., Chang W., McGowan, L. D., Francois R., Grolemund G., Hayes A., Henry, L., Hester J., Kuhn M., Pedersen T. L., Miller E., Bache, S. M., Muller, K., Ooms J., Robinson, D., Seidel, D. P., Spinu, D.,. . . Yutani, H. (2019). Welcome to the Tidyverse. The Journal of Open Source Software. Retrieved from https://joss.theoj.org/papers/10.21105/joss.01686

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

10. Xie, Y. (n.d.). Knitr v1.30. Downloaded October 18, 2020, from https://www. rdocumentation.org/packages/knitr/versions/1.30

11. Alexander, R., and Caetano, S.(2019, Sept 16). "gss_cleaning.R". Used December 19, 2020, from https://www.tellingstorieswithdata.com/01-03-r_essentials.html

12. Kassambara, A. ggpubr package. Retrieved from https://github.com/kassambara/ggpubr