Forecasting if 2019 Canadian Federal Election would have been different if everyone had voted Using Multilevel Regression with Post-Stratification

SHIYUN TANG

12/22/2020

The code and data that support this analysis can be found at: https://github.com/SHIYUN-TANG/STA304FINAL

Abstract

Since not everyone vote in the Canadian federal election 2019, result of election doesn't show the actual voting choice of all the Canadian citizen. To fix this problem we are going to do the forecasting to assume 'everyone' vote in 2019 Canadian Federal Election to see if there would be a difference Using Multilevel Regression with Post-Stratification. We are going to use a build a multilevel logistic model by using the predictor sex, age, education and province from the survey data and to do the post-stratification using the census data.

Keywords

Canadian federal election, Liberal party, Conservative party, Multilevel regression and post stratification, Election forecasting

Introduction

Canada is constitutional monarchy, composed of the Queen of Canada, who is officially represented by the Governor General, the Senate and the House of Commons(Laws-lois.justice.gc.ca. 2020). The House of Commons has 338 seats, held by members elected by citizens who vote in general elections and there are 338 electoral districts, each with a corresponding seat in the House of Commons (the Constitution Act, 1867). Every Canadian citizen equal and over 18 has the right to vote for Canadian federal election (Laws-lois.justice.gc.ca. 2020), but not everyone who has the right to vote will vote. The data shows 77 percentage of Canadians reported voting in the 2019 federal election, which is same as 2015 election (Statistic Canada, 2019). The Liberal party own 157 seats win the election and the Conservative party own 121 seats was the 2nd in the election (Global News, 2019). The difference of number of seats owned by these two party is smaller than in previous elections (Zimonjic, Peter 2019). Since lower participation could lead to changes in election result. (Archer 2003), we are going to find out whether Laberal party would win if everyone has the right to vote vote in the Canadian federal election in 2019.

To forecast if there's a difference if 'everyone' had voted in 2019 Canadian federal election, we use Multilevel Regression with Post-Stratification (MRP). MRP is a common technique when it comes to forecasting issues

concerning politics. It is used to obtain accurate micro-level subgroups estimates, which are then combined with subgroups' weight in the population by certain mathematical formula to get a macro-level estimate (Ghitza & Gelman, 2013; Lax & Phillips, 2009; Park et al., 2004). To using MRP, we first build a multilevel logistic model by using a survey data then post-stratification by using a census data. Age, sex and education has been found to matter a great deal to the voting option (Elections Canada 2010) so that we use this two variable to build the model to predict the result. In addition, each province has different number of electoral districts would have different voting intentions thus we decide to also use it as a predictor.

We will describe the data we use and the model we build in the Methodology section and the result we get after using multilevel regression with post-stratification will be shown in the Result section. Also, we are going to present the whole analysis and conclude in the Discuss section.

Methodology

Data

For the survey data, we use the survey from Canadian Election Study which is the Online Survey which shows the attitudes of Canadians during and after the 2019 election (Stephenson, 2020). For the census data, we use the Education Highlight Tables from Statistic Canada which provide counts and percentage distributions for various geographic levels by highest level of educational attainment, sex and selected age groups for the 2016 Census (Statistic Canada, 2019). Since the census program provides a statistical portrait of the country every five years, The census data from 2016 is the latest census data we can find. We make a assumption that the data stay the same from 2016 to 2019. Also, the census data is unbiased and representative.

We know that certain demographic characteristics, such as age, sex, province, can significantly influence people's voting preferences (Kaufmann & Petrocik, 1999; Penney et al., 2016; Godek, 2018). We pick variable sex, age groups, province and education level from both survey data and census data and these four variables appears in both data. Also in we pick the categorical variable vote choice and only use the data who vote Liberal party and Conservative party because we want to know if Liberal party will still have more seat than Conservative party if 'everyone' vote in the election. Since age is a numerical variable in survey data but is categorical (collect as group) in census data, we group the age in the survey data. Also, we rename the factor name in each variable to make the factor name be the same. Finally Variable age groups, province and education level are categorical variables and sex is a dummy variable.

Model

The four predictor we use in the model is sex, age groups, province and education level. First, we partition the survey data into different level 2 cells based on two variables, province and sex. In our model, the slope and intercept of variable education are vary between cells which means between different province with different sex. The variable education is both a fixed and a random effect. The other variable sex, province, agegroup are fixed in the model. We use these predictors to build a multilevel regression model.

The formula of the multilevel regression model is $Y_{ij} = \alpha+(\beta+b_j)\, x_{ij}+a_j+\epsilon_{ij}$, where i stands for the individual cell member, j stands for the cell and α and β are coefficient baselines which do not vary across the cells. b_j and a_j are random variables that make our estimator's intercept and slope vary across different cells; ϵ_{ij} is the residual term, which is different in different cells. The model we create here is $Y_{ij} = \alpha+(\beta+b_j)\, x_{ij}+a_j+\epsilon_{ij}$.

The model we create can be written as:

$VoteChoice = logit^{-1}(Sex+Province+AgeGroup+Education+Education|cells)$

Which also can be written as:

$log(\frac{p}{1-p}) = \alpha + a_j + \beta_1\ x_1 + \beta_1\ x_1 + \beta_2\ x_2 + \beta_3\ x_3 + \beta_4\ x_4 + \beta_5\ x_5 + \beta_6\ x_6 + \beta_7\ x_7 + \beta_8\ x_8 + \beta_9\ x_9 + \beta_{10}\ x_10 + \beta_{11}\ x_11 + \beta_{12}\ x_{12} + \beta_{13}\ x_{13} + \beta_{19}\ x_{19} + \beta_{20}\ x_{20} + b_{ij}\ x_{ij} + \epsilon_{ij}$.

where:

p: the probability people vote liberal party.

$log(\frac{p}{1-p})$: log odds of voting liberal party.

α: log odds of voting liberal party when person is female in Alberta with low eduction between 25 to 34. β_1: log odds of voting liberal party when person is male in Alberta with low education between 25 to 34.

β_2: log odds of voting liberal party when person is female in British Columbia with low education between 25 to 34.

β_3: log odds of voting liberal party when person is female in Manitoba with low education between 25 to 34.

β_4: log odds of voting liberal party when person is female in New Brunswick with low education between 25 to 34.

β_5: log odds of voting liberal party when person is female in Newfoundland and Labrador with low education between 25 to 34.

β_6: log odds of voting liberal party when person is female in Northwest Territories with low education between 25 to 34.

β_7: log odds of voting liberal party when person is female in Nova Scotia with low education between 25 to 34.

β_8: log odds of voting liberal party when person is female in Nunavut with low education between 25 to 34.

β_9: log odds of voting liberal party when person is female in Ontario with low education between 25 to 34.

β_{10}: log odds of voting liberal party when person is female in Prince Edward Island with low education between 25 to 34.

β_{11}: log odds of voting liberal party when person is female in Quebecland and Labrador with low education between 25 to 34.

β_{12}: log odds of voting liberal party when person is female in Saskatchewan with low education between 25 to 34.

β_{13}: log odds of voting liberal party when person is female in Yukon with low education between 25 to 34.

β_{14}: log odds of voting liberal party when person is female in Alberta with mid education between 25 to 34.

β_{15}: log odds of voting liberal party when person is female in Alberta with mid-high education between 25 to 34.

β_{16}: log odds of voting liberal party when person is female in Alberta with high education between 25 to 34.

β_{17}: log odds of voting liberal party when person is female in Alberta with very high education between 25 to 34.

β_{18}: log odds of voting liberal party when person is female in Alberta with low education between 35 to 44.

β_{19}: log odds of voting liberal party when person is female in Alberta with low education between 45 to 54.

β_{20}: log odds of voting liberal party when person is female in Alberta with low education between 55 to 64.

x_1: Equals to one when sex is male, 0 otherwise.

x_2: Equals to one when province is British Columbia, 0 otherwise.

x_3: Equals to one when province is Manitoba, 0 otherwise.

x_4: Equals to one when province is New Brunswick, 0 otherwise.

x_5: Equals to one when province is Newfoundland and Labrador, 0 otherwise.

x_6: Equals to one when province is Northwest Territories, 0 otherwise.

x_7: Equals to one when province is Nova Scotia, 0 otherwise.

x_8: Equals to one when province is Nunavut, 0 otherwise.

x_9: Equals to one when province is Ontario, 0 otherwise.

x_{10}: Equals to one when province is Prince Edward Island, 0 otherwise.

x_{11}: Equals to one when province is Quebecland and Labrador, 0 otherwise.

x_{12}: Equals to one when province is Saskatchewan, 0 otherwise.

x_{13}: Equals to one when province is Yukon, 0 otherwise.

x_{14}: Equals to one when education level is mid education, 0 otherwise.

x_{15}: Equals to one when education level is mid-high education, 0 otherwise.

x_{16}: Equals to one when education level is high education, 0 otherwise.

x_{17}: Equals to one when education level is very high education, 0 otherwise.

x_{18}: Equals to one when age is between 35 to 44, 0 otherwise.

x_{19}: Equals to one when age is between 45 to 54, 0 otherwise.

x_{20}: Equals to one when age is between 55 to 64, 0 otherwise.

After cell division and modeling, we will perform post-stratification. Performing post-stratifiction is to get the population estimate by calculating the weighted average of all cell-level estimates obtained from our built multilevel logistic model with the following mathematical formula: $\hat{y}^{PS} = \frac{\sum_j N_j \cdot \hat{y_j}}{\sum_j N_j}$ whrere $\hat{y}^{PS}$ is the estimate in each hell and N_J is the population size of the J^{th} cell based off demographics. It shows how entire population will vote.

Results

$\alpha = -0.91994$, $\beta_1 = -0.28117$, $\beta_2 = 1.33178$, $\beta_3 = 0.83265$, $\beta_4 = 1.64576$, $\beta_5 = 1.92745$, $\beta_6 = 1.99772$, $\beta_7 = 2.03865$, $\beta_8 = 0.5851$, $\beta_9 = 1.50622$, $\beta_{10} = 1.87134$,$\beta_{11} = 1.95906$, $\beta_{12} = -0.27678$, $\beta_{13} = 1.18107$, $\beta_{14} = -0.55237$, $\beta_{15} = -0.28051$, $\beta_{16} = 0.17606$, $\beta_{17} = 0.471$, $\beta_{18} = -0.15899$, $\beta_{19} = -0.27135$, $\beta_{20} = -0.18921$. β_8, β_{12}, β_{13}, β_{15}, β_{16}, β_{17} has p-values are greater than 0.05 which means does not have a significant impact on the voting choice. $\alpha = -0.91994$ means that when the values of all explanatory variables are 0, our estimated value for log odds of voting is -0.91994; however, this number does not have any practical meaning here. For all those numerical values of β_i, where i = 1, ..., 20, they represent the value of change when there is a one unit increase in their assigned variables.

The $\hat{y}^{PS}$ value is 0.52 which means according to our post-stratification analysis after building a multilevel logistic regression model using the predictors age, education level, province and sex, the expected proportion of entire population will vote Liberal Party is 0.52.

predicted numbers of vote in each province table:

Province	Liberal Party	Conservative Party	Winner	Seats
Alberta	0	2257820	Conservative Party	34
British Columbia	1069210	1463765	Conservative Party	42
Manitoba	167390	487265	Conservative Party	14
New Brunswick	288845	111915	Liberal Party	10

Province	Liberal Party	Conservative Party	Winner	Seats
Newfoundland and Labrador	264005	23860	Liberal Party	7
Northwest Territories	23120	520	Liberal Party	1
Nova Scotia	494530	0	Liberal Party	11
Nunavut	460	16030	Conservative Party	1
Ontario	4790285	2438855	Liberal Party	121
Prince Edward Island	67025	7285	Liberal Party	4
Quebecland and Labrador	4371950	0	Liberal Party	78
Saskatchewan	0	566235	Conservative Party	14
Yukon	7435	13605	Conservative Party	1

In total, Liberal party own 232 seats, and Conservative party own 106 seats where liberal party has a huge advantage in Quebec and Nova Scotia and Conservative party has a huge advantage in Alberta and Saskatchewan. The final result is the same as actual result in 2019 and the the result per province is very similar to the actual result. Therefore, although 'everyone' vote in the election in 2019, the result might not change.

Discussion

Summary

To see if 2019 Canadian Federal Election would have been different if everyone had voted, we use the census data Highest level of educational attainment by sex and selected age groups ans use the survey data from Canadian Election Study which shows the attitudes of Canadians during and after the 2019 election to create a multilevel regression model. We use sex, age, education, province as the predictors to analyze the model. The slope and intercept of predictor education are vary between cells which contains sex and province and all the predictors are fixed while education is both fixed and random effect. After building the model, We carried out the post-stratification in which we applied the model to the census data. Eventually, we predicted the proportion of 'everyone' would vote the Liberal party is 0.52 and the Liberal party will win the election.

Conclusions

Even though the proportion of people vote might make the result of the election different. The result might not change while 'everyone' who has the right to vote vote in 2019 which Liberal party will win the election, even the advantage of the Liberal party has grown.

Weakness & Next Steps

During our work, there's several weakness. First the census dataset we find might cause some bias. The variable age in our census data groups by every ten years but since only person who is 18 years old or older is qualified as an elector. We can only avoid the agegroup that contains people from 15 years old to 25 years old or contains people under 18 years old who actually don't have the right to vote. By this action, we would miss the data that people who vote between 18 and 25. Also, the survey dataset is not big enough so that the model we built might be inaccurate. Furthermore, since people who vote Liberal Party and Conservative Party might vote other party other than these two, but the logistic model does not cover this situation.

To improve our study, we may find a census dataset or survey dataset which have more predictor we can use since in the model we create, we use only 4 predictors. It also enable us to create a more complicated model contains more cells so that we can build a more accurate model. Another way to find a more accurate model is to find a larger dataset and use larger number of cells to see if the result stay the same or not.

Also, we can use Multilevel Regression with Post-Stratification by Bayesian approach to see if a proper prior distribution can give as a more accurate result. Furthermore, we might use the non-logistic multiple model to predict the election result to see the estimate proportion for all the party includes in the election to see if the estimate result is different.

References

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

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

Statistics Canada. Education Highlight Tables, 2016 Census, "Highest level of educational attainment (general) by sex and selected age groups" https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/edu-sco/index-eng.cfm

Joseph Larmarange (2020). labelled: Manipulating Labelled Data. R package version 2.7.0. https://CRAN.R-project.org/package=labelled

Hadley Wickham and Evan Miller (2020). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 2.3.1. https://CRAN.R-project.org/package=haven

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

Statistics Canada."Reasons for not voting in the federal election, October 19, 2015" Component of Statistics Canada catalogue no. 11-001-X

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.

Ghitza, Y., & Gelman, A. (2013). Deep interactions with MRP: election turnout and voting patterns among small electoral subgroups. American Journal of Political Science, 57(3), 762–776.

Lax, J. R., & Phillips, J. H. (2009). How should we estimate public opinion in the states? American Journal of Political Science, 53(1), 107–121.

Park, D. K., Gelman, A., & Bafumi, J. (2004). Bayesian multilevel estimation with poststratification: state-level estimates from national polls. Political Analysis, 12(4), 375–385.

Kaufmann, K. M., & Petrocik, J. R. (1999). The changing politics of American men: understanding the sources of the gender gap. American Journal of Political Science, 43(3), 864–887.

Penney, J., Tolley, E., & Goodyear-Grant, E. (2016). Race and gender affinities in voting: experimental evidence. Queen's University. https://www.econ.queensu.ca/research/working-papers

Godek, P. E. (2018). Determining state preferences for the electoral college: 1788-2016. The Cato Journal, 38(3), 631+.

Laws-lois.justice.gc.ca. 2020. Canada Elections Act. [online] Available at: https://laws-lois.justice.gc.ca/eng/acts/e-2.01/fulltext.html

"Canada election: The 2019 results by the numbers". Global News. Retrieved October 27, 2019.

Zimonjic, Peter (October 22, 2019). "Liberals take losses but win enough in Quebec and Ontario to form minority government". Canadian Broadcasting Corporation. Retrieved October 21, 2019.

Archer, Keith. 2003. "Increasing youth voter registration: Best practices in targeting young electors." Electoral Insight.Vol. 5, no. 2. July. Elections Canada. p. 26-30. (accessed February 8, 2012).

Elections Canada. 2010. Estimation of Voter Turnout by Age Group at the 2008 Federal General Election. Working Paper Series. Ottawa. 12 p. (accessed February 8, 2012).

The Constitution Act, 1867, 30 & 31 Vict, c 3, http://canlii.ca/t/ldsw retrieved on 2020-12-22