Predicting the Percentage of Votes Obtained by Liberal in 2019 Canadian Federal Election If Everyone Can Vote Using Logistic Regression With Post-stratification

Xi Zheng 1005153628

2020/12/21

Code and data supporting this analysis is available at: https://github.com/xixiaoguai727/304FinalProject

Abstract

Liberal won the 2019 Canadian Federal Election, but Liberal receives fewer votes compared to Conservative. Some people may wonder, if everyone in Canada has the right to vote, would the Liberal Party still have a lower vote rate than the Conservative? In this report, a logistic regression model with post-stratification is built and aim to estimate Liberal's percentage vote obtained if everyone in Canada has the right to vote. After doing the statistical analysis, we found that if everyone can vote in Canada, liberals would still receive less vote than conservatives.

Key Words

Canadian Federal Election, Logistic Regression, Akaike Information Criterion, Stepwise regression, Post-stratification

Introduction

As we all know Liberal won the 2019 Canadian Federal Election but got fewer votes than Conservative. To register and vote in a federal election, "voter must be a Canadian citizen aged 18 or older on election day and provide acceptable proof of identity and address"[1]. "Canada is globally known for being a welcoming and accepting country and now it is being recognized as one of the most diverse countries in the world"[2]. Many people living in Canada have not yet obtained a nationality. According to 2016 Census data[3], around 7% of people in Canada are not Canadian citizens, however, the election results would have a great impact on the future lives of these people.

Some people may wonder, if everyone in Canada has the right to vote, will the percentage vote obtained by Liberal be different? This report aims to investigate whether the election results will be different if everyone in Canada has the right to vote by logistic regression with post-stratification.

Two data sets will be used to do the prediction. Campaign Period Survey in 2019 Canadian Election Study Online Survey[5] is used as a survey sample to build the logistic regression model. 2016 General social survey on Canadians at Work and Home[4] is used as census data for the post-stratification part. The methodology section below describes the specific steps used to build the logistic model with post-stratification based on the above two data sets. The results obtained by the model and statistical analysis are in the result section. Related discussion and limitations are presented in the discussion section.

Methodology

Data

Campaign Period Survey(CPS) in 2019 Canadian Election Study Online Survey contains 37,822 observations are used as a survey sample to build the logistic regression model instead of the Post-Election Survey(PES). The reason why use CPS instead of PES is CPS contains nearly four times more observations than PES, a larger sample size will produce a better model. People may vote for a different party than they stated in the Campaign Period Survey, this is one of the drawbacks of using the Campaign Period Survey.

2016 General social survey on Canadians at Work and Home with 18,249 observations is used as census data for post-stratification part. The reason why choose the 2016 General social survey(GSS) on Canadians at Work and Home instead of actual Canada census data is GSS has lots of same answer options as the 2019 Canadian Election Study Online Survey, so it will be more convenient to use GSS for analysis than the actual census data.

Vote intention for all types(citizen, pr, non-citizen) of respondents is recorded in a new variable named vote. To form the logistic regression, variable voteLiberal is created and it stores binary value. 1 stands for the corresponding respondent who wants to vote for Liberal and 0 stands for he/she does not want to vote for Liberal. The dependent variable being used in the regression is voteLiberal.

Gender, marital status, employment, province, education, and income are six independent variables selected to form the logistic regression. Observations choose NA(not available) or reject to answer questions related to those six variables are removed from the data set. Besides, answer options of CPS and GSS data sets are rearranged to the same format. For example, the answer option to the province in the GSS data set does not contain territories(like Northwest Territories, Yukon), thus, observations choose territory in the CPS data set are removed.

Those independent variables are all categorical variables, and five of them are shown in Table 1. Province variable includes Alberta, British Columbia, Manitoba, New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario, Prince Edward Island, Quebec, and Saskatchewan.

Table 1: Five Independent Variables Used in the Regression Model

Gender	Marital	Employment	Education	Income
male	Married	Working at a paid job or self-employed	Less than high school diploma or its equivalent	below $60,00
female	Living common-law	Looking for paid work	High school diploma or high school equivalency certificate	$60,000 - $10,000
	Divorced	Going to school	Trade certificate or diploma	above $10,000
	Separated	Household work	College, CEGEP or other non-university certificate	
	Widowed	Retired	University certificate or diploma below the bachelor's level	

Gender	Marital	Employment	Education	Income
	Single, never married	Long term illness	Bachelor's degree (e.g. B.A., B.Sc., LL.B.)	
		Other	University certificate, diploma or degree above the bach. . .	

Model

We are interested in how independent variables, like gender, marital status, employment, province, education, income, affect people's vote intention in the 2019 Canadian Federal Election. Thus, the logistics regression model is the proper model to see the probability of a random person in Canada who would like to vote for Liberal.

We build the logistic regression using RStudio. We firstly build a model using all of those six predictor variables we have. Then, we use backward stepwise regression by the Akaike Information Criterion(AIC) to get our final model. "AIC is a criterion for assessing models, and it balances the goodness of fit and a penalty for model complexity"[7]. The smaller the value of AIC, the better the model. "Backwards stepwise regression starts with all the potential terms in the model, then removes the term with the largest p-value each time to give a smaller information criterion".[7]

Our original model containing six variables has an AIC value of 9231.3, and after doing the step-wise regression we get a new model with only four variables (gender, marital, edu, province), and it has a smaller AIC value which is 9223.31. So we choose to use the new model to do the left analysis.

Model Specifics

This is the equation we get:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{male} + \beta_2 x_{Livingcommon-law} + \beta_3 x_{maritalMarried} + ... + \beta_{21} x_{Saskatchewan}$$

p stands for the probability of a random person in Canada would vote for Liberal. Variable Living common-law means, if the person's marital status is Living common-law, then the value of $x_{Livingcommon-law}$ will be 1, otherwise, it will be 0. The same word applies to all 21 x variables. β_1 to β_{21} in Table 2(result section) represents 21 beta coefficients in our equation. β_0(-1.61948) is the intercept parameter. When all 21 x variables equal to zero, only β_0 is left, after taking log-off, we get p equals to 0.17. This can be interpreted as the probability of a widowed female with a Bachelor's degree lived in Alberta vote a Liberal is about 17%.

Next, we perform a post-stratification analysis using the same four predictor variables in our cleaned GSS data set and the regression model we got above to get the estimated proportion of voting for Liberal.

Results

Table 2: Beta Estimates

		Estimate			Estimate
B0	intercept	-1.61948	B12	eduUniversity certificate, diploma or degree above	0.10430
B1	gendermale	-0.08174	B13	provinceBritish Columbia	0.64057
B2	maritalLiving common-law	-0.27822	B14	provinceManitoba	0.50147
B3	maritalMarried	-0.06183	B15	provinceNew Brunswick	0.65709
B4	maritalSeparated	-0.01682	B16	provinceNewfoundland and Labroador	1.21394
B5	maritalSingle, never married	-0.06587	B17	provinceNova Scotia	1.07504
B6	maritalWidowed	0.02460	B18	provinceOntario	1.00616
B7	eduCollege, CEGEP or other non-university certificate	-0.45218	B19	provincePrince Edward Island	0.76154
B8	eduHigh school diploma or high school equivalency certificate	-0.44551	B20	provinceQuebec	0.86776
B9	eduLess than high school diploma or its equivalent	-0.54169	B21	provinceSaskatchewan	-0.18219
B10	eduTrade certificate or diploma	-0.32776			
B11	eduUniversity certificate or diploma below the bachelor's level	-0.05432			

The post-stratification result we get is 0.4698, which means the estimated proportion of voting for Liberal is 46.98% if everyone in Canada would vote. This is based off our post-stratification analysis of the proportion of voters in favour of Liberal modelled by our logistic model, which accounted for gender, marital status, education, province. The result is 13.86% higher than the actual vote percentage(33.12%)[8] Liberal received in the election (46.98% - 33.12% = 13.86%).

To refer, we repeated the same steps in the Model section to get a logistic regression model for Conservative and after doing the post-stratification, we get the result 0.4906, which means the estimated proportion of voting for Conservative is 49.06%.

It's clearly that vote obtained by Liberal would increase if everyone in Canada can vote, but Liberal still win 2.08% less votes than Conservative(46.98% - 49.06% = -2.08%). The vote percentage difference is 0.86% larger than the actual difference 1.22% (2.08% - 1.22% = 0.86%).

Discussion

Summary

We firstly clean CPS and GSS data set to our desired form, then we build a logistic regression by RStudio and do the backward stepwise regression to obtain the best model we could have by using six independent variables. Next, we using the prediction model for the GSS data set to get the estimated proportion of voting

for Liberal is 46.98%. We repeat the same steps for Conservatives and get 49.06% as the estimated voting proportion.

Conclusion

Our result shows that Conservative get 2.08% more votes than Liberal. Compared to the actual percentage difference of 1.22%, Conservative still be more popular than Liberal. However, the election result is not based on how many votes each party wins, is based on how many seats each party wins. Based on the actual results, Liberal got 184 seats compared to Conservative got 99, but the probability of 0.86% vote change the election result is very low. Therefore, no matter whether everyone in Canada can vote, Conservative is always more popular than Liberal, but Liberal would win the election anyway.

To sum up, whether there exists eligibility requirements for voters will not change the final election result and vote percentage. If everyone in Canada can vote, Liberal would still win the election with the percentage vote obtained less than Conservative. Thus, existing policies in Canada would continue, such as the legalization of cannabis, and would not be improved.

Weakness & Next Steps

There are still some weaknesses and limitations in our analysis. As we mentioned in the data section, people may vote for a different party than they stated in the Campaign Period Survey, which lowers the result accuracy of our prediction model. The census data set we use is from the 2016 General social survey, which is obtained three years before the election happens, demographic characteristics may change in those three years. Also, the number of observations in the cleaned GSS data set is smaller than the cleaned CES data set, which makes our post-stratification part become less accurate. Therefore, the 2016 GSS survey is not a very perfect census data set.

There are still a few things we could do in the future to make improvements. We can find a larger and nicer data set as our census data, find more significant predictor variables, and do the regression again to get a better result. We also can compute the number of seats each party would obtain if every in Canada can vote to see if the result would be different.

References

[1]Facts about voter registration, citizenship and voter ID. (n.d.). Retrieved December 09, 2020, from https://www.elections.ca/content.aspx?section=med

[2]Government of Canada, S. (2019, June 18). Census Profile, 2016 Census Canada [Country] and Canada [Country]. Retrieved December 21, 2020, from https://www12.statcan.gc.ca/census- recensement/2016/dp-pd/prof/details/Page.cfm?Lang=E

[3]Dailyhive. (2019, June 08). News. Retrieved December 21, 2020, from https://dailyhive.com/vancouver/canada-most-diverse-countries-ranking-2019

[4]2016 General social survey on Canadians at Work and Home. Retrieved December 21, 2020, from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/hsda?harcsda4+gss30

[5]2019 Canadian Election Study Online Survey. Retrieved December 21, 2020, from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DUS88V

[6]Wikipedia contributors. (2020, December 17). 2019 Canadian federal election. In Wikipedia, The Free Encyclopedia. Retrieved 14:47, December 21, 2020, from https://en.wikipedia.org/w/index.php?title=2019_Canadian_federal_election&oldid=994820597

[7]Sue-Chee, S. (2020) STA302/1001-Methods of Data Analysis I[PowerPoint presentation]. Retrieved from https://q.utoronto.ca

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