Canadian Federal Election Result Everyone Inclusive Predicted by Logistic Regression Model With Post-stratification Technique

Michael Huang

December 7, 2020

Code and data supporting this analysis is available: https://github.com/michaelhuang1/canadian-election-result-if-everyone-voted

Abstract

In this analysis, the outcome of interest is the voting outcome of the 2019 Canadian federal election if votes from everyone are counted for including all ages. Several predictors will be used such as age, sex,province and education in the logistic model then a post-stratification technique will be used. It was found that if everyone voted, the liberal party would have gain slightly more votes. Knowing this helps with making more precise prediction of the next election and possibly help candidates know their positions.

Keywords

Election, Liberal , everyone voted, logistic model, post-stratification

Introduction

Voting is a way for citizens to select the leader for their country, however different countries have different laws against voting. For Canadians, citizenship and the age requirement of eighteen must be met in order to vote. Statistically, the percentage of voting turnout is only around 60% to 70%. Which means that's a big proportion of the population decides not to vote in an election. Logistic model is the correct model to apply when there is a binary outcome. In this case, the outcome is either vote for liberal party or do not vote for liberal party. Since the outcome is either 0 or 1, linear models will not give accurate predictions as most of the observations are far away from the line, hence logistic regression makes predictions more precise. Next,post-stratification technique will be applied to decrease the variance and bias of the predictors selected in the mode Two data set will be used to investigate the outcome of 2019 election if everyone had voted. Including people that did not vote and people that are under age. In the next section(Methodology Section), it will include description of the regression model and data that are being used to perform this analysis. Result of whether Liberal party wins the election will be shown in the Result section. Any inferences and conclusions of the analysis will be included in the discussion section of this report.

Methodology

Data

In this analysis two data set are being used. The gss dataset from 2017 from the chass website have been used as the census data and the ces web survey has been used as the survey data. The gss data is a general social survey collected from telephone interview and the web where the ces data are collected from purely online questionnaires. The population for the gss data is all Canadians over 15 years old with a sample of 20,000 individuals and the frame being telephone numbers. Both data have been cleaned down to having the same variables with the same possible values. Note that there is an extra variable of "liberal" in the survey data which represents the person having liberal party as their first choice if the value is one, otherwise it is a 0. Those variables are chosen because they are highly relevant to an individual's voting choice.Also need to take into account that the age in the data set includes ages below 18 which is how everyone is defined in this analysis.

Model

In this report, the purpose is fo find out what would have happened if everyone voted. A logistic regression model would be the technique to use for binary outcomes. Historically, the voting outcome differs significantly by province and age. Also sex and education are some variables that would effect the election. To take into account this difference in sample and target population, it suffice to employ a post-stratification technique to decrease the variance and bias of the predictors selected in the model. First, partition the population into cells based on multiple demographic and geographic characteristics. Next, use the sample (survey data) to estimate the response variable (probability of Liberal party winning the election) within each cell. Finally use the census data from 2017 to aggregate the cell-level estimates up to a population-level estimate by weighting each cell by it's relative proportion to the entire population. In the following subsections model specifics and the post-stratification calculation will be discussed in detail. This prediction uses a logistic regression model to model the proportion of Canadians who have liberal party as their first voting choice with the software R studio. In order to model the proportion of voters that vote for liberal party and to obtain relatively precise conclusion, some re-arrangements are made on these data. It is reasonable to see if one person votes for liberal party or not(which is a binary response) so all votes that don't have liberal party as their first choice are represented with0, also "“Don't know/ Prefer not to answer" have been filtered out as well. A model is build using of the glm() function in the R package glm2 (Marschner,2011) to build the logistic regression model which is:

$$Pr(y_i) = log(\frac{y_i}{1-y_i}) = \beta_o + \beta_1 x_{age} + \beta_2 x_{sex} + \beta_3 x_{province} + \beta_4 x_{education} + \epsilon$$

Here y_i represents the proportion of voters who will vote for Liberal party. Four predictors are included in this model: age, sex, province and education. β_0 represents the intercept of the model, and is the probability of voting for liberal party at age 0. Additionally, β_1 represents the slope of the first predictor. For everyone unit increase in age, it is expected to have an increase in the probability of voting for Liberal party. Similarly, β_2 represents the slope of the sex variable. Since β_3 refers to the slope of province, one of the province will be chosen as base line where model outcome changes for all other provinces. β_4 represents the slope of education, similarly it also chooses one value as the base line where other values will have some impact on the model.

The post-stratification estimate can be noted as:

$$\hat{y}^{ps} = \frac{\sum N_j \hat{y_j}}{\sum N_j}$$

where $\hat{y_j}$ represents the estimate in each cell j, and N_j represents the size of the j^{th} cell in the population.

We are including all the variables to create cells because they are all likely to influence voter outcome for this election. For instance, if the candidate is from a certain city and mostly likely that city would have higher

percentage of people voting for him. Thus, people from different provinces will make different decisions on whether they should vote for this candidate or not. Education level of voter can cause large divergence as well since people at different education level will have different interpretation on politics. Likewise, people from similar age range might have similar preferences on candidate selection. Also, the gender group of one candidates' voters can be influenced by which gender group the candidate is in. People are likely to support the candidate who has the same gender as them.

#Result The outcome of the analysis shows that the percentage of vote liberal party receives if everyone had voted is around 33.5%. This is based off the post-stratification analysis of the proportion of voters in favor of liberal party modeled by a logistic regression model, which accounted for 4 variables, age, sex, the province they live in and their education levels in the model. As in real life data, the liberal party received 33.1%(Britneff 2019), slightly less than the predicted value.

From the above table, age and sex do have somewhat small p value. However p-values for provinces are quite large for most of the provinces. Moreover most education levels have p-values around -0.5.

In order to determine the significance of four independent variable (age, sex, province and education), an ANOVA table for this logistic regression model is applied. From this above table, it illustrates that the residual deviance of the null model with just an intercept is 33701. Every addition of predictors decreases the residual deviance. As more predictors of age,sex,province and educations are being added in, the residual deviance decrease from 33701 to 33659 to 33650 to 32792 and finally 32519.

Discussion

Since there are around 30% population that does not vote in federal election, it would be interesting to investigate if there are any impact on the result if everyone votes. In this analysis everyone is accounted for. The first choice of all persons are used in the model despite the person actually voted or not hence everyone is included. Four predictors were selected, age, sex, province and education level. The estimator is the proportion of votes that the liberal party would receive in 2019. A logistic regression model was used for this binary estimator. Observing the p-value of each predictor of the model, it shows that four predictors all have an effect on the proportion of voting for the liberal party. Nevertheless, some of the predictors contain a somewhat large p-value. In particular, we apply an ANOVA table. It shows that compared to the null model, the residual deviance of the logistic regression model decreases by the addition of these four predictors. which suggests that the chosen variables can be used to predict the election. After applying the logistic model, we make use of the post-stratification technique, finally we obtain the proportion of votes liberal party would receive is 33.5%.

In conclusion, since the proportion of votes for liberal party is 33.5%, the liberal party would still win the primary vote and the actual proportion of voters in favor of voting for liberal party is 33.1% which is slightly less than the predicted value. This suggest that the more people who did not actually vote had liberal party in mind as their first choice. Since liberal party in reality has received more votes than other parties, it suffices to say that most likely people who did not vote had a family member or close friend who voted for liberal party. Hence if their votes were to be counted, liberal party would be their choice. Also since this analysis included people from age 15 to 18. They fall into the group of people that have no opinion on politics. Therefore if their votes were to be counted there would have a decent chance of being liberal party. As the party leader is Justin Trudeau. On top his good looks, he also have a strong politic background since his father is a well respected politics person. Hence the increase in the proportion can be explained.

Weaknesses

There is only four predictors in the model which means there might be some important factors that are left out and should be included. Since the definition of everyone includes all ages, in the data used in this analysis it only included observations from 15 and up. Some small age groups are being left out which is

a lot of people and their votes would make a difference. Also their might exist variable correlations which should be accounted for and might slightly affect the prediction.

Next steps

It is sufficient for us to select more predictors in this study, which can help reduce the errors for prediction. It would also be a good idea to gain data regarding people from smaller age group. The correlation between each variables should be checked. The best way to study the result if everyone voted for the election is to get everyone to vote. More people should be encouraged to participate in voting. In addition, AIC and BIC model selection techniques could be applied to select the best model for our study.

References

General Social Survey-Family(GSS). Statistics Canada. https://bit.ly/2T8PrNa

General Social Survey on Family (cycle 31), 2017. Statistics Canada. http://dc.chass.utoronto.ca/myaccess.html

Hosmer, D. & Lemeshow, S. (2000). Applied Logistic Regression (Second Edition). New York: John Wiley & Sons, Inc.

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: AGrammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr

Beatrice Britneff (2019) Canada election: The 2019 results by the numbers https://globalnews.ca/news/6066524/canada-election-the-2019-results-by-the-numbers/

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." Journal of Open Source Software, 4(43), 1686. doi: 10.21105/joss.01686.

Hadley Wickham and Evan Miller (2020). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version2.3.1. https://CRAN.R-project.org/package=haven Marschner, I. C. (2011). glm2: Fitting Generalized Linear Models with Convergence Problems. The R Journal 3(2): 12-15

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.

R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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