Conservative Party estimated to win 2019 Canadian Federal Election with full voter turnout

Muhammad Tsany

12/8/2020

Abstract

An analysis of the 2019 Canadian Federal Election and the outcome that would have occurred if every eligible voter had voted is done. Specifically, the chosen predictor variables that will be used for analysis are an individual's unique characteristics, such as education and income. This report will demonstrate that the election would have had similar outcomes, but more importantly, it sheds light upon the importance of voter turnout.

Introduction

Voting eligibility has been more important in the current decade more than ever. Voter turnout is affected highly by "a number of sociodemographic characteristics, particularly age and education," (Blais, 2005). So, there are numerous variables that affect voter turnout. In this paper, the 2019 Canadian Federal Election will be analyzed such that every eligible voter will have voted. In this case, voter turnout for this election will be nearly 100%.

So, a multilevel logistic regression with poststratification (MRP) model based off survey data given by Canadian Election Study (CES) and census data given by General Social Survey (GSS) will be post-stratified. The CES dataset will be based off the online survey. This allows estimates based on certain characteristics, such as education and income levels, that can infer voter's intentions based off the CES data. As such, full voter turnout in Canada will simulate and predict the 2019 Canadian Federal Election with full voter turnout. Code and data supporting this is available at: https://github.com/mtsany/Final-PS-Option-B. For legal reasons, the cleaned GSS data cannot be uploaded.

Data

The population of this report is the Canadian population. The frame are individuals who currently reside in Canada and participated in both surveys. The sample are individuals who currently reside in Canada and participated both surveys prior to cleaning.

There are two datasets used: survey (CES) and census (GSS). The survey data was received from an online survey in 2019 that was led by Laura Stephenson, Allison Harell, Daniel Rubenson, and Peter Loewen. There were 37,822 observations in the dataset. The CES team collected the data by using a "two-wave panel with a mdoified rolling cross-section during the campaign period and a post-election recontact wave" (Stephenson et al. 2019). The purpose of the CES is to However, after cleaning the data, the number of observations decreased to 16,518.

The post-stratified data is the census data by GSS is provided by Statistics Canada. The GSS data includes variables relating to family, such as educational attainment and ethnic diversity. The GSS survey surveyed "approximately 43,000 people, 15 years of age and older, living in the 10 provinces" (Statistics Canada, 2017). Statistics Canada randomly chose a sample of households and "uses a frame that combines landline and cellular telephone numbers from the Census and various administrative sources with Statistics Canada's dwelling frame" (Statistics Canada, 2017). So, Statistics Canada collected the data through a telephone interview to one random eligible person in a household that resides in Canada.

According to Stephenson et al. (2020), the dataset consisted of approximately 50% men and 50% women. The dataset had 28% of its respondents aged 18 - 34, 33% aged 35 - 54, and 39% aged 55 and higher. The CES team also ensured that "provincial quotas were split evenly" in the survey.

The chosen predictor variables were: sex, income, province, age, and education level. As stated previously, age and education are particularly important in voter intentions. Other social chracteristics, such as income, province, and sex were chosen because these are key features that affect an individual's voter intentions.

Model

A multilevel logistic regression is used to analyze the probability of the Liberal Party to win in the 2019 Canadian Federal Election. Certain survey response variables, such as age, province, education level, income, and sex is chosen to represent the Canadian population. In this analysis, post-stratification using census data is used to predict the outcome of the election if every individual had voted. The model is appropriate because it allows the estimation of the probability that the Liberal Party will be voted by an individual in Canada. The model will be based off survey data given by CES.

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age,i} + \beta_2 x_{prov,i} + \beta_3 x_{educ,i} + \beta_4 x_{inc,i} + \beta_5 x_{sex,i}$$

Age from the survey was transferred to become numerical and then categorized into age groups. The age brackets were divided by intervals of 20, with the exception of age group 1. This age group included individuals that were 19 to 40 years old. The census data was coded in a similar way for post-stratification

Province from the survey data was coded to be categorized as the five regions in Canada. For example, New Brunswick, Newfoundland and Labrador, Nova Scotia, and Prince Edward Island was coded as group 1. In this case, the group is the Atlantic region of Canada. The data was cleaned such that Nunavut, Yukon, and Northwest Territories were not included in the model. This is due to the low observation count of these individuals in the survey. There were 26, 38, and 25 observations, respectively, in the survey dataset. Each of the omitted provinces represented 0.1% of the dataset (Stephenson et al. 2019). So, omitting these observations will only lead to a very slight decrease that in statistical power in the report.

Education was divided into three categories: lower, middle, and higher education. Lower education refers to no schooling up to some secondary/ high school. Middle-level education refers to completion of secondary/ high school up to some university. Higher education refers to Bachelor's degree up to Professional degree or doctorate. This is an important distinction to make because education level is an integral part of the voting decision of an individual.

Income was categorized into income brackets. It is important to note that the survey data and the census data did not have matching income data. The income brackets for the survey data had intervals of $30,000. Meanwhile, the interval for the census data was $25,000. To remedy this, the income categories were divided into lower, middle, and upper. Lower was categorized by income of less than $100,000. Middle were observations that were between $60,001 to $150,000 in the survey data. However, in the census data, this was coded as $75,000 to $124,999. Observations that were greater than $150,001 in the survey data and greater than $125,000 in the census data was categorized as upper. There is no perfect solution for this problem. This will create a bias in the estimate for the groups in this variable.

Gender was quantified as a binary variable in the survey data. Sex was quantified as a binary variable in the census data. There were no gender variable in the census data. "imputation using the gender reported by an individual in the sample to predict their potential responses for sex" (Kennedy et al. 2020). Statistically, this will have no issue in the modelling and proceeding analysis.

Table 1: GLM model coefficients

Term	Estimate	Std. Error	z-value	p-value
(Intercept)	0.587	0.096	6.087	1.15e-09
age_bracket	-0.045	0.020	-2.317	0.0205
province_q	-0.400	0.016	-25.226	$< 2e\hat{}$-16
education_q	0.471	0.030	15.711	$< 2e\hat{}$-16
income_bracket	-0.106	0.020	-5.345	9.05e-08
sex	-0.304	0.033	-9.304	$< 2e\hat{}$-16

Using a generalized linear model (GLM), the voter intentions were modeled against the predictor variables mentioned previously. All predictor variables were statistically significant at the 5% level. The model had a Fisher Scoring iteration of 4, which implies 4 iterations were required until convergence. The AIC value of the model is 21854,

According to Akaike's Information Criteria (AIC), there is no better fit by removing predictor variables. The current model is the best fit. However, according to Bayesian Information Criterion (BIC), the best fit removes one predictor variables: age bracket. So, our model should be based off of the remaining four predictor variables: province, education, income bracket, and gender.

In backward elimination using AIC, it starts with all of the potential predictor variables and omits the variable with the largest p-value every time. This p-value was carried out by a partial F-test. Furthermore, this creates a smaller information criterion. In backward elimination using BIC, it penalizes the model complexity more heavily than AIC. This makes the model much more simpler than backwards AIC. It can be seen that the model using AIC does not omit any predictor variables, while BIC omits age brackets (Sue-Chee, 2020).

The new model is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{prov,i} + \beta_2 x_{educ,i} + \beta_3 x_{inc,i} + \beta_4 x_{sex,i}$$

Results

From the post-stratification calculation, the $\hat{y}^{PS}$ calculation is: $\hat{y}^{PS} = \frac{\sum N_j \hat{y}_j}{\sum N_j} = 0.471$

Table 2: Voter intention by sex characteristic

Sex	Estimate
0	0.506
1	0.436

Table 2 depicts voter intention by sex. It is seen that females are more likely to vote for the Liberal Party at 0.506, while males are less likely to vote at 0.436.

Table 3 depicts voter intention using income and education levels. As a household's income increases, the household is more likely to vote for the Conservative Party over the Liberal Party. Furthermore, it is seen that a higher education correlates to a higher probability of voting for the Liberal Party.

Table 4 depicts voter intention by Canadian province. The voter intentions in each region varies drastically. The estimate for the regions range from 0.656 to 0.291.

Table 3: Voter intention by income and education levels

	Group	Estimate
Income		
Lower	1	0.494
Middle	2	0.471
Upper	3	0.448
Education		
Low-level	1	0.364
Middle-level	2	0.470
Higher-level	3	0.579

Table 4: Voter intention by Canadian provinces

Province	Estimate
Atlantic Region	0.656
Central Canada	0.565
Ontario	0.469
Prairie Region	0.376
West Coast	0.291

Discussion

The preceding analysis demonstrates that social demographics such as education and income have a correlation with the probability of voting for the Liberal Party. Voter intention for households differ drastically with different probabilities of voting for each region The most important points found in the preceding analysis were:

1) The Liberal Party would have still lost in the 2019 Canadian Federal Election if voter turnout was 100% with 47.1% of total ovters.

2) Higher education levels increase probability of a household voting for the Liberal Party.

Given that this report simulates the 2019 Canadian Federal Election with 100% voter turnout, it can be concluded that there will be no change in the election outcome.

Weaknesses and Next Steps

During the cleaning process, a large sum of observations were omitted from both the survey data and census data. In the survey data, observations reduced from 37,822 to 16,518. In the census data, observations reduced from 20,602 to 18,687. Therefore, the large decrease in observations demonstrates that there is a large loss in statistical power when proceeding with the analysis. The loss in statistical power will decrease the probability of detecting an effect.

For future reference, a survey and census data with a greater observation will be better for modelling and analysis, as a greater sample size imply greater statistical power and accuracy. However, a larger census data is hard to accumulate because it will take a much longer time and greater monetary investment.

In both datasets, it is seen that the income brackets in the survey data and census data do not match up. The intervals in the survey data had intervals of $30,000, while the intervals of the census data was $25,000. This implies that there will be bias because the multilevel logistic regression model based off the survey data

will have bias because it will not reflect the true voting intention of households that are categorized under the incorrect income bracket.

For a future report, a census data that has a similar income bracket should be used. However, this variable could have also been omitted from the model and analysis.

Although only five predictor variables were picked initially, more social characteristics could have been chosen for the model. Furthermore, predictor variables, such as education may not have been cleaned properly for both survey and census data because the cells may have been misclassified. Categories in education, such as, "No schooling" were categorized into the same cell as households with "Some secondary/ high school". It had to be assumed that households in this cell will have similar voting intentions.

Gender and sex were imputed as a binary variable in both the survey and census data. Males were coded as "1" and females were coded as "0". In the post-stratification, gender and sex were considered and assumed to be interchangeable. This is an ethical issue because sex refers to a "set of biological attributes in humans and animals" (Kennedy et al. 2020). However, gender refers to "the socially constructed roles, behaviours, expressions and identities" (Kennedy et al. 2020). According to Kennedy et al. (2020), it raises ethical concerns, despite the imputation not raising any statistical concerns, that people have the right to self-identify, which is ignored in this report. In the future, other measures to post-stratify sex and gender will be considered.

The predictive power of this model seems to complement the actual result of the 2019 Canadian Federal Election. However, the model itself should be re-applied to past elections to understand and assess the actual predictive power of this model. If the model can be validated by previous elections, then the model has great predictive power. Other modifications to the model can include including or omitting certain predictor variables, which can improve the accuracy of the model.

References

Blais, A. (2006). WHAT AFFECTS VOTER TURNOUT? Annual Review of Political Science, 9(1), 111-125. https://doi.org/10.1146/annurev.polisci.9.070204.105121

Kennedy, Lauren, et al. "Using Sex and Gender in Survey Adjustment." ArXiv.org, 30 Sept. 2020, arxiv.org/abs/2009.14401.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rohan, Alexander. 2020 "GSS_Cleaning". STA304

Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020. Canadian Election Study - 2019 Online Survey Codebook. doi.org/10.7910/DVN/DUS88V

Statistics Canada. (2020, April 30). General Social Survey – Family (GSS). Retrieved December 20, 2020, from https://www.statcan.gc.ca/eng/survey/household/4501

Statistics Canada. (2017, March 31). Factors associated with voting. Retrieved December 20, 2020, from https://www150.statcan.gc.ca/n1/pub/75-001-x/2012001/article/11629-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

Sue-Chee, S. (2020). *Week11_302_a20.pdf*. University of Toronto

Xie, Y, Dervieux, C, Riederer, E. (2020). *R Markdown Cookbook*. R package version 1.3.1.

Zhu, H. (2020). *Create Awesome LaTeX Table with knitr::kable and kableExtra*. R package version 1.3.1.