The Importance of Turnout of the Election

FanxiZhou

2020/12/9

Abstract

We investigate the results of the 2019 Canadian federal election by analysing the dataset of canadian election study 2019-phone survey and general social survey. By analyzing, results shows that the outcome of the election would be similar as the outcome of the 2019 Canadian federal election even if the turnout of the election is 100%. These results are important because the turnout of each election keeps changing and it is difficult to predict that how would the result change if turnout increase. Our analysis provides an outcome for people as a reference.

Contents

Keywords

Canadian elections, Turnout, MRP techinque, Logistic Regression.

Introduction

The 2019 Canadian federal election (43rd Canadian general election) was held on October 21, 2019. This election elected the Canada's 43rd Parliament. The Liberal party, which is led by Justin Trudeau, gained 157 seats in the House of Commons(Krystyne Manzer 2019). Although the Liberal party won the election, they only gained the 33.0% of 17.9 million national votes so they losed the majority and formed a minority government(Krystyne Manzer 2019). The Conservatives party captured 121 seats after the Libreal party. The Québécois party and the NDP party gained 32 seats and 24 seats respectively. The turnout of the 2019 Canadian federal election is nearly 66%, not as high as in 2015 Canadian federal election(CBC 2019).

We analyzed the vote results of Canadian who voted in the 2019 Canadian federal election through several factors by building a binary logistic regression model based on the dataset form canadian election study 2019-phone survey(CES dataset). The goal of our analysis is to find whether it would affect the vote results if all Canadians vote in 2019 Canadian federal election by performing Multilevel regression and post-stratification (MRP) analysis on the CES dataset and General Social Survey (GSS dataset).

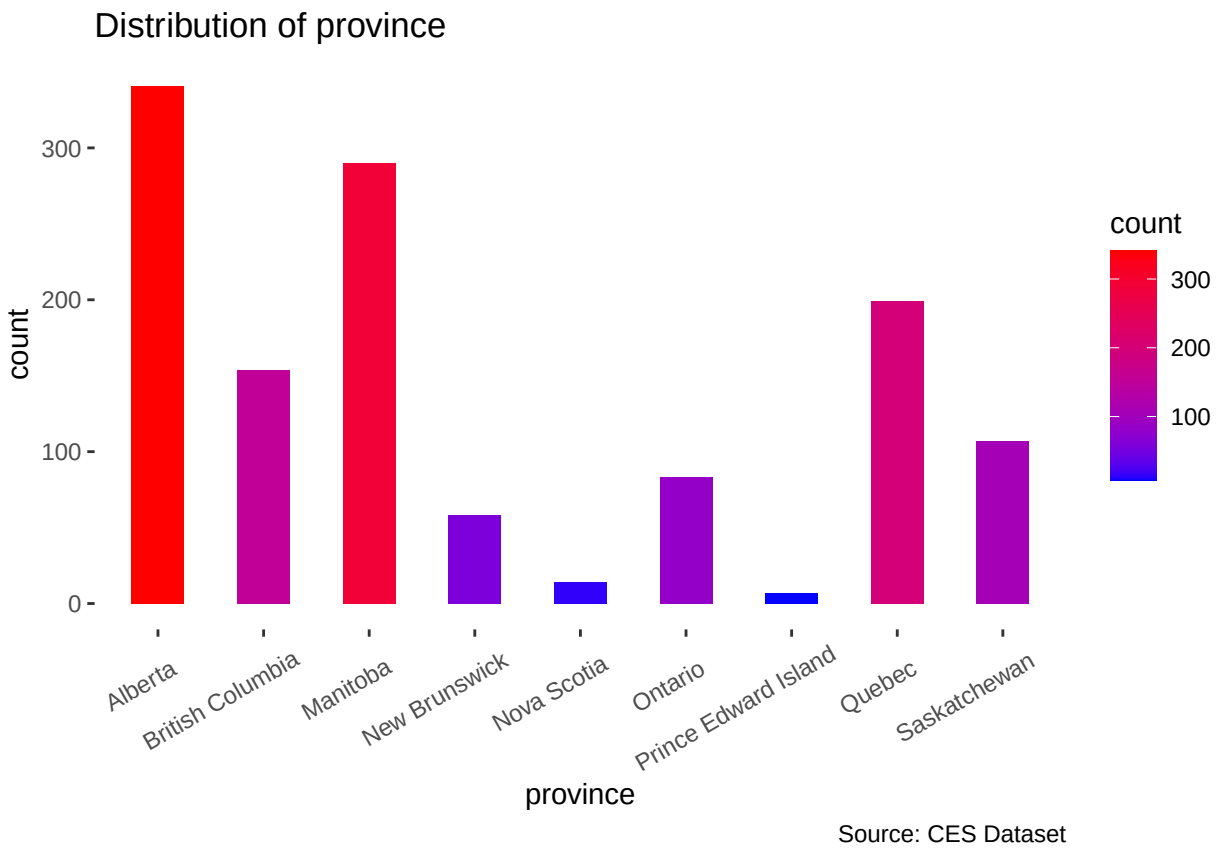
In our study, we found that the total number of children, sex, marital status significantly effect people's vote opinion. As a Canadian, people would more likely to vote Liberal with less children in their life. Male have lower votes on Libral party compared to female. Furthermore, people who live in common-law or married would more likely to support the Conservatives party. By observinf the result based on the model from the CES dataset, our result is similar to the result of the 2019 Canadian federal election. It means that eventhough all of canadians votes in the election, the result would not change a lot.

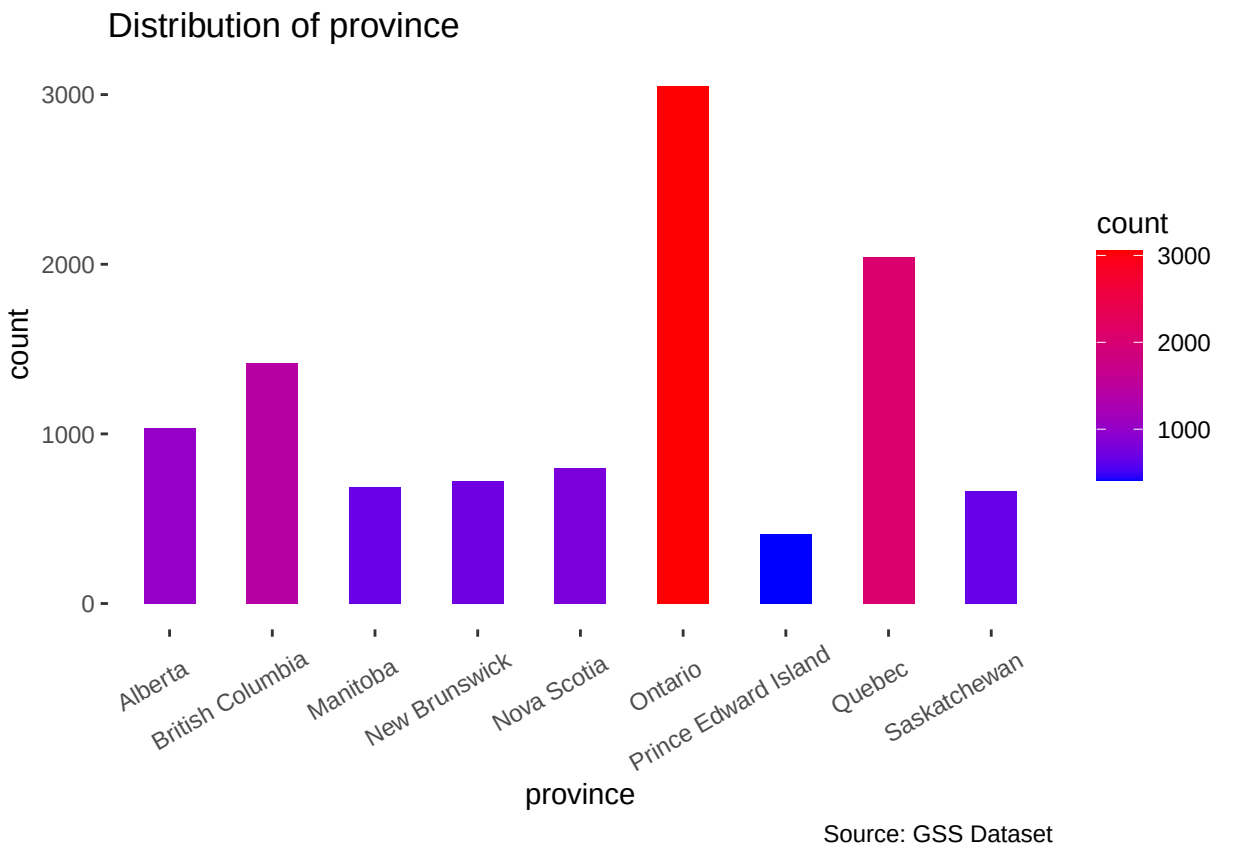
Data discussion

The GSS dataset is based on the General Social Survey (GSS), which aimed to gather information about the changes in living conditions of Canadians and to provide information on specific policy issues. The original data contains 20602 observations and 81 variables. In this dataset, the target population includes the person who is 15 years old and older,living in 10 provinces in Canada. we tried to analysis the distribution of age, sex, education and so on. The frame of the survey was to combine the telephone numbers (landline and cellular) with Statistics Canada's Address Register. During the survey, 91.8% of the telephone numbers reached the household. The overall response rate of the survey is 52.4%(cite).

The CES dataset is based on the Canadian Election Study 2019- phone survey. The purpose of this survey is to represent the adult population of Canada in the 43rd Canadian general election which is the Canadian citizens over 18 in 10 different provinces. The original dataset contains 4021 observations and 278 variables. In this dataset, the target population is the Canadian citizens over 18 in 10 different provinces. The frame of the survey is constructed by 34% landline phone number and 66% wireless phone numbers. The estimated number of eligibles is 72241, dividing 4021 people who complete this survey gives the final response rate of 5.6%.

In our study, we made two subsets by selecting some common variables in CES dataset and gss dataset that may affect the people's vote decision. There are total 9 common variables in CES dataset and GSS dataset: age,sex,education,has_religion,religion_importance,income_family,employment,born_canada and province. The age is numeric variable and other 9 variables are categorical. The CES also has a unique variable vote_Libreal which is a binary varible contrains 1 and 0. Also, the unqiue variable of GSS dataset is count which represent the count of each cell. It was used for doing post-stratification process(see model section).


Distribution of province
count
300
200
100
0
Alberta
British Columbia
Manitoba
New Brunswick
Nova Scotia
Ontario
Prince Edward Island
Quebec
Saskatchewan
province
count
300
200
100
Source: CES Dataset



According the figures above, most of respondents in CES data comes from Alberta, Manitoba and Quebec. Only few respondents in Nova Scotia and Prince Edward Island. However, the distribution is different in GSS dataset. Ontario province has most of respondents in GSS dataset. Quebec has the second highest populaton of respondents.

Model

The purpose of our study is to determine whether the turnout of the election could affect the actual vote outcome. We used multilevel regression and post-stratification technique for this analysis. In the following sub-sections I will describe the model specifics and the calculation for the post-stratification process.

Model Specifics

Before making a model, we need to do the variable selection so that we can choose a subset of the predictors which make the model fit better. In general, when we add more predictors to the model, the bias of the predictions gets the smaller but the variance of the estimated coefficients gets bigger. In other words, it is necessary for us to select peoper predictor for our final model. In the beginning, we used the all of 9 variables as the predictors of our full model: age,sex,education,has_religion,religion_importance,income_family,employment,born_canada and province. The vote_Liberal is the response variable of our model. It is a binary variable so we dicided to build a binary logistic regression model. Then, we used the backward eliminations for variable selection, which starts with all potential predictors in the model, then remove the predictor with the largest p-value each time to give a smaller information criterion. We used Akaike information criterion (AIC) for this part. AIC uses the number of independent variables that are used to make a model and the maximum likelihood estimate of the model to get a value. Here, the maximum likelihood estimate tells us how well the model reproduces the data. After finishing this process, we had our final model. The response variable was vote_Liberal, and the 5

predictors of model were age, sex, has_religion, born_canada and province. We will use this final model based on CES dataset to predict the probability of voting for Liberal party in GSS dataset.

Then, we used R(cite) to run the binary logistic regression models to model the proportion of voters who will vote for Liberal party to assume all of people vote in the election, that is, turnout is 100%. The equation below is the binary logistic regression model:

$$\pi_i = Pr(Y_i = 1|X_i = x_i) = \frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)}$$

or

$$logit(\pi_i) = log(\frac{\pi_i}{1 - \pi_i}) = \beta_0 + \beta_1 x_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_q x_{iq}$$

We assume that Y_i is a binary response variable for i = 1,... n and takes on value 0 or 1 with P(Y_i = 1) = pi_i. Suppose X is a set of explanatory variables, x_i is the observed value of the explanatory variables for observation i = 1,... q. From the above formula, we can also get:

$$\frac{\pi}{1 - \pi} = e^{\beta_0} e^{\beta_1 x_1} ... e^{\beta_q x_q}$$

Then the β_0 is the baseline odds and β_1 can be interpreted as holding predictors constant, a one-unit increase in x_1 increases the probability of voting for Liberal party by a factor of e^{β_1}.

Post-Stratification

Multilevel regression and post-stratification (MRP) combines two statistical techniques to determine the relationship between the response variable of our interest and predictors we chose. Unlike the normal multilevel regression analysis, we add a post-stratification process base on the previous multilevel regression analysis. We used the sample data to train a regression model and then we would use this trained model to predict the outcome in the population dataset which would be a large population. The MRP requires the data to be demographic. In our study, we chose 5 predictors which are mentioned in the previous section as the key demographic features of the sample. However, MRP also has some limitations. As we mentioned before, the MRP requires the data to be demographic. Also, if the sample data is not sufficient enough or the demographic predictors are not enough, the outcome would be biased and can even be failed. In our sample data GSS Dataset, we could find some key demographic features as our predictors, so we chose to use MRP for our analysis.

In the post-stratification process, in order to estimate the proportion of voters who will vote for Liberal party. We performed a post-stratification analysis on the GSS dataset(citation). We created many cells based on different age, sex, has_religion, born_canada and province. Performing the model described in the above section, we estimated the proportion of voters in each cell. Then, we calculate the proportion of voters estimate for each cell by using the respective population size of that cell and sum those values and divide that by the whole population.

```
##    win_prob
## 1 0.2611661
```

Results

Table 1: Summary of Model Results

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-14.9719	384.3982	-0.0389	0.9689
age	0.0146	0.0040	3.6482	0.0003

	Estimate	Std. Error	z value	Pr(>|z|)
sexMale	-0.2599	0.1248	-2.0817	0.0374
has_religionYes	13.7395	384.3982	0.0357	0.9715
born_canadaborn outside canada	0.7168	0.1574	4.5534	0.0000
provinceBritish Columbia	0.5031	0.1999	2.5170	0.0118
provinceManitoba	-0.6167	0.1785	-3.4555	0.0005
provinceNew Brunswick	-0.3364	0.3126	-1.0762	0.2819
provinceNova Scotia	0.0568	0.5619	0.1011	0.9195
provinceOntario	-0.7241	0.2883	-2.5119	0.0120
provincePrince Edward Island	-1.1738	1.0936	-1.0734	0.2831
provinceQuebec	-0.2993	0.1939	-1.5435	0.1227
provinceSaskatchewan	-0.3817	0.2428	-1.5723	0.1159

By doing the binary logistic model we get the model results. From the model results above, we can see that age,sex, born_canada and province are significant predictors. However, the p-value for has_religionYES is very big which means it may not significant. Also, some of provinces has higher p-value. The probablity of vote for Liberal is 0.26 which means around 26% people would vote for Liberal party if the all of Canadians vote in the election. Which is similar as the actual vote outcome. The liberal party is still a minorty government.

Discussion

weakness

nextstep

Appendix

References