Will the turnout of the election affect the election result?*

FanxiZhou

2020/12/20

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.

Keywords: Canadian election, Turnout, MRP technique, Logistic Regression

Contents

*Code and data are available at: https://github.com/zhoufanx/STA304-Final-Paper

1 Introduction

Canada's election system comes from the United Kingdom. It is a constitutional monarchy, composed of the Queen of Canada, and is formally represented by the Governor (or Provincial Lieutenant Governor), the Senate and the House of Commons. The Senate has 105 seats and its members are appointed by the Governor on the recommendation of the Prime Minister. The House of Commons has 338 seats, held by members elected by citizens who voted in the general election or by-election. The number of electoral districts is determined according to the rules stipulated in the 1867 Constitution ("the "representation formula"). There are 338 electoral districts, and each electoral district has a corresponding seat in the House of Commons. In each electoral district, the candidate with the most votes will win a seat in the House of Commons and represent the electoral district as its member of parliament.(("Home" 2020))

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((Manzer 2019)). Although the Liberal party won the election, they only gained the 33.0% of 17.9 million national votes so they lost the majority and formed a minority government((Manzer 2019)). The Conservatives party captured 121 seats after the Liberal party. The Québécois party and the NDP party gained 32 seats and 24 seats respectively. Voter turnout is the percentage of eligible voters who votes in an election. The turnout of the 2019 Canadian federal election is nearly 66%, not as high as in 2015 Canadian federal election((News 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)((Stephenson and Loewen 2019)). 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)((Gagné and Keown 2014)).

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 Liberal party compared to female. Furthermore, people who live in common-law or married would more likely to support the Conservatives party. By observation, 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 even though all of Canadians votes in the election, the result would not change a lot.

2 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%((Gagné and Keown 2014)).

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%((Stephenson and Loewen 2019)).

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).

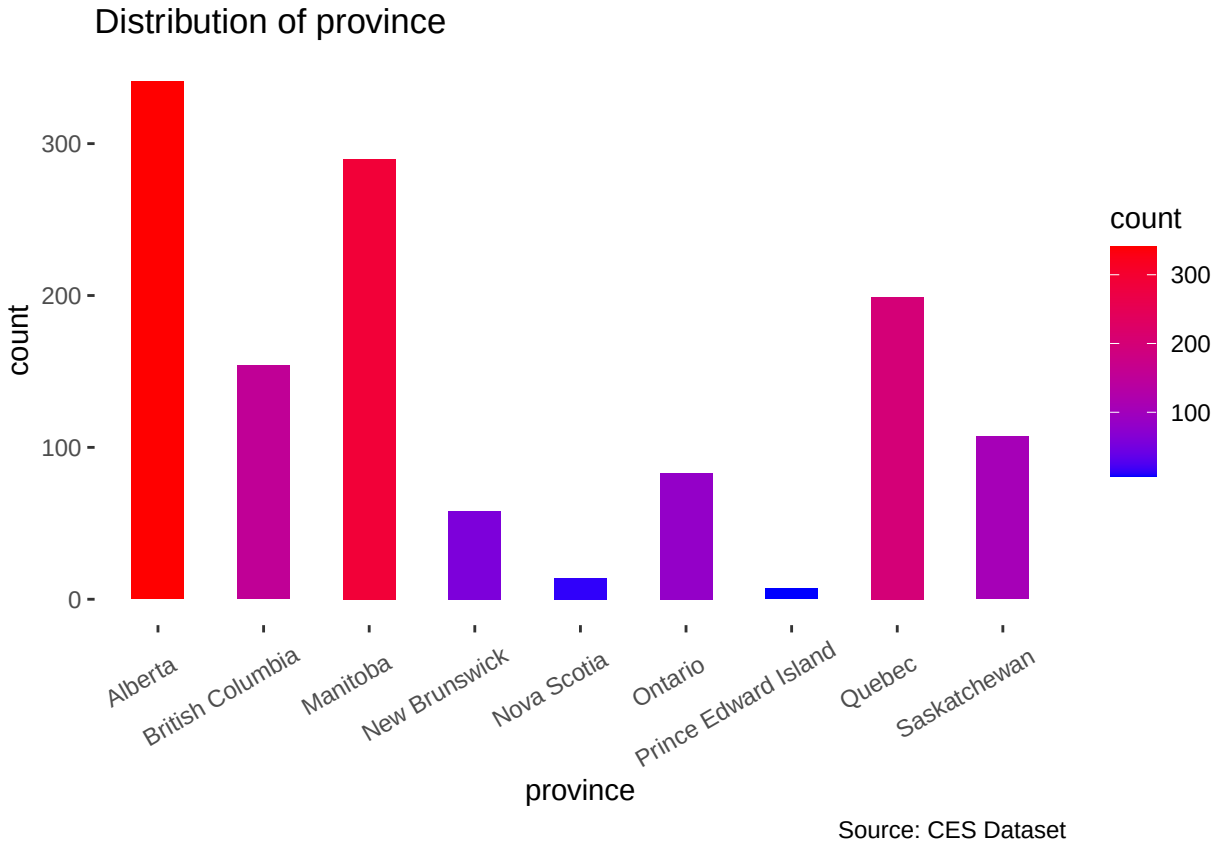

Figure 1: Distribution of province in CES

According the figure1 and figure2 , 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.

From the figure3 and figure4, the age distribution in two datasets is little bit differnet. The young people owns larger proportion in the GSS dataset. However, 50-60 age people occupies most of proportion of people who votes in the election in both of these two datasets

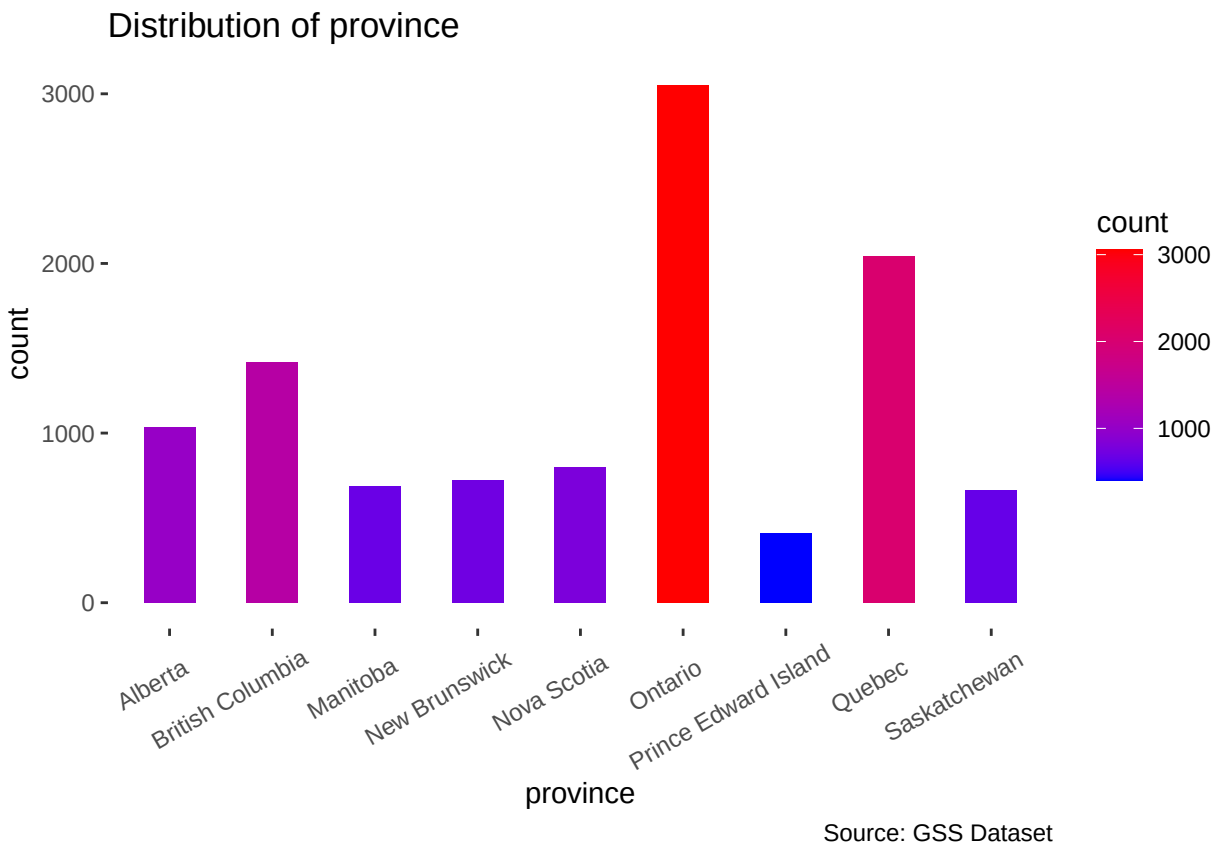

Figure 2: Distribution of province in GSS

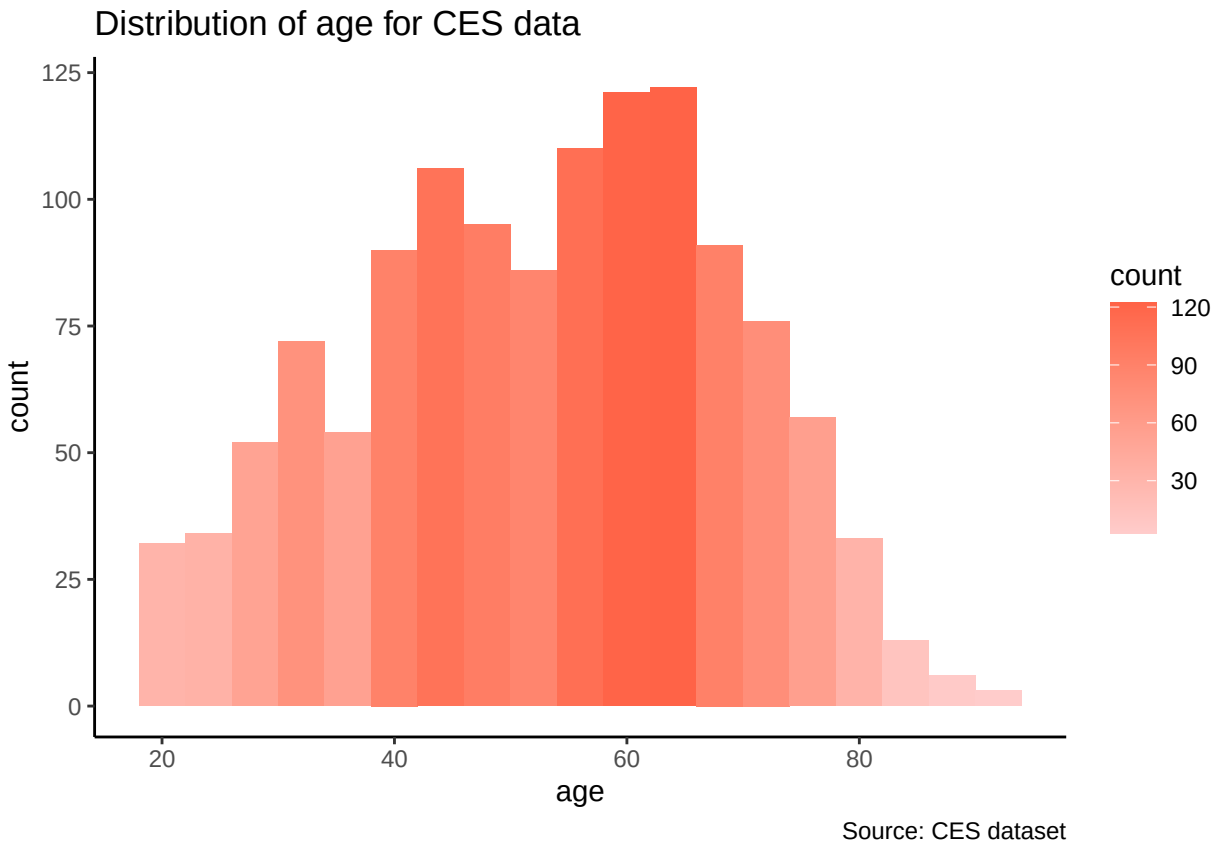

Figure 3: Distribution of age in CES

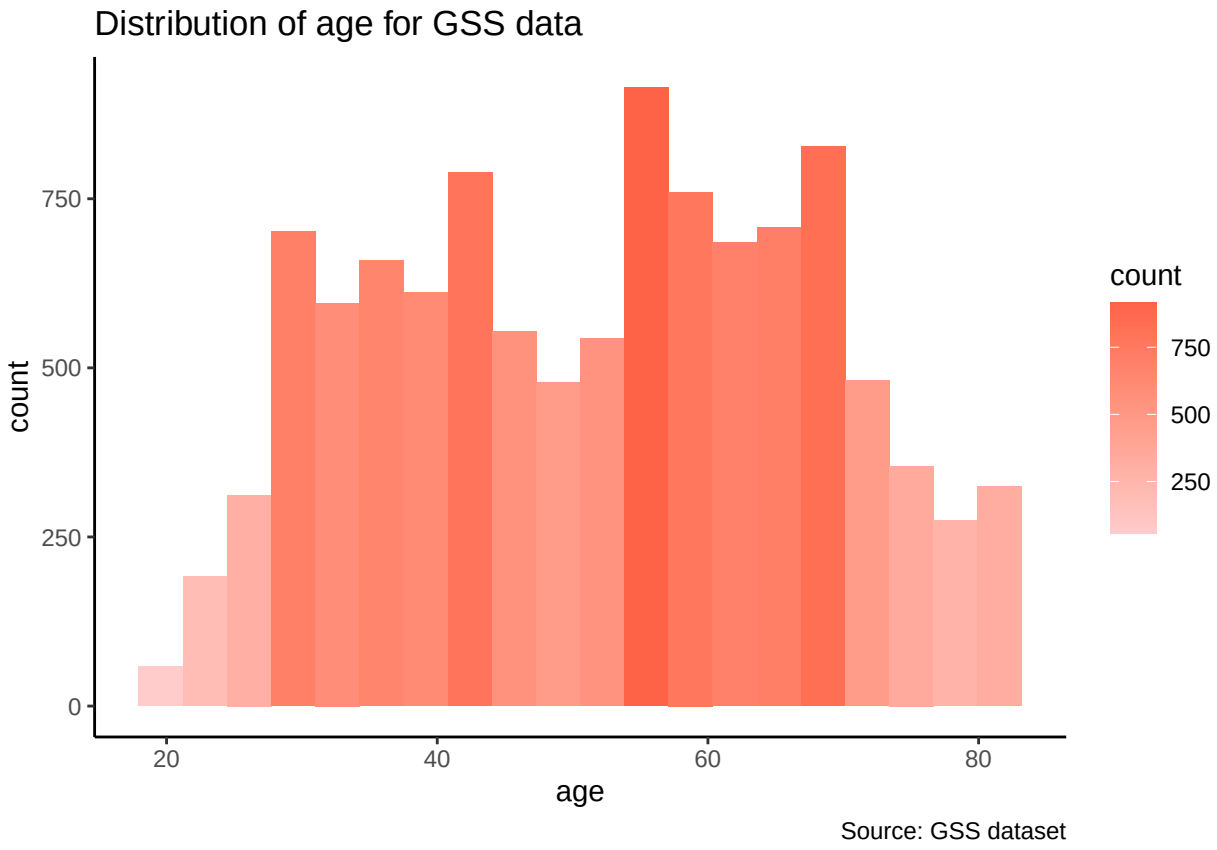

Figure 4: Distribution of age in GSS

3 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.

3.1 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 proper 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}.

3.2 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.

4 Results

Table 1: Summary of prediction of vote outcome

	Probability
Vote for Libreal Party	0.261

Table 2: 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
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 table2 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 probability 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 minority government.

From the figure5, we could see that people in the elder age group are more likely to vote for the Liberal party, young people are more likely to choose to not vote for liberal party.

The figure6 is the distribution of sex for people's vote choice. The red parts represents the proportion of people who votes for Liberal party. We could clearly see that the proportion of male and female is almost the same. The number of males in this data is more than that of females.

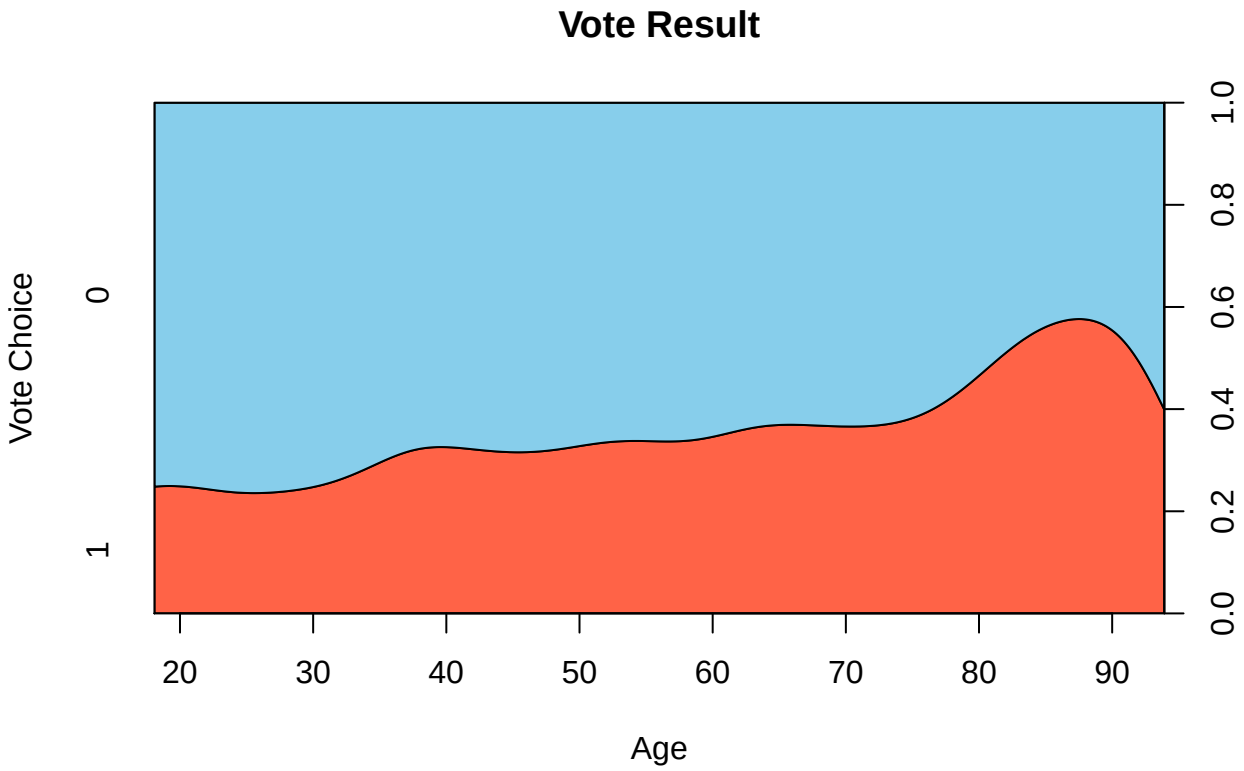

Figure 5: Vote results for different age

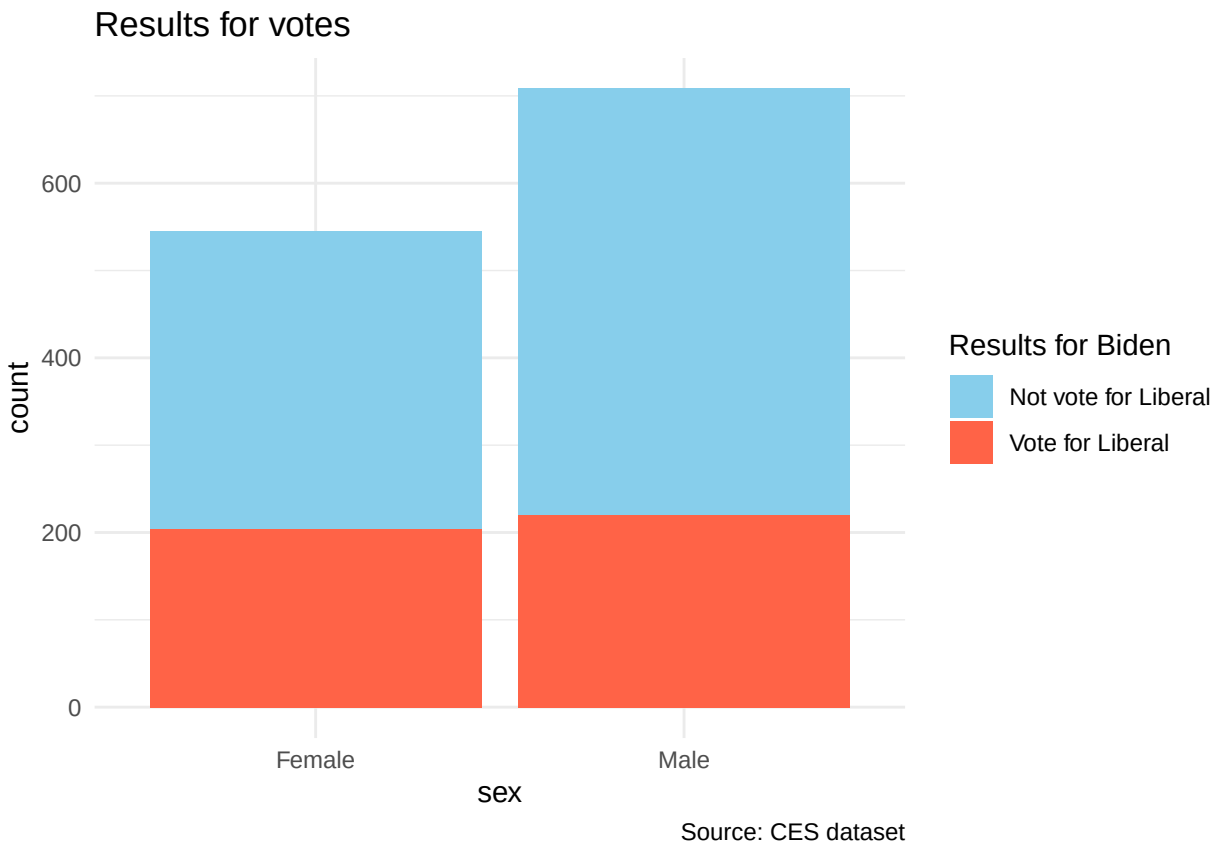

Figure 6: Vote results for different gender

5 Discussion

In this model, the p-value helps us to test the null hypothesis so that we can indicate whether the factors have a correlation with our predictor in the whole population. The null hypothesis for our model is that there is no correlation with our response variable. If the p-value is smaller than 0.05, it rejects the null hypothesis so there may be a correlation between that factor and our response variable. The smaller the p-value, the stronger evidence for rejecting the null hypothesis. However, if the p-value is greater than 0.05, it supports the null hypothesis, which means that there may not be a correlation between the factor and our response variable. From the model results above(), we can find that age, sex, born_canada, province are significant predictors since their p-value is less than 0.05. However, has_religion is not a significant predictor, the p-value is 0.97015 which is higher than 0.05.

The Odds are defined as the ratio of the probability of success and the probability of failure.((Bruin 2011))The probability can be reconstructed as

$$probability = \frac{odds}{1 + odds}$$

In other words, we could calculate its probability by using this equation. Each unit change in age in the model increases the log odds of voting for the Liberal party by 0.0146. Being sex male, versus being female, changes the log odds of voting for the Liberal party by -0.675. In addition, the p-value for born_canada is very small which means there is a strong evidence based on our model that reject our null hypothesis. People born outside Canada, compare to born in Canada, change the log odds of voting for the Liberal party by 0.7168. Province British Columbia, Ontario and Manitoba have significant value. It means the odds of these province are also significant. It means that people in British Columbia, Ontario and Manitoba, versus people in Alberta, change the odds of voting for the Liberal party by 0.5031,-0.7241 and -0.6167 respectively.

Table 3: Summary of odds ratio and 95%CI

	OR	2.5 %	97.5 %
(Intercept)	0.0000	NA	6.695466e+10
age	1.0147	1.0068	1.022800e+00
sexMale	0.7711	0.6037	9.850000e-01
has_religionYes	926820.7399	0.0000	NA
born_canadaborn outside canada	2.0478	1.5037	2.788800e+00
provinceBritish Columbia	1.6539	1.1184	2.450600e+00
provinceManitoba	0.5397	0.3793	7.639000e-01
provinceNew Brunswick	0.7143	0.3797	1.301200e+00
provinceNova Scotia	1.0584	0.3351	3.172600e+00
provinceOntario	0.4848	0.2695	8.387000e-01
provincePrince Edward Island	0.3092	0.0161	1.875300e+00
provinceQuebec	0.7413	0.5053	1.081400e+00
provinceSaskatchewan	0.6827	0.4203	1.091200e+00

The table3 above shows the odds ratio and 95% confidence interval for each predictor. Odds ratios are used to compare the relative odds with the occurrence of the outcome. If the odds ratio is greater than 1, it means that exposure associated with higher odds of outcome. However, if the odds ration less than 1, it means the exposure associated with lower odds of outcome. Also, if odds ratio equal 1 it means exposure would not affect the odds of outcome. The 95% confidence interval means that 95% probability that the population would fall within the interval. For instance, the odds ratio of age is greater than 1, and it means that the age would associated with the higher odds of outcome. The odds ratio here indicates that for every 1 unit increase in the age, the likelihood that voting for Liberal party is present increases by approximately 1.0147 times. Also, we are 95% confident that the proportion of people who votes for Liberal party contains the value of the odds ratio for the population.

To sum up, We post-stratify based on age, sex, born_canada, has_religion and province. The expected

probability of voting Liberal party is about 26.1%, the proportion for actual election is around 33%. The actual proportion is higher, and it means that if turnout of the election becomes 100%, the proportion of people who would be more likely to vote for Liberal party would decrease. The Liberal government would still be a minority government. The age of people, gender of people, born in Canada or not, and people in different provinces would affect the probability that voting for the Liberal party in the election. We assumed that all of people in the survey would vote in the actual election and get these results. These results are believable since we selected proper predictors by using AIC which provided a proper fitted model for us to doing analysis. However, there are also some flaws in this analysis. I would discuss about them in the below sub section.

5.1 weakness

- According to the General Social Survey (GSS), 91.8% of the telephone numbers reached the household during the survey, but the survey's response rate is only 52.4%. Also the response rate in Canadian Election Study (CES) is also low, which is only 5.6% A low response rate can produce sampling bias if the non-response outcomes are unequal among the participants since the data would be less representative.
- During the data clean data process we remove a lot of missing values(NA), and also we delete some category in a variable like we removed people who refused to response some question in the survey. So this clean process may increase the bias of the results, since even if they refused to response some of questions in the survey, their vote choice still could affect the final results.
- People who completed the survey were mainly distributed in the 30-45 and 50-60 age range, which means older people are more likely to complete the survey. From the result, we concluded that people more likely to vote for Liberal party as their age increases. However, the result may not be accurate since our sample can not represent the whole population.

5.2 nextstep

- More predictors could be added to model, and build the model based on different subsets of selected predictors. Then we could compare different models and choose the one that best fits our data.
- We could use a bigger dataset as our post-stratification data for analysis. A smaller dataset may cause the bias, and it may not represent the actual population which would affect the accuracy of the results.
- We could improve our data clean methods so that we could avoid doing some unnecessary deletions on the original data, which would make our results more believable.
- We also could apply more techniques like chi-squre test, Anova, and adjusted R square to test the goodness of model. Also, we may use different kinds of model like lm, brms and glmer to determine which kinds of model would be better.

6 Appendix

We use R (R Core Team 2020) and packages tidyverse (Wickham et al. 2019), dplyr (Wickham et al. 2020), kableExtra (Zhu 2020) CES dataset (Stephenson and Loewen 2019) and GSS dataset(Gagné and Keown 2014) for this analysis.

References

Bruin, J. 2011. "Newtest: Command to Compute New Test @ONLINE." February 2011. https://stats.idre.ucla.edu/stata/ado/analysis/.

Gagné, Roberts, C., and L.-A. Keown. 2014. "Weighted Estimation and Bootstrap Variance Estimation for Analyzing Survey Data: How to Implement in Selected Software." http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=12-002-%20X20040027032&lang=eng.

"Home." 2020. – *Elections Canada*. https://www.elections.ca/content.aspx?section=res&dir=ces&document=part1&lang=e.

Manzer, Krystyne. 2019. "2019 Canadian Election Results." *RBC Global Asset Management*. https://www.rbcgam.com/en/ca/article/canadian-federal-election-2019/detail.

News, CBC. 2019. "Canadian Election Drew Nearly 66." *CBCnews*. CBC/Radio Canada. https://www.cbc.ca/news/canada/voter-turnout-2019-1.5330207.

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

Stephenson, Allison Harell, Laura B., and Peter John Loewen. 2019. "Data:The 2019 Canadian Election Study – Phone Survey."

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686.

Wickham, Hadley, Romain Fran?ois, Lionel Henry, and Kirill Müller. 2020. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Zhu, Hao. 2020. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.