How could 2019 Canadian federal election be changed if everyone has voted

Code and data supporting this analysis is available at:
https://github.com/Hiraethwly/STA304_Final_Project

Leyi Wang (1006318682)

22/12/2020

Abstract

The purpose of this report is to investigate the outcome of the 2019 Canadian federal election if every eligible voter has participated. This report discovers the relationship between people's votes and their gender, income level, and education through the logistic regression model; then, the model is applied by post-stratification to eliminate bias. Voting and elections are the most basic elements of democracy; this report studies voter turnout for a better understanding of electoral democracy in Canada.

Keywords

- 2019 Canadian Federal Election, voter turnout, votes, Liberal Party, Conservative Party.
- Gender, Income level, Education level.
- Logistic regression model, Post-stratification.

Indroduction

On October 21, 2019, The 2019 Canadian federal election was held to elect members of the House of Commons to the 43rd Canadian Parliament. As everyone knows, Prime Minister Justin Trudeau who is from the Liberal Party won in this election.

Inspired by Jack Bailey who has found out Prime Minister of a minority Labour government would win if everyone had voted in the 2017 General Election by building and analyzing Multilevel regression with post-stratification. (Bailey) Thus, we are quite interested in the results of the Canadian 2019 Federal Election if everyone had voted since only 67.0% eligible voters cast a ballot last year in this election. (Elections Canada)

The purpose of this report is to investigate how the election result would change if everyone has voted in 2019, whether the Liberal Party would still win or not. Since the major competitor of the Liberal Party in 2019 is the Conservative Party, there two parties hold 67.5 vote share totally. Hence,the main idea is to focus on the proportion of voters who support the Liberal Party and the proportion of voters who support the Conservative Party.("Federal Election 2019 Live Results")

In order to accomplish this goal, we fit a logistic regression model with the post-stratification technique for the proportion voting for Liberal Party based on 2019 Canadian Election Study (CES) Data and 2017 General Social Survey: Families Cycle 31 (GSS) Data in the Methodology section. In Result second, the estimated proportion of voters in favour of the Liberal Party is provided; then, inference of the data and result with conclusions, weakness and next steps are presented in the Discussion section.

Methodology

Data

In this report, two data sets will be used to investigate which party would win the election if everyone had voted. Thus, this report uses the 2019 Canadian Election Study (CES) Data as survey data and the 2017 General Social Survey: Families Cycle 31 (GSS) Data as census data.

Introduction of the Survey Data

In order to easily access CES survey data, this report uses R package cesR to obtain an subset of the 2019 online survey called "ces2019_web". (Paul and Alexander)

CES 2019 was composed of a two-wave panel with a modified rolling-cross section during the campaign period from September 13th to October 21st in 2019 and a post-election recontact wave from October 24th to November 11th in 2019. This survey target on respondents who are aged 18 or order and need to be Canadian citizens and permanent residents. There are 37,822 online samples through Qualtrics, with targets stratified by region and balanced on gender and age within each region. (Stephenson et al.)

Cleaning data

Since this report aims to investigate the proportion of voters who support liberal party if everyone has voted and how voting for LP is affected by voters' gender, income level and education level.

This report select variables: cps19_gender, cps19_province, cps19_education, cps19_income_cat, cps19_v_likely, cps19_votechoice, cps19_vote_unlikely, cps19_v_advance, cps19_vote_lean from ces2019_web sample data.

Base on cps19_v_likely,cps19_votechoice, cps19_vote_unlikely, cps19_v_advance, cps19_vote_lean, these variables provide detailed information about which parties that each voters would most likely choose in the election day; then, a varibale called vote_LP is recorded as 1 if people would vote for Liberal Party and recorded 0 if people would vote for conservative party.

Then, cps19_gender was reclassified as Male, Female and Other gender. In addition, cps19_gender and cps19_province were renamed as gender and province for simplification.

For cps19_education, cps19_education records the highest level of education that voters have completed. For simplification, it was renamed as education and reclassified into three education levels. For those people who doesn't complete secondary school, they have below upper-secondary education level. For those tho complete secondary school but doesn't go to college or university, they have upper secondary education level. For those who pursues any education that beyond high school, they have tertiary education level.

For cps19_income_cat, cps19_income_cat records household income of voters and was classified into "low-income","middle income", and "high income".(cite)(show that how we classify)

Please refer to R scrip "1.Cleaning_Survey_data.R" for cleaning survey data.

After Cleaning Data

Response variable: vote_LP(binary variable): if people vote for Liberal party, vote_LP is 1; otherwise, if people vote for Conservative party, vote_LP is 0.

Predictor variables:

1. gender(Categorical variable): Female or Male

2. province(Categorical variable): Newfoundland and Labrador, New Brunswick, Nova Scotia, Prince Edward Island, Quebec, (Manitoba, Saskatchewan, Alberta, and British Columbia.

3. education(Categorical variable): Below Upper-Secondary, Upper Secondary, and Tertiary Education

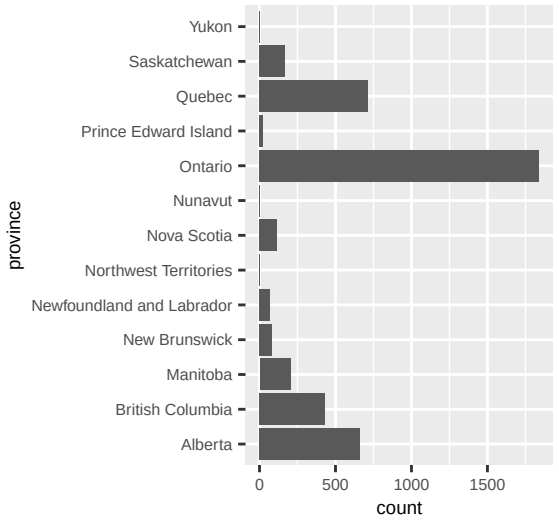
4. income(Categorical variable): Low-income, Middle-income, High-income

Weakness of Survey data

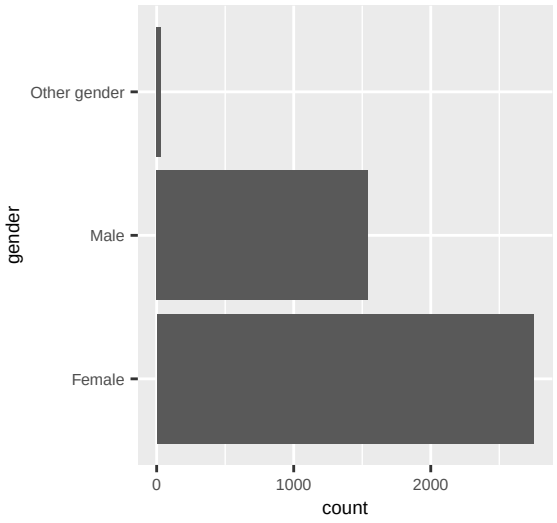
1. There exist many NA in CES data. When the missing value is removed, the accuracy of data is reduced.
2. In the survey, the respondent can choose "Don't know/ Prefer not to answer" options which cause the missing response in this component of the question thus, lacking of responses may cause non-response bias.

Demonstration of Survey Data

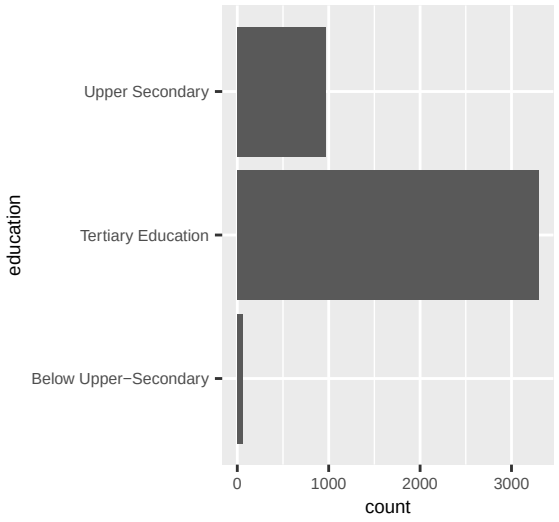
Graph 1: The bar plot of province



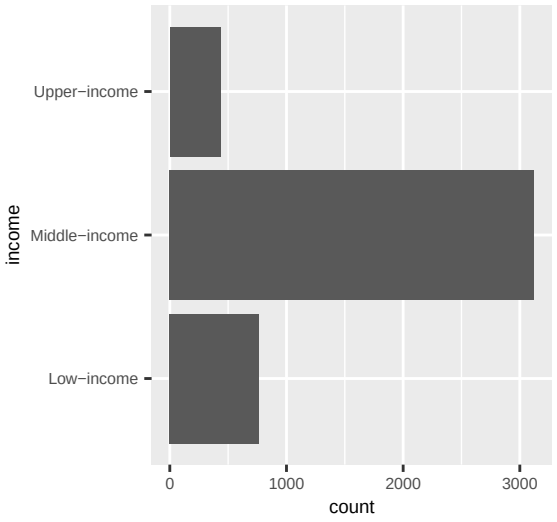
Graph 2: The bar plot of gender



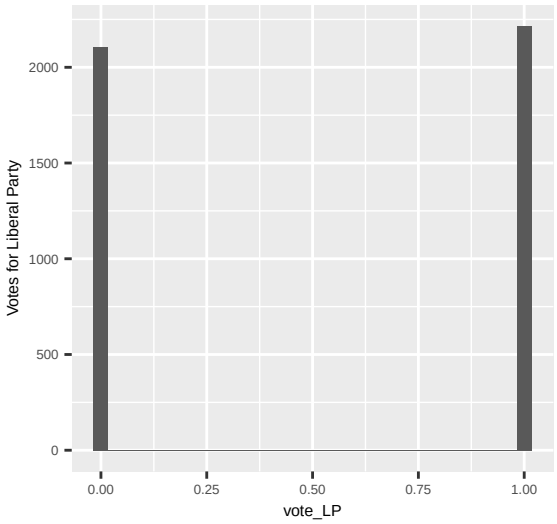
Graph 3: The bar plot of education



Graph 4: The bar plot of income



Graph 5: The bar plot of votes for Liberal Party



Introduction of the Census Data

This report uses 2017 General Social Survey: Families Cycle 31 (GSS) Data as census data.

The GSS data monitors the overall standard of living and well-being in Canadian families by collecting information on conjugal and parental history (chronology of marriages, common-law unions and children), family origins, children's home leaving, fertility intentions, and other socioeconomic characteristics.(Surveys And Statistical Programs - General Social Survey - Family (GSS))

It provides information of social trend in many aspect such as province, income, gender, education level that can affect people's decision to vote.

Please refer to R scrip "2.Cleaning_Census_Data" and "3.Cleaning_Post_Stratification" for cleaning census data.

The target population: All non-institutionalized persons 15 years of age or older, living in the 10 provinces of Canada excluding residents living in Yukon, Northwest Territories and Nunavut

Sampling frame: This survey uses a frame that combines landline and cellular telephone numbers from the Census and various administrative sources with Statistics Canada's dwelling frame. Records on the frame are groups of one or several telephone numbers associated with the same address (or single telephone number in the case a link between a telephone number and an address could not be established). This sampling frame is used to obtain a better coverage of households with a telephone number.(Surveys And Statistical Programs - General Social Survey - Family (GSS))

The frame population: the population that can be covered by sampling frame.

Sample: respondent is randomly selected in each sampled household.

Sampling method: Stratified random sampling is used to collect samples that divides the population into smaller subgroups by homogeneous features called "strata". Therefore, the ten provinces are divided into 27 strata by geographic areas. In each stratum, simple random sampling without replacement is used to generate 2017 GSS dataset.

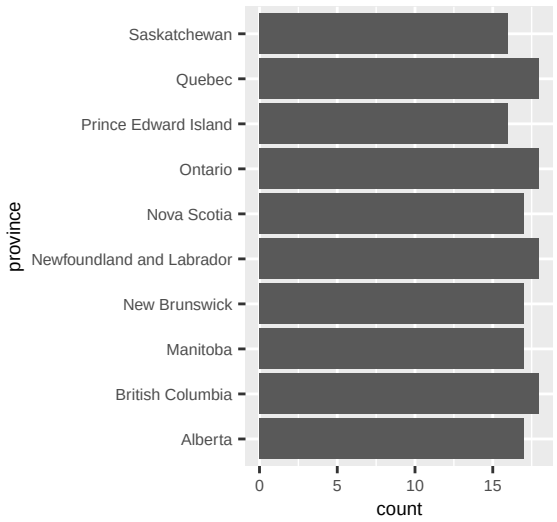
Non-response Problems and solutions: There are at least two times for people to participate in this survey; for example, if the people miss the interview due to a conflict of timing, they are able to make another appointment of an interview call.

Weakness of Census data

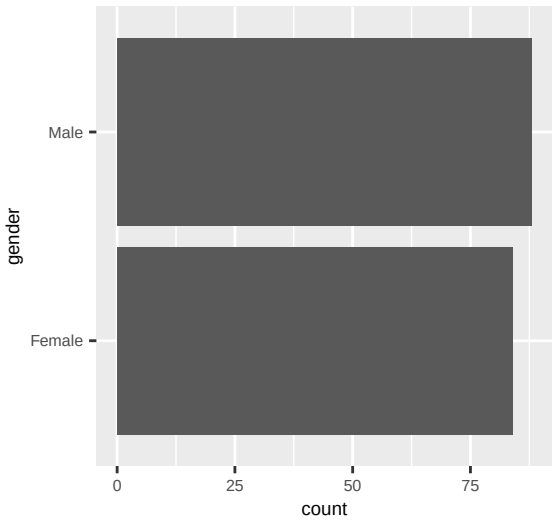
1. There exist many NA in GSS data. When the missing value is removed, the accuracy of data is reduced.
2. The overall response rate is only 52.4%; thus, lacking of responses may cause non-response bias.

Demonstration of Census data

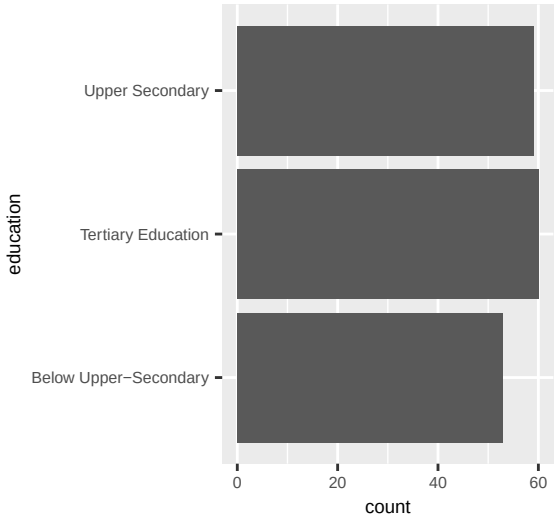
Graph 5: The bar plot of province



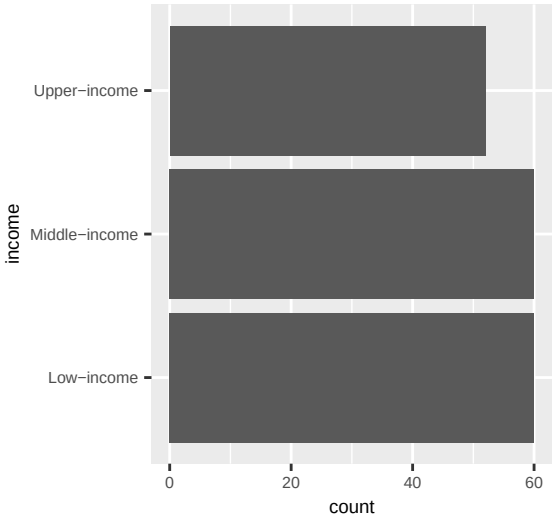
Graph 6: The bar plot of province



Graph 7: The bar plot of education



Graph 8: The bar plot of income



Model

Model selection:

This report uses logistic regression since it suitable to conduct conduct when the dependent variable is dichotomous. Logistic regression can be used to establish a relationship between binary response variable vote_LP and a group of predictor variables such as gender, income, education.

Model identification:

The logistic regression model we are using is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{Male} + \beta_2 x_{OtherGender} + \beta_3 x_{Middle-income} + \beta_4 x_{Upper-income}$$
$$+\beta_5 x_{TertiaryEducation} + \beta_6 x_{UpperSecondary} + \epsilon$$

Where:

- p: the proportion of voters who will vote for Liberal Party.
- $log(\frac{p}{1-p})$: the logit of the probability of people votes for Liberal Party
- β_0: the intercept of the model.
- β_1: fix other predictor variables constant, if people is Male, compared to Female, the logit of the probability of people votes for Liberal Party changes by β_1.
- β_2: fix other predictor variables constant, if people has other gender, compared to Female, the logit of the probability of people votes for Liberal Party changes by β_2.
- β_3: fix other predictor variables constant, if people belongs to middle-income level, compared to people who belongs to low-income level, the logit of the probability of people votes for Liberal Party by β_3.
- β_4: fix other predictor variables constant, if people belongs to upper-income level, compared to people who belongs to low-income level, the logit of the probability of people votes for Liberal Party changes by β_4.
- β_5: fix other predictor variables constant, if people has Tertiary education level, compared to people has under upper secondary education, the logit of the probability of people votes for Liberal Party changes by β_5.
- β_6: fix other predictor variables constant, if people has upper secondary education level, compared to people has under upper secondary education, the logit of the probability of people votes for Liberal Party changes by β_6.

Then we fit the logistic model.

```
## # A tibble: 7 x 5
##   term                         estimate std.error statistic   p.value
##   <chr>                           <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)                    0.208     0.267     0.780  0.435
## 2 genderMale                    -0.225     0.0641   -3.51   0.000444
## 3 genderOther gender            -0.0155    0.370    -0.0419 0.967
## 4 incomeMiddle-income           -0.244     0.0835   -2.93   0.00343
## 5 incomeUpper-income            -0.488     0.124    -3.93   0.0000833
## 6 educationTertiary Education    0.241     0.263     0.917  0.359
## 7 educationUpper Secondary      -0.154     0.268    -0.575  0.566
```

Post-Stratification

The post-stratification provide an effective approach for correcting bias from overrepresented and underrepresented samples. In logistic model, the response variables are estimated for each cell; thus, the post-stratification aggregates the cell-level estimates up tp a population-level estimate by weighting each cell by its relative proportion in the population. The technique can also help discern the degree to which bias exists should a researcher choose to compare weighted versus unweighted results.(Royal)(cite)

This report analyze the sample data into cells by five weighting factors, which includes: province, gender, income, and education. There three variables are both related to one's voting decision and they both exist in survey data and census data.

$$\hat{y}^{PS} = \frac{\sum N_j \hat{y}_j}{\sum N_j}$$

where:

- $\hat{y}^{PS}$: the estimate in each cell.
- N_j : the population size of j^{th} cell based off demographics.

Result

Table 1: Logistic Model

term	estimate	std.error	statistic	p.value
(Intercept)	0.2079831	0.2665978	0.7801381	0.4353096
genderMale	-0.2250099	0.0640573	-3.5126377	0.0004437
genderOther gender	-0.0155105	0.3700053	-0.0419197	0.9665627
incomeMiddle-income	-0.2442388	0.0834735	-2.9259429	0.0034341
incomeUpper-income	-0.4880871	0.1240508	-3.9345750	0.0000833
educationTertiary Education	0.2410580	0.2629195	0.9168508	0.3592208
educationUpper Secondary	-0.1537840	0.2676738	-0.5745203	0.5656158

Table 2: Hypothesis Testing

term	estimate	p.value	threshold	significant
(Intercept)	0.2079831	0.4353096	0.05	FALSE
genderMale	-0.2250099	0.0004437	0.05	TRUE
genderOther gender	-0.0155105	0.9665627	0.05	FALSE
incomeMiddle-income	-0.2442388	0.0034341	0.05	TRUE
incomeUpper-income	-0.4880871	0.0000833	0.05	TRUE
educationTertiary Education	0.2410580	0.3592208	0.05	FALSE
educationUpper Secondary	-0.1537840	0.5656158	0.05	FALSE

From the summary of the model we have fitted above (Table 1), the logistic regression is:

$$log(\frac{p}{1-p}) = 0.2080 - 0.2250x_{Male} - 0.0155x_{OtherGender} - 0.2442x_{Middle-income} - 0.4880x_{Upper-income}$$
$$+0.2410x_{TertiaryEducation} - 0.1538x_{UpperSecondary}$$

(estimated value of parameters is rounded to 4 decimal places)

Where:

- p: the proportion of voters who will vote for Liberal Party.
- $log(\frac{p}{1-p})$: the logit of the probability of people votes for Liberal Party
- β_0: the intercept of the model.
- β_1: fix other predictor variables constant, if people is Male, compared to Female, the logit of the probability of people votes for Liberal Party changes by -0.2250.
- β_2: fix other predictor variables constant, if people has other gender, compared to Female, the logit of the probability of people votes for Liberal Party changes by -0.0155.
- β_3: fix other predictor variables constant, if people belongs to middle-income level, compared to people who belongs to low-income level, the logit of the probability of people votes for Liberal Party by -0.2442.
- β_4: fix other predictor variables constant, if people belongs to upper-income level, compared to people who belongs to low-income level, the logit of the probability of people votes for Liberal Party changes by -0.4880.

- β_5: fix other predictor variables constant, if people has Tertiary education level, compared to people has under upper secondary education, the logit of the probability of people votes for Liberal Party changes by 0.2410.
- β_6: fix other predictor variables constant, if people has upper secondary education level, compared to people has under upper secondary education, the logit of the probability of people votes for Liberal Party changes by -0.1538.

Based on table 2, it tests whenever there is a significant logistic relationship between voting for the Liberal Party and voter's gender, income level and education level. As we can see from the table, the estimated slope intercept and β_2, β_4, β_5 are not significant since their p-value are much larger than 0.05.

Table 3: The post-stratified estimates for each province

province	alp_predict	lower	upper
Alberta	0.5205895	0.5205895	0.5205895
British Columbia	0.5273675	0.5273675	0.5273675
Manitoba	0.5250255	0.5250255	0.5250255
New Brunswick	0.5303817	0.5303817	0.5303817
Newfoundland and Labrador	0.5291499	0.5291499	0.5291499
Nova Scotia	0.5299593	0.5299593	0.5299593
Ontario	0.5276288	0.5276288	0.5276288
Prince Edward Island	0.5335294	0.5335294	0.5335294
Quebec	0.5337072	0.5337072	0.5337072
Saskatchewan	0.5213603	0.5213603	0.5213603

Table 4: The post-stratified estimates

alp_predict
0.5282606

Basing off the logistic regression model with predictors variable gender, income and eduation that combined with summary table of post-stratification, the proportion of voters in favour of voting for Liberal Party is 0.5282606 as the table shows.

Discussion

Summary

The purpose of this report is to forecast the outcome of the 2019 Canadian federal election that whether the Liberal Party would still win or not if every eligible voter has voted in 2019. Thus, this report focuses on the difference in the proportion of votes between two major parties, the Liberal Party and the Conservative Party. Besides, this report also investigates the relationship between voting for the Liberal Party and voters' gender, education level and income level. Based on the two data sets, 2019 Canadian Election Study (CES) Data and 2017 General Social Survey: Families Cycle 31 (GSS) Data, the report uses a logistic regression model with predictor variables gender, income, and education to estimate the proportion of people in favour of Liberal Party. Then, post-stratification is used to adjust the sampling weights that the report concludes that the estimated percentage of voters who support the Liberal Party would be around 52.83%, which indicates that Liberal Party would still win the Federal Election if everyone has voted.

Conclusions

From the fitted logistic regression model and estimated parameters with their corresponding p-value, voters' gender as Male or Female and their income level may influence their votes for the Liberal Party or not, while voters' gender as other gender and their education level has no relationship with whether or not supporting the Liberal Party.

The post-stratification analysis shows that the proportion of voters in favour of voting for Liberal Party is 52.83%. Since our data only contains people who votes for the Liberal Party and who votes for the Conservative Party, the proportion of votes of Liberal Party exceed 50% proportion; thus, Liberal Party would still win in 2019 Federal Election in this prediction.

Weaknesses

We cannot ignore the weaknesses of our analysis. Here is a list of them:

1. Our survey data was collected during the election; thus, the response that respondent filled out in this survey may not be their final answer which reduced the accuracy of the analysis.
2. The survey data was collected in 2019 and the census data was collected in 2017; thus, the analysis might be affected by time.
3. Our model is not complex enough to accurately capture relationships between a dataset's features and a target variable since there are tons of other factors that impact the result of 2019 Federal Election.
4. The proportion of votes who votes for the Liberal Party exceeds the proportion of voters who support the Conservative party doesn't necessarily make the Liberal Party win. Since Canadian electoral systems are methods of choosing political representatives. Elections in Canada use a first-past-the-post system, whereby the candidate that wins the most votes in a constituency is selected to represent that riding.("Canadian Electoral System | The Canadian Encyclopedia")

Next Steps

For improve our analysis in the future,

1. Accessing more survey and census data always is the best way to improve the accuracy of the analysis.
2. By report analysis, education level doesn't significantly affect people's decision of voting; thus, we improve the accuracy of logistic model by removing this varibale.
3. After the final election outcome is released, we can compare our conclusions with the actual election results and do a post-hoc analysis to improve our estimation method.

References

1. Stephenson, Laura B., Allison Harell, Daniel Rubenson and Peter John Loewen. The 2019 Canadian Election Study – Online Collection. [dataset] (Stephenson et al.)

2. Paul A. Hodgetts and Rohan Alexander (2020). cesR: Access the CES Datasets a Little Easier.. R package version 0.1.0.

3. General social survey on Family (cycle 31), 2017 Retrieved Dec 01,2020 from: http://dc.chass.utoronto.ca/myaccess.html

4. Surveys And Statistical Programs - General Social Survey - Family (GSS). Www23.Statcan.Gc.Ca, 2020, https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4501#a1).

5. Bailey, Jack. "If Everyone Had Voted In 2018, Jeremy Corbyn Would Probably Be Prime Minister". Twitter, 2020, https://twitter.com/PoliSciJack/status/1327920037198499840. Accessed 22 Dec 2020.

6. Elections Canada. "Voter Turnout At Federal Elections And Referendums – Elections Canada". Elections.Ca,2020,https://www.elections.ca/content.aspx?section=ele&dir=turn&document=index&lang=e.

7. Royal, KennethD. Survey Research Methods: A Guide For Creating Post-Stratification Weights To Correct For Sample Bias. 2020.

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

9. 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.

10. Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.

11. David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.3.https://CRAN.R-project.org/package=broom

12. "Federal Election 2019 Live Results". CBC News, 2020, https://newsinteractives.cbc.ca/elections/federal/2019/results/.

13. "Canadian Electoral System | The Canadian Encyclopedia". Thecanadianencyclopedia.Ca, 2020, https://www.thecanadianencyclopedia.ca/en/article/electoral-systems.