Statistical Analysis of the outcome of 2019 Canadian Election if everyone voted

Yichen Su

Contents

1 Information

Topic: Statistical Analysis of the outcome of 2019 Canadian Election if everyone voted

Author: Yichen SU

Date: December 22, 2020

Code and data for this analysis is available at: https://github.com/YichenSu529/STA304-final-project-Statistical-Analysis-of-the-outcome-of-2019-Canadian-Election-if-everyone-voted.git

2 Abstract

Politicians always aim for an increased turnout rate in the election, which has brought the curiosity of the outcome of having everyone votes for the election. In this study, we will analyze the proportion of people voting for Liberal and Conservative Parties in the 2019 Canadian Federal Election if having everyone votes. Since there are many factors such as sex, age, marital status, people's interest in politics, and other individuals' opinions that could affect their vote choice in the election, a combination of multiple logistic models and post-stratification method will be used in our study. The outcome suggests that the Conservative Party has more popular vote than the Liberal Party, which is consistent with the results in the 2019 Canadian Federal Election.

Keywords: 2019 Canadian Federal Election, Election, Turnout, multiple logistic model, post-stratification, Stepwise selection

3 Introduction

The statistical technique and knowledge have been commonly used to produce the analysis regarding political outcomes or issues. Since there are more data regarding political science that combined many people's attitudes, opinions, political and economic measures, the election prediction can be made based on these motivations. The decreasing voters' turnout for the Canadian Federal Election becomes one of the major issues. And as we know the 2019 Canadian federal election, the Liberal Party 33.1% of the popular vote, which is lower than the Conservative Party with 34.4% despite the winning of the overall Federal Election (Dunham, B 2019). Therefore, it is interesting to know whether the outcome for the popular vote of these two parties is different from the actual results if consider "everyone" is voting.

The way to provide the analysis in predicting the election result is to apply the statistical technique called multilevel regression and post-stratification (MRP) methods. This similar approach has been used by many professional analysts in predicting the national votes using past election polling and many sources of information(Lauderdale, Bailey, Blumenau & Rivers). Therefore, in this report, the multilevel regression model and post-stratification method can be used in predicting the proportion of people voting Liberal Party and people voting Conservative Party combined with factors of voters' attitudes. In this case, we can determine whether the sample is representative, and the factors are responsible for people's election decision.

In this report, two datasets are being used to navigate the proportional is voter's voting for two main political parties respectively if having everyone voting and how it is consistent with the actual result. One is obtained from the Canadian Election Study (CES) and another is obtained from the Canadian General Social Survey (GSS). In the Methodology section(section2), we would discuss in detail the data cleaning process and how the MRP model is stimulated. And as moving on to the Result section (Section3), the performance of our model in determining the proportion of votes for two parties will be shown. Lastly, in the Discussion section(section4), the statistical finds and consequences will be discussed and the potential weakness of the analysis and further improvement for the investigation will also be addressed.

4 Model

4.1 Data

Two datasets are being used in this report. One dataset is retrieved from the 2019 Canadian Election Study (CES). The `ces2019_web` dataset is the sub-dataset of the 2019 CES online that includes some behaviors between voting and election. This dataset is used as the survey dataset because it contains numerous variables about the opinions and attitudes of the 2019 federal election. It also includes a lot of variables associated with social, economic, and political issues. And the other is retrieved from the 2013 Canadian General Social Survey (GSS) that contains lots of information about the political and social interests of the Canadians and the raw dataset of the 2013 GSS can be further used as the census data. The target population of the study is all Canadian who are allowed to vote. And the frame population is the people who do the survey online and by phone.

4.2 Data Cleaning

The `ces2019_web` dataset and the raw dataset from the 2013 GSS are both raw data that contain a large number of variables, observations, and missing values. So, the data cleaning process is crucial. We will select some variables regarding the social, economic, and political opinions that could contribute to people's voting choices. The chosen 10 variables are selected and renamed in the survey dataset in R script. The important variables `vote_liberal` and `vote_conservative` are also created based on the variables `votechoice`, which contain only 1 and 0. The original value 1 in `votechoice` means voting for liberal is assigned as 1 in `vote_liberal` and value 2 in `votechoice` means voting for conservative is referred to as 1 in `vote_conservative`. Since we are considered everyone is voting, people who are defined as `somewhat_unlikely` in `likely_to_vote` have NA in `votechoice`, which means they are unable to choose a value in `votechoice`. Therefore, another variable `unlikely_vote` has those people's vote choice if they decide to vote. In this case, `vote_liberal` and `vote_conservative` should have also included those who are somewhat unlikely in `likely_to_vote` but has chosen to vote for these two parties if they vote. These have been produced in R script.

Before running the model selection, we need to mutate the variables of survey datasets; so that it is consistent with the census dataset to proceed post-stratification. The variables that are collected and renamed are `sex`, `age`, `if_married`, `likely_to_vote`, `edu_level`,`household_income`, `participate_volunteer`, `interest_politic`, `votechoice`, and `unlikely_vote`. Besides, census data are cleaned by partially using the R script that previously provided for GSS2017 and will be used for post-stratification later. And the mutation of variables should be similar to the survey dataset.

Table 1: baseline characteristics of the survey data

	Overall(N=23445)
sex	
Female	12734 (54.3%)
Male	10711 (45.7%)
age	
15 to 34 years	5129 (21.9%)
35 to 54 years	8258 (35.2%)
55 to 74 years	8968 (38.3%)
75 years and over	1090 (4.6%)
if_married	
Mean (SD)	0.472 (0.499)
Median [Min, Max]	0 [0, 1.00]
likely_to_vote	
Don't know or not eligible	1714 (7.3%)

	Overall(N=23445)
somewhat likely	3118 (13.3%)
somewhat unlikely	726 (3.1%)
very likely	17541 (74.8%)
very unlikely	346 (1.5%)
edu_level	
Bacholor's degree or higher	11293 (48.2%)
Post-secondary dipolma	7727 (33.0%)
equal or lower than highschool	4425 (18.9%)
participate_volunteer	
no	10355 (44.2%)
yes	12228 (52.2%)
interest_politic	
not at all interested	383 (1.6%)
not very interested	4287 (18.3%)
somewhat interested	16384 (69.9%)
very interested	2391 (10.2%)
vote_liberal	
Mean (SD)	0.271 (0.444)
Median [Min, Max]	0 [0, 1.00]
vote_conservative	
Mean (SD)	0.260 (0.439)
Median [Min, Max]	0 [0, 1.00]

Table1 demonstrates the baseline chracteristics of the cleaned survey data from 2019CES, which hase the reduced amounts of variables. It shows the total 23445 observations of the data without the missing values. It shows the proportion of the each subcategory's observations among the total observation. For example, it presents that 13.3% of people pose that they are somewhat likley to vote for the 2019 election. This similar interpretations can apply to all the variable's subcategory.

4.3 Model Specifics

The proportion of people voting for the Liberal Party and Conservative Parties in the 2019 Federal Election can be estimated using the Multiple Logistic Regression Model with Post-Stratification. The Multiple Logistic Regression Model is usually used in producing analysis of binary outcomes with many independent variables. Since the response variables for the analysis are `vot_liberal` and `vote_conservative`, which only have 0 and 1. Also, the variables that are chosen for the analysis are mainly independent categorical variables; therefore, the Multiple Logistic Regression Model is the primary choice. Since our sample size for the Multiple Logistic Regression Model is much larger than the number of variables, using the stepwise selection can make the model more generalized. And the stepwise backward has the advantage to consider all the variables at the same time that can keep more variables in the model as possible. Therefore, the final model is determined by using the Backward stepwise selection. This function starts with the model with full variables and then removes the least significant variables one at a time until the model reaches the specific threshold of 0.05 or 0.2 (n.d.). It provides the model in predicting the votes for Liberal with the lowest possible AIC value 25610. And the final model in predicting the votes for Conservative with the AIC value 24820.

This is the model of predicting the votes for Liberal:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_{1age_{35-54}} + \beta_2 X_{2age_{55-74}} + \beta_3 X_{3age_{75+}} + \beta_4 X_{4marr_1} + \beta_5 X_{5likely_{SL}}$$

$$+\beta_6 X_{6likely_{SU}} + \beta_7 X_{7likely_{VL}} + \beta_8 X_{8likely_{VU}} + \beta_9 X_{9likely_{VL}} + +\beta_1 0 X_{10edu_{PSD}} + \beta_1 1 X_{11edu_H} + \beta_1 2 X_{12politic_{NVI}}$$

$$+\beta_{13}X_{13politic_{SI}} + \beta_{14}X_{14politic_{VI}} + \beta_{15}X_{15volunteer_{na}} + \beta_{1}6X_{16volunteer_{yes}} + \epsilon$$

(see notation in Appendex A)

This is the model of predicting the votes for Conservative:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_{sex_{Male}} + \beta_2 X_{2age_{35-54}} + \beta_3 X_{3age_{55-74}} + \beta_4 X_{4age_{75+}} + \beta_5 X_{5marr_1} + \beta_6 X_{6likely_{SL}}$$

$$+\beta_7 X_{7likely_{SU}} + \beta_8 X_{8likely_{VL}} + \beta_9 X_{9likely_{VU}} + \beta_{1}0X_{10edu_{PSD}} + \beta_{1}1X_{11edu_H} + \beta_{1}2X_{12politic_{NVI}}$$

$$+\beta_{1}3X_{13politic_{NVI}} + \beta_{14}X_{14politic_{SI}} + \beta_{15}X_{15politic_{VI}} + \beta_{16}X_{16volunteer_{na}} + \beta_{1}7X_{17volunteer_{yes}} + \epsilon$$

log is the natural logarithm. p is the proportion of people voting Liberal or Conservative Parties in 2019 Canadian Federal Election. $\frac{p}{1-p}$ as the "odd ratio" and $log(\frac{p}{1-p})$ is the log odds ratio X_i represents the predictor variables in our model. e.g. $X_{1age_{35-54}}$ is the input variable of age between 35 to 54 years old, similarly to the rest X_i in the equation. The X_i here is the dummy variables that only takes value 0 or 1. β_is (for i from 1 to 17/18) are the coefficients for each of the subcategories, which represent the average difference in the log of odds ratio between $X_i = 0$ and $X_i = 1$ when having other variables constant. e.g. for the model of voting Conservative, β_8 represents the average difference in the log of odds ratio between married and not married people holding other variables constant. $\beta_0$is the constant term that represents the intercept at time zero; so, β_0 is the value of $logit(p)$ when having every $X_i = 0$. ϵ is the error term of the model.

4.4 Post-Stratification

To provide a more accurate estimate of the probability of people voting Liberal and Conservative party., we will produce the analysis using the post-stratification method. Since the data is composed of variables that contain different subcategories and each subcategory has different sizes, we are partitioning them into different cells to provide rebalance later. The basic idea of post-stratification is to correct the imbalance between different strata(subcategory). So, by using a weighted average of the averages in each subcategory, we would have the rebalanced estimate of the population mean(Reilly, Gelman, Katz, C. n.d.). In our study, the census data has been cleaned and is ready for the use of post-stratification. In this case, we can use the above models to estimate the probability of people voting in the Liberal and Conservative Party based on each subcategory. Here is the mathematical calculation for the above procedures.

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

For the above equation, $\hat{y}^{PS}$ can be defined as the proporion of voters voting for Liberal or the proporion of voters voting for Conservative with regarf to everyone would vote. $\hat{y}_j$ is the estimated probability in j^{th} cell. N_j is the population size in j^{th} cells. $\sum N_j$ is the overall population.

5 Results

Table 2: The Predicted Proportion of Voters for Liberal

predict
0.22

We estimate the proportion of people voting for Liberal Pary in the condition of everyone is voting with an equation of $\hat{y}^{PS} = \frac{\sum N_j \hat{y}_j}{\sum N_j}$ (Caetano, 2020), is estimated to be **0.22**. It means that there are 22% of the people voting for the Liberal Paty. This prediction value is based on the post-stratification analysis of the proportion of voters voting the Liberal Party from a logistic regression model, `survey_model`, based on the variables `age`, `if_married`,`likely_to_vote`, `edu_level`, `interest_politic`,and `participate_volunteer`.

Table 3: The Predicted Proportion of Voters for consservative

predict
0.264

We predict the proportion of people voting for Conservative Pary in the condition of everyone is voting with an equation of $\hat{y}^{PS} = \frac{\sum N_j \hat{y}_j}{\sum N_j}$ (Caetano, 2020), is estimated to be **0.264**. It means that there are 26.4% of the people voting for the Conservative Paty. This estimate is based on the post-stratification analysis of the proportion of voters voting the Conservative Party from a logistic regression model, `survey_model2`, based on the variables `sex`, `age`, `if_married`,`likely_to_vote`, `edu_level`, `interest_politic`,and `participate_volunteer`.

Table 4: Summary - AIC Score of final Models

Model Name	AIC
model for voting liberal	25610
model for voting conservative	24820

By using the stepwise selection from part I, we can obtain the the final models `survey_model` and `survey_model2` with the lowest possible AIC values, which are 25610 and 24820 respectively.

Table 5: Summary for the Survey Logistic Regression Model of voting Liberal

Term	Estimate	Standard Error	Test Statistic	P-Value
(Intercept)	-17.960	94.332	-0.190	0.849
age35 to 54 years	0.031	0.042	0.743	0.458
age55 to 74 years	0.092	0.043	2.166	0.030
age75 years and over	0.150	0.077	1.939	0.053
if_married	-0.026	0.031	-0.838	0.402
likely_to_votesomewhat likely	16.764	94.331	0.178	0.859
likely_to_votesomewhat unlikely	16.514	94.331	0.175	0.861
likely_to_votevery likely	16.734	94.331	0.177	0.859
likely_to_votevery unlikely	0.356	229.603	0.002	0.999
edu_levelPost-secondary dipolma	-0.428	0.035	-12.332	0.000
edu_levelequal or lower than highschool	-0.556	0.044	-12.687	0.000
interest_politicnot very interested	0.353	0.176	2.000	0.046
interest_politicsomewhat interested	0.729	0.174	4.180	0.000
interest_politicvery interested	0.786	0.180	4.366	0.000
participate_volunteernot answer	-0.181	0.086	-2.108	0.035
participate_volunteeryes	-0.191	0.032	-6.048	0.000

Table 6: Summary for the Survey Logistic Regression Model of voting Conservative

Term	Estimate	Standard Error	Test Statistic	P-Value
(Intercept)	-18.705	93.336	-0.200	0.841
sexMale	0.389	0.032	12.271	0.000
age35 to 54 years	0.104	0.044	2.341	0.019
age55 to 74 years	0.092	0.045	2.046	0.041
age75 years and over	0.309	0.078	3.982	0.000
if_married	0.523	0.032	16.261	0.000
likely_to_votesomewhat likely	16.482	93.336	0.177	0.860
likely_to_votesomewhat unlikely	16.386	93.336	0.176	0.861
likely_to_votevery likely	16.679	93.336	0.179	0.858
likely_to_votevery unlikely	0.120	227.935	0.001	1.000
edu_levelPost-secondary dipolma	0.383	0.035	10.833	0.000
edu_levelequal or lower than highschool	0.478	0.043	11.177	0.000
interest_politicnot very interested	0.100	0.165	0.605	0.545
interest_politicsomewhat interested	0.298	0.163	1.828	0.068
interest_politicvery interested	0.446	0.169	2.635	0.008
participate_volunteernot answer	0.070	0.087	0.805	0.421
participate_volunteeryes	0.190	0.032	5.867	0.000

Because `survey_model` and `survey_model2` are two models that predicting the proportion of poppular vote for the Liberal and Conservative Parties, we can do summary tables that form the estimated value for β_i.In Table5 and Table6, we have the estimate values for each input and also estimated value for $\hat{\beta}_0$, which these values are shown in the fitted logistic regression models below.

$$log(\frac{p}{1-p}) = -17.96023+0.03134X_{1age_{35-54}}+0.09236X_{2age_{55-74}}+0.14957X_{3age_{75+}}-0.02625X_{4marr_1}+16.76414X_{5likely_{SL}}$$

$$+16.51438X_{6likely_{SU}}+16.73371X_{7likely_{VL}}+0.35580X_{8likely_{VU}}-0.42751X_{10edu_{PSD}}-0.55589X_{11edu_H}+0.35277X_{12politic_{NVI}}$$

$$+0.72852X_{13politics_{SI}} + 0.78568X_{13politic_{VI}} - 0.18115X_{15volunteer_{na}} - 0.19057X_{16volunteer_{yes}} + \epsilon$$

`survey_model`is the model that are finally selected in predicting the proportion of people voting for the Liberal Pary in 2019 Canadian Federal Election. Based on the summary table for the model, we can determine the estimated values of the intercept, $\hat{\beta}_0$ and estimated slopes $\hat{\beta}_i$ (for i = 1 to 16). Therefore, the above is the fitted logistic regression for voting Liberal.

$$log(\frac{p}{1-p}) = -18.70489+0.38948X_{sex_{Male}}+0.10394X_{2age_{35-54}}+0.09181X_{3age_{55-74}}+0.30885X_{4age_{75+}}+0.52315X_{5marr_1}$$

$$+16.48228X_{6likely_{SL}}+16.38572X_{7likely_{SU}}+16.67933X_{8likely_{VL}}+0.12037X_{9likely_{VU}}+0.38277X_{10edu_{PSD}}+0.47766X_{11edu_H}$$

$$+0.1X_{12politic_{NVI}}+0.29828X_{13politics_{SI}}+0.44613X_{14politic_{VI}}+0.06968X_{15volunteer_{na}}+0.19041X_{16volunteer_{yes}}+\epsilon$$

`survey_model2` is the model that are finally selected in predicting the proportion of people voting for the COnservative Pary in 2019 Canadian Federal Election. Based on the summary table for the model, we can determine the estimated values of the intercept, $\hat{\beta}_0$ and estimated slopes $\hat{\beta}_i$ (for i = 1 to 17). Therefore, the above is the fitted logistic regression for voting Conservative.

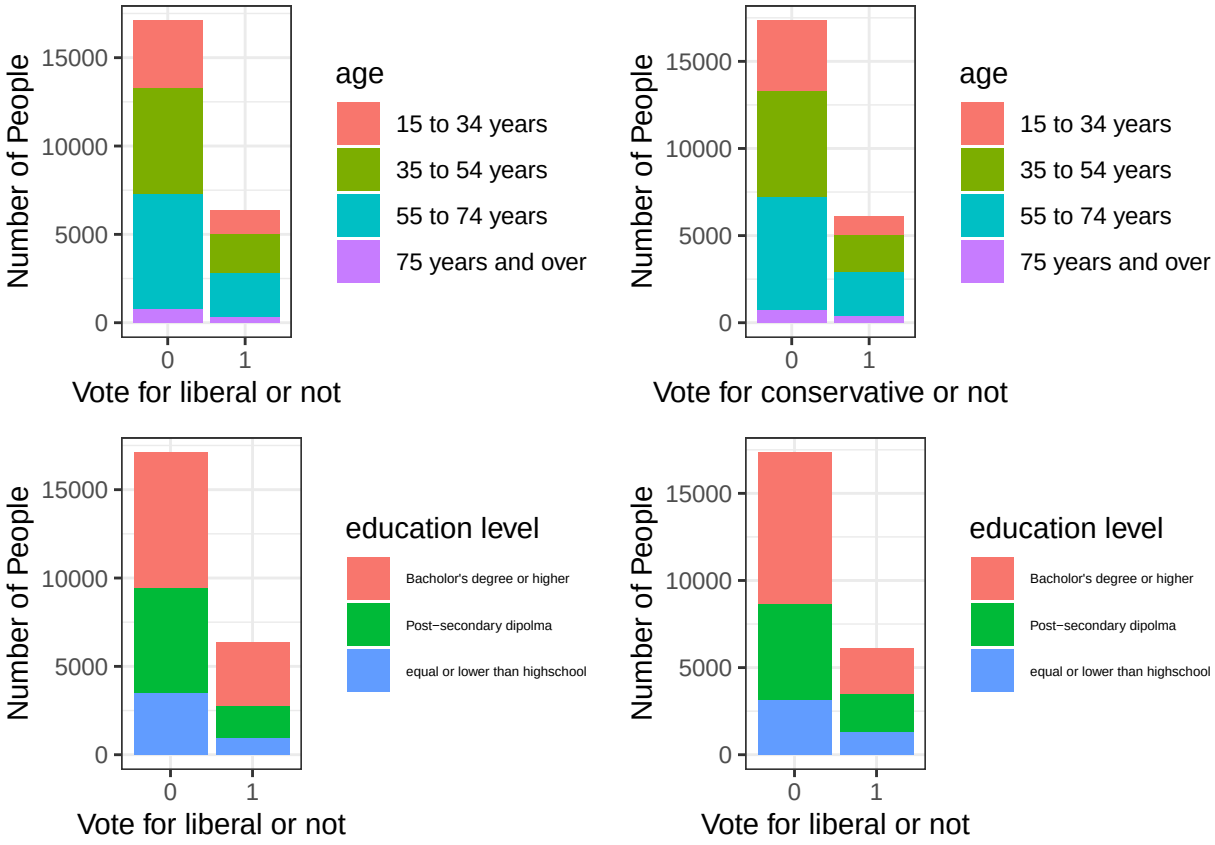

Figure 1: Relationship between vote choice and number of voters by age, ehucation level

In Figure1, four bar plots above illustrate the relationships between the voters' choice and number of voters in the 2019 federal Election, grouped by the independent variables in the logistic `age` and `edu_level`. Each variable are shown in two bar plots, one is for voting to Liberal and one is for voting to Conservative.

In Figure2, four bar plots above illustrate the relationships between the voters' choice and number of voters in the 2019 federal Election, grouped by the independent variables in the logistic `interest_politic` and `Likely_to_vote`. Each variable are shown in two bar plots, one is for voting to Liberal and one is for voting to Conservative.

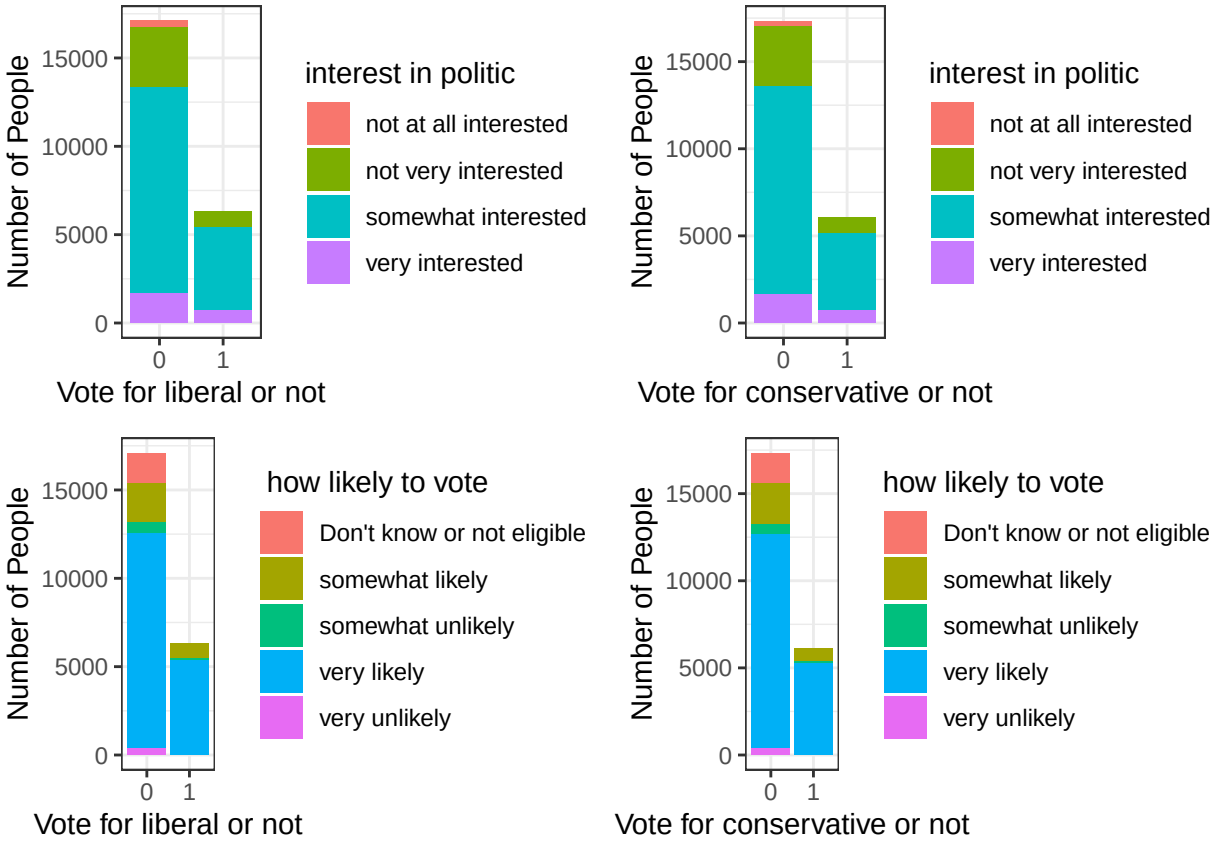

Figure 2: Relationship between vote choice and number of voters by interest_politic, liklihood to vote

6 Discussion

6.1 Summary and Conclusion

The objective of this report is to demonstrate the outcome of the popular vote between Liberal and Conservative Parties for the 2019 Canadian Federal Election if having the condition that everyone would vote. The dataset we used is survey and census data. The survey data retrieved from 2019 CES is used for doing a multilevel logistic model. And the census data is from the GSS2013 is used for post-stratification. Since the study is assuming everyone would vote, we would combine another variable `unlikley_vote` that collect vote choice from those who are unlikely to vote. This is because they do not have the vote choice in 'votechoice' previously, we now should also be included in `vote_liberal` or `vot_conservative` if any of them is voting for them. By using the stepwise backward method, the variables in ultimate models for both parties are finalized as well as their logistic regression model. Then, the prediction on the popular vote for both parties, with the condition of everyone is voting, can be done by apply post-stratification.

According to the outcome of our study, there are 22% of the people voting for Liberal and 26% of the people voting for Conservative, considering that everyone is voting. Based on the outcome, we can see that none of the parties holds the large majority of voters that could form the majority government directly. This is consistent with the fact that the Liberal party ends up holding minority government in the 2019 Canadian Federal election. Also, the result presents that the Conservative Party has more popular vote than the Liberal Party, which matches with the actual results of the 2019 Canadian Federal election. Therefore, even with assuming everyone would vote, the predicted outcome has no difference from the actual outcome in the 2019 Canadian Federal election. This can relates Figure2 and Table1 that we can see a small portion of people that is somewhat unlikely to vote. So, as we consider them into the proportion of them voting for Liberal or Conservative, there is no surprising change to the result when having those people's votes of the two major parties. Besides, for Figure1 and Figure2, we are also interested in looking at the proportion of votes for both parties concerning the subcategories of some variables from the models. And there is no major difference in the proportion of the votes between two parties when we compare them with the same variables.

Overall, the actual voters' turnout for the 2019 Canadian Federal Election was 66%, which is slightly lower than the turnout in 2015 (Stright News). However, the outcome for the study stays consistent with the actual outcome in 2019, in which the predicted proportion of people voting Conservative is slightly higher than the proportion of people voting Liberal. This can be due to the fact of the proportion of people that are unlikely to vote occupies a small proportion of eligible votes. If that is the case, the survey data that we use for the study is pretty representative of the population.

6.2 Weakness

There are many limitations to using the logistic regression model. Logistic regression has the assumption of having a linear association between the dependent variables and the independent variables(GeeksForGeeks 2020). Therefore, it presents the restriction if we have a non-linear relationship between the variables. Besides, before we do the logistic regression model, we would require carrying out the variable selections. This process will require the identification of potential variables (Torsten, H. & Jürgen, D. 2013) So, logistic regression doesn't have a strict role in determining the best model. Since the method that we use to finalize the variables is stepwise regression, there is no guarantee of selecting the best variables, especially with a small dataset. In the study, we do not provide a linearity check for the variables. But since all the variables for the model are categorial, there should not have a large issue with the linearity problem. Also, the sample size of the model is large; therefore, it should decrease the instability of using stepwise regression(Choueiry G.). But it does not mean to guarantee the best model. On the other hand, since we aim to select the variables that can be matched in both datasets. Therefore, some potential variables can be missed in our model.

6.3 Next Step

After discussing some limitations and weaknesses of our model, there exists some problems that we can improve. If we are predicting a similar study in the future, it is important to check the linearity of variables, which we can make scatterplots to check the patterns. Besides, since we are only dealing with few predicting variables from both datasets and the census dataset is collected from 2013 GSS, it might not be representative enough to the analysis dealing with the 2019 election. Since the limited number of variables, we can use, the prediction of the election might be inefficient. Therefore, the improvements can be having a more updated census dataset that contains more variables regarding sociology, economics, and political issues. In this case, we would have more variables that could affect the election results.

7 Reference

7.1 Survey Data and Census Data(2013GSS and 2019CES)

General Social Survey, cycle 27, 2013 (version 2): Social Identity.Retreived from

https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/hsda?harcsda4+gss27v2

Stephenson, Laura B., Allison Harell, Daniel Rubenson and Peter John Loewen.

The 2019 Canadian Election Study – Online Collection. [dataset]

7.2 Other referneces

Dunham, B. (2019, October 22). Conservatives win popular vote but lose election. Retrieved from

https://election.ctvnews.ca/conservatives-win-popular-vote-but-lose-election-1.4649651

Lauderdale, B., Bailey, D., Blumenau, J., & Rivers, D. (2019, October 15). Model-based pre-election

polling for national and sub-national outcomes in the US and UK. Retrieved December 22, 2020, from

https://www.sciencedirect.com/science/article/pii/S016920701930189X?casa_token=wXkw-NnFllYAAAAA%3AkCS95QZqU6b3lfudLl7I1fLaUFjYEcgwObDUYCok6woovSIRmQM7D8N3Ic2d29WYkF-kERR1fva7

Understand Forward and Backward Stepwise Regression. (n.d.). Retrieved from

https://quantifyinghealth.com/stepwise-selection/

Reilly, Gelman, Katz, C. (n.d.). Poststrati cation Without Population Level Information on the

Poststratifying Variable, With Application to Political Polling. Retrieved from

https://stat.columbia.edu/~gelman/research/published/aprvlRv1.pdf

Voter turnout for Canada's 2019 federal election was, well, meh. (2019, October 23). Retrieved from,

https://www.straight.com/news/1316836/voter-turnout-canadas-2019-federal-election-was-well-meh

Torsten Heyer & Jürgen Stamm (2013) Levee reliability analysis using logistic regression models – abilities,

limitations and practical considerations, Georisk: Assessment and Management of Risk for Engineered

Systems and Geohazards, 7:2, 77-87, DOI: 10.1080/17499518.2013.790734

GeeksForGeeks (2020, September 02). Advantages and Disadvantages of Logistic Regression.

Retrieved from, https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

Choueiry George (n.d) Understand Forward and Backward Stepwise Regression. Retrieved from,

https://quantifyinghealth.com/stepwise-selection/

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

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

Xie,Y. (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Xie,Y. (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

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

Wickham,H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Auguie,B (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

8 Appendix

8.1 Appendix A

$log(\frac{p}{1-p})$ is the log odd ratio we estimate the log odd proportion of people voting the Liberal Party.

$X_{sex_{Male}}$ is 1 if the input variable of Male and 0 otherwise.

$X_{2age_{35-54}}$ is 1 if the input variable of age between 35 to 54 years old and 0 otherwise.

$X_{3age_{55-74}}$ is 1 if the input variable of age between 55 to 74 years old and 0 otherwise.

$X_{4age_{75+}}$ is 1 if the input variable of age equal or over 75 years old and 0 otherwise.

X_{5marr_1} is 1 if the input variable is married with the value 1 and 0 otherwise.

$X_{6likely_{SL}}$ is 1 if the input variable `likely_to_vote` is somewhat likely and 0 otherwise.

$X_{7likely_{SU}}$ is 1 if the input variable `likely_to_vote` is somewhat unlikely and 0 otherwise.

$X_{8likely_{VL}}$ is 1 if the input variable `likely_to_vote` is very likely and 0 otherwise.

$X_{9likely_{VU}}$ is 1 if the input variable `likely_to_vote` is very unlikely and 0 otherwise.

$X_{10edu_{PSD}}$ is 1 if the input variable `edu_level` is Post-secondary diploma and 0 otherwise.

X_{11edu_H} is 1 if the input variable `edu_level` is equal or lower than high school and 0 otherwise.

$X_{12politic_{NVI}}$ is 1 if the input variable `interest_politic` is not very interested and 0 otherwise.

$X_{13politic_{SI}}$ is 1 if the input variable `interest_politic` is somewhat interested and 0 otherwise.

$X_{14politic_{VI}}$ is 1 if the input variable `interest_politic` is very interested and 0 otherwise.

$X_{15volunteer_{na}}$ if the input variable `participate_volunteer` is not anser and 0 otherwise.

$X_{16volunteer_{yes}}$ if the input variable `participate_volunteer` is yes and 0 otherwise.

β_is represent the average difference in the log of odds ratio $X_i = 0$ and $X_i = 1$ when having other variables constant, where X_i matches above.

$\beta_0$is the constant term that represents the intercept at time zero.