Prediction on 2019 Canadian Election Result using MRP

Yao He

12/20/2020

Abstract

Canadian election result is closely related to all Canadian citizens as the party in charge determines the public policies and public benefits. In this study, I will predict the 2019 Canadian election result - the probability of Justin Trudeau and the Liberal Party winning the election, using the multilevel regression & post stratification based on the data from General social survey Family 2017(Statistics Canada, 2017) and 2019 Canadian Election Study - Online Survey(Stephenson; Harell; Rubenson; Loewen, 2019). From the study, the predicted result is 30%.

Keywords

Canada 2019 Election; Justin Trudeau; Liberal Party; Multilevel regression; Post-stratification.

Introduction

Every Canadian citizen must have some public issues that they care, such as social benefits, climate changes, and so on. Meanwhile, citizens from different provinces, such as Quebec and Alberta, tend to care about more province specific issues, such as Quebec independence and trans Mountain oil pipelines programs. These issues largely depend on the federal policies designed by the party in charge(BBC, 2019). Hence, the election is actually related to all Canadian citizens and determinant to their daily life. It is likely that citizens with different age, gender, and education, living in different provinces might weigh differently about the polices that the elected party may implement. Thus, the study is about the election result prediction and I am going to predict the probability of Trudeau winning the election using the factors age, gender, province, and education.

I will build a model using the survey data obtained from 2019 Canadian Election Study - Online Survey, and perform a post-stratification to census data obtained from General social survey Family 2017 to test the model and make a prediction.

The report can be break down into the following four parts: introduction to the data sets I obtained, model and methodology that I used, results from the model, and further discussions.

Data

1. Survey data - CES dataset

The CES data (2019 Canadian Election Study - Online Survey, 2019) is obtained through Harvard dataverse.

1.1 Target Population

The target population, according to the user's guide, is all eligible voters in Canada, who are Canadian citizens and permanent residents aged 18 or over.

1.2 Frame Population

The survey is conducted through Qualtrics, hence the frame population is Canadian citizens and permanent residents aged 18 or over who have access to Qualtrics.

1.3 Sampling Approach

The samples are taken by voluntary sampling method. All cases from the frame population have the chance to participate in the survey and they may choose to participate in the survey or not.

1.4 Sample Population

The sample data, which is the dataset we obtained, are the Canadian citizens and permanent residents aged 18 or over who have access to Qualtrics and choose to take the survey voluntarily.

1.5 Methodology of collecting data

Data were collected by online surveys where voters who wish to take the survey may take at any time and data are recorded by Quualtrics. The target result is to get 100 responses per day.

1.6 Strength of the dataset

The dataset only includes cases where they respondents answer the survey questions faithfully and correctly. Incomplete responses, duplicate responses of previous respondents, speeders, those who "straight-lined" grid questions ("straightliners"), and respondents whose postal code didn't match their province have all been removed from the data file, and are excluded from numbers reported in this codebook(Stephenson; Harell; Rubenson; Loewen, 2020).

1.7 Weakness of the dataset

The survey is designed to enable respondents to refuse to answer any given question. In order to achieve this goal, the survey extensively use "Don't know/ Prefer not to answer" options, For example, some questions had the text "If you do not know, or prefer not to answer, please click ???" at the end of them. For those questions, if the respondent did not respond to that question or a component of that question (for example, did not click on or move a slider), then their response to that question, or that component of the question, was recorded as missing(Stephenson; Harell; Rubenson; Loewen, 2020). Although this gives respondents high freedom of answering the survey, this may create many cases that may not be useful for our study.

1.8 Sample data

I am going to predict the probability of Trudeau winning the election by the variable age, sex, education and province, hence I select these variables as well as their vote choice from the dataset. I create one variable that specifically indicate whether the respondent will vote for Trudeau or not. In order to get a more accurate prediction, I eliminate the cases with NA variables in the dataset.

- Age: Ages of respondents are recorded as numbers directly.

- Education: In order to match the variables between the survey data and the census data, I combine several categories of education, and get the following five different categories for education level (NA cases where the respondent's education may not correctly respond to the five levels are also eliminated):
1. "Less than high school diploma or its equivalent";
2. "High school diploma or a high school equivalency certificate";
3. "College, CEGEP or other non-university certificate";
4. "University certificate, diploma or degree above the bachelor's level";
5. "Bachelor's degree (e.g. B.A., B.Sc., LL.B.).

- Sex: Sex were considered as two different values: male and female.

- Province: There are only ten provinces that are included in the dataset, excluding the three territorries. They are: Newfoundland and Labrador, Prince Edward Island, Nova Scotia, New Brunswick, Quebec, Ontario, Manitoba, Saskatchewan, Alberta, and British Columbia.

- vote_Trudeau: This variable is a binary variable with only values of 0 and 1, where 1 represents the voter will vote for Trudeau and 0 represents the voter will not.

2. Census data - GSS dataset

The GSS data (General social survey on Family (cycle 31), 2017) is obtained through Statistics Canada.

2.1 Target Population

The target population, according to the user's guide, is all non-institutionalized persons 15 years of age and older, living in the 10 provinces of Canada, not including people who lived in Yukon, Northwest Territories, and Nunavut; and full-time residents of institutions.(statistic Canada, 2020, p.11).

2.2 Frame Population

The study creates a new frame that combines telephone numbers (landline and cellular) with Statistics Canada's Address Register, and collects data via telephone. And each case is assigned into a stratum within its province.

2.3 Sampling Approach

The samples are taken by stratified sampling method. All cases from the frame population are assigned to a stratum within its province and then sample population are taken using a simple random sample without replacement from each strata in the frame population.

2.4 Sample Population

The sample data, which is the dataset we obtained, are the households where there is at least one person and are selected using the sampling approach mentioned above from the frame population.

2.5 Methodology of collecting data

Data were collected by interviewing these selected family units via Computer assisted telephone interviews. All telephone interviews took place from 9:00 a.m. to 9:30 p.m. Mondays to Fridays; Saturdays from 10:00 a.m. to 5:00 p.m. and Sundays from 1:00 a.m. to 9:00 p.m..
To encourage the participation of the household who refused the interview, they would be reached up to two more times to explain the importance of the survey. Households whose timing was inconvenient will be able to reschedule another call. Households who missed the call would be contacted numerous times.
The variables we used for this study are age, sex, education, and province.
These socia-demoraphic classification were conducted under Statistics Canada's harmonized content questions, which are standardized modules for household survey variables.

2.6 Strength of the dataset

Interviewers try to get as much participation as possible such that the data is representative of the target population.
In order to keep the accuracy of the data, the survey contains multiple questions related to the same topic as the way to prevent non-thinking answers (i.e respondent may randomly choose an option without thinking). For example, instead of only asking what institution that response attended/is attending, the survey also asked them what the highest certificate, diploma or degree that they have completed is. By comparing the answers of these two questions, the data reflected more accurate responses.

2.7 Weakness of the dataset

The survey is a long phone interview which interviewees may get tired and would cause the non-meaningful response. The response rate of gss.csv was 52.4%.
Moreover, the questions in the survey may be improved through the way that questions are worded. For example, when asking the education level, the question used "What type of educational institution (are you attending/did you attended)?" followed by some types of institution. But the options for the interviewees are "Yes/No". It is unclear for the respondent to use a binary option to answer an open type questions. It would be better to make the question like "Are you attending/did you attended ..." followed by some types of institution.

2.8 Sample data

I am going to predict the probability of Trudeau winning the election by the variable age, sex, education and province, hence I select these variables from the dataset "gss.csv". In order to get a more accurate prediction, I eliminate the cases with NA variables in the dataset.

- Age: Age respondents are recorded as numbers directly.

- Education: In order to match the variables between the survey data and the census data, I combine several categories of education, and get the following five different categories for education level:
1. "Less than high school diploma or its equivalent";
2. "High school diploma or a high school equivalency certificate";
3. "College, CEGEP or other non-university certificate";
4. "University certificate, diploma or degree above the bachelor's level";
5. "Bachelor's degree (e.g. B.A., B.Sc., LL.B.).

- Sex: Sex were considered as two different values: male and female.

- Province: There are only ten provinces that are included in the dataset, excluding the three territorries. They are: Newfoundland and Labrador, Prince Edward Island, Nova Scotia, New Brunswick, Quebec, Ontario, Manitoba, Saskatchewan, Alberta, and British Columbia.

Model/Methodology

I will use a logistic regression model by the glm function in R to model the proportion of voters who will vote for Justin Trudeau with factors of age, sex, province, and education. And I will use backward elimination using AIC to get an alternative model. I will then compare the results we get from the two models. The main reason for choosing the logistic regression is that the dependent variable that we care about is the probability of voting for Justin Trudeau (vote_Trudeau). It is a binary variable. I want to make a prediction of the probability of Trudeau winning the election and it makes no sense if the predicted value is greater than 1 or less than 0.

The explanatory variables are:
• Age is a numerical variable that represents the age of a given person. From the two datasets, age is surveyed by inputting the numbers, hence we continue to use to numerical values here. Meanwhile, numerical values are more accurate than just putting age groups for the prediction. Age is included because people of different ages may weigh the political policies differently and tend to have different opinions on the Party in charge.
• Sex is a categorical variable with two possible values: "Male" and "Female". In the model, it is treated as a dummy variable. Female has a value of 0 and male has a value of 1.
• Province is a categorical variable that consists ten possible values: Newfoundland and Labrador, Prince Edward Island, Nova Scotia, New Brunswick, Quebec, Ontario, Manitoba, Saskatchewan, Alberta, British Columbia, Nunavut, Northwest Territories, and Yukon Territories. In the model, they are treated as 12 seperate dummy variables, where 0 represents the province of Alberta and 1 represents each province in each dummy variable. Citizens from different provinces, such as Quebec and Alberta, tend to care about more province specific issues, such as Quebec independence and trans Mountain oil pipelines programs. These issues largely depend on the federal policies designed by the party in charge(BBC, 2019).
• Education is a categorical variable that consists five possible values: Less than high school diploma or its equivalent, High school diploma or a high school equivalency certificate, College, CEGEP or other non-university certificate, University certificate, diploma or degree above the bachelor's level, and Bachelor's degree (e.g. B.A., B.Sc., LL.B.). Education is also treated as 4 seperate dummy variables. 0 is corresponding to the degree of bachelor and 1 represents the rest four degrees accordingly. Voters with different education also tend to value policies differently. It is likely that people of higher education level may see different sides of the coin.

Model Setup

This is the full logistic regression model I am using:
Voting for Trudeau:

$$log(\frac{p_{JT}}{1-p_{JT}}) = \beta_{t0} + \beta_{t1}x_{age} + \beta_{t2}x_{male} + \beta_{t3}x_{BC} + \beta_{t4}x_{MB} + \beta_{t5}x_{NB} + \beta_{t6}x_{NL} + \beta_{t7}x_{NT} + \beta_{t8}x_{NS} + \beta_{t9}x_{NU}$$
$$+\beta_{t10}x_{ON} + \beta_{t11}x_{PE} + \beta_{t12}x_{QC} + \beta_{t13}x_{SK} + \beta_{t14}x_{YT} + \beta_{t15}x_{C} + \beta_{t16}x_{HS} + \beta_{t17}x_{LHS} + \beta_{t18}x_{U} + \epsilon$$

From the equations:
• p_{JT} represents the proportion of voters who will vote for Justin Trudeau.
• β_{t0} represents the intercept of the models, and is the probability odds of voting for Justin Trudeau at age 0, sex is female, province is Alberta, and education is Bachelor's degree.
• β_{t1} represents the slope of the model w.r.t. to age and x_{age} is the age of respondent. For everyone one unit increase in age, we expect a β_{t1} increase in the probability odds of voting for Justin Trudeau.
• β_{t2} is the slope of Male for Justin Trudeau's model and x_{male} is an indicator variable that represents the respondent who was indicated as male in column of sex. The beta shows the mean difference between the probability odds for the two models when the voter is a male and a female.
• β_{t3} to β_{t14} are the slopes of different provinces for Justin Trudeau's model. $x_{province}$ is an indicator variable showing whether the voter is in the province whose abbreviation is in the subscript. The beta shows the mean difference between the probability odds for the two models when the voter is in Alberta and other 12 regions in Canada.

• β_{t15} to β_{t18} are the slopes of different education for Justin Trudeau's model. $x_{education}$ is an indicator variable whose subscript corresponds to the education level. The beta shows the mean difference between the probability odds for the two models when the voter has a Bachelor's degree and other four other degrees. In this case, C = College, CEGEP or other non-university certificate, HS = High school diploma or a high school equivalency certificate, LHS = Less than high school diploma or its equivalent, and U = University certificate, diploma or degree above the bachelor's level.
• ϵ is the error term of the models.
The full model is comprehensive given the data that I obtained for this study. However, one drawback related to that is the occurrence of overfitting. As more variables fitted in the model, the R^2 will definitely increase although some of the variables may be useless.

Hence, I choose to run a backward elimination using AIC to factor out the variables that may not be useful and overfit the model. After running a backward elimination using AIC, the alternative model is:

$$log(\frac{p_{JT}}{1-p_{JT}}) = \beta_0 + \beta_3 x_{BC} + \beta_4 x_{MB} + \beta_5 x_{NB} + \beta_6 x_{NL} + \beta_7 x_{NT} + \beta_8 x_{NS} + \beta_9 x_{NU}$$
$$+\beta_{10} x_{ON} + \beta_{11} x_{PE} + \beta_{12} x_{QC} + \beta_{13} x_{SK} + \beta_{14} x_{YT} + \beta_{15} x_C + \beta_{16} x_{HS} + \beta_{17} x_{LHS} + \beta_{18} x_U + \epsilon$$

From the equations:
• p_{JT} represents the proportion of voters who will vote for Justin Trudeau.
• β_0 represents the intercept of the models, and is the probability odds of voting for Justin Trudeau at age 0, sex is female, province is Alberta, and education is Bachelor's degree.
• β_3 to β_{14} are the slopes of different provinces for Justin Trudeau's model. $x_{province}$ is an indicator variable showing whether the voter is in the province whose abbreviation is in the subscript. The beta shows the mean difference between the probability odds for the two models when the voter is in Alberta and other 12 regions in Canada.
• β_{15} to β_{18} are the slopes of different education for Justin Trudeau's model. $x_{education}$ is an indicator variable whose subscript corresponds to the education level. The beta shows the mean difference between the probability odds for the two models when the voter has a Bachelor's degree and other four other degrees. In this case, C = College, CEGEP or other non-university certificate, HS = High school diploma or a high school equivalency certificate, LHS = Less than high school diploma or its equivalent, and U = University certificate, diploma or degree above the bachelor's level.
• ϵ is the error term of the models.

Comparing to the full model, the alternative model should have a lower AIC, showing that the variables in the alternative model are more carefully selected.

Post-Stratification

In the above section, I created two models based on the survey data demographics and its stated voting choice. In this section, I will perform post-stratification analyses, which "incorporating population distributions of variables into survey estimates" (Little, 1993) for Trudeau's model to estimate the proportion of voters who will vote for him.
In the analyses, I will group similar units together in order to reduce the variance. The variables that I use to construct the groups are: age, sex, province, and education.
I will then weigh each proportion estimate (within each bin) by the respective population size of that bin and sum those values and divide that by the entire population size.

Results

Logistic Regression Model Results

Table 1: Full Logistic Regression Model of vote_Trudeau

term	estimate	std.error	statistic	p.value
(Intercept)	-0.9659997	0.1009686	-9.5673309	0.0000000
age	0.0017037	0.0012926	1.3180135	0.1874991
sexMale	-0.0418705	0.0463498	-0.9033582	0.3663358
provinceBritish Columbia	0.2140975	0.0885431	2.4180039	0.0156059
provinceManitoba	0.1652600	0.1001271	1.6505015	0.0988404
provinceNew Brunswick	0.2737261	0.1147667	2.3850654	0.0170761
provinceNewfoundland and Labrador	0.5472698	0.1216449	4.4989145	0.0000068
provinceNorthwest Territories	0.4162840	0.4740983	0.8780542	0.3799143
provinceNova Scotia	0.4944977	0.1071002	4.6171487	0.0000039
provinceNunavut	0.6193492	0.4619627	1.3406909	0.1800208
provinceOntario	0.3476912	0.0821185	4.2340169	0.0000230
provincePrince Edward Island	0.3772465	0.2054452	1.8362391	0.0663223
provinceQuebec	0.4081749	0.0842905	4.8424802	0.0000013
provinceSaskatchewan	-0.6815668	0.1261935	-5.4009654	0.0000001
provinceYukon	-0.5258979	0.4982673	-1.0554534	0.2912180
educationCollege, CEGEP or other non-university certificate	-0.1730347	0.0682653	-2.5347369	0.0112532
educationHigh school diploma or a high school equivalency certificate	-0.2072609	0.0643867	-3.2190035	0.0012864
educationLess than high school diploma or its equivalent	-0.4682827	0.0887872	-5.2742160	0.0000001
educationUniversity certificate, diploma or degree above the bachelor's level	0.0252217	0.0724512	0.3481199	0.7277501

• Based on Table 1, the intercept term is -0.9660, showing that if the voter is 0 years old, female, living in Alberta and has a bachelor's degree, then the probability odds of voting for Trudeau is -0.9660. The p-value is almost 0, meaning that the intercept term is significant in this model.

• The slope of age is 0.0017 which shows that keeping other variables constant, as age increases by 1 year, the mean probability odds of voting for Trudeau will increase by 0.0017 and people are more likely to vote for Trump. The slope of male is -0.0419 which suggests that keeping other variables constant, the mean difference of probability odds of voting for Trudeau between men and women are 0.0419 and males are less likely to vote for Trudeau. However, the p-value for age is 0.1875 and the p-value for sex is 0.3663, which shows that the two factors are not significant.

• For the slopes of provinces, 7 out of the 12 factors are significant. Significant terms include provinces British Columbia, New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario, Quebec and Saskatchewan. To interpret, for example, the slope for the term British Columbia is 0.2141, indicating that if the voter lives in British Columbia and other variables remain constant, then the mean of probability odds is higher by 0.2141 for the voter than another voter living in Alberta. Other provinces have the similar interpretation.

• For the slopes of education, 3 out of the 4 factors are significant, including College, CEGEP or other non-university certificate, High school diploma or a high school equivalency certificate, and Less than high school diploma or its equivalent. To interpret, for example, the slope for the term College, CEGEP or other non-university certificate is -0.1730, indicating that if the voter has a degree of College, CEGEP or other

non-university certificate and other variables remain constant, then the mean of probability odds is lower by 0.1730 for the voter than another voter who has a degress of bachelor. Other degrees have the similar interpretation.

Table 2: Alternative Logistic Regression Model of vote_Trudeau

term	estimate	std.error	statistic	p.value
(Intercept)	-0.9011713	0.0765953	-11.7653585	0.0000000
provinceBritish Columbia	0.2171363	0.0884744	2.4542281	0.0141187
provinceManitoba	0.1676516	0.1001059	1.6747429	0.0939847
provinceNew Brunswick	0.2754511	0.1147381	2.4006938	0.0163640
provinceNewfoundland and Labrador	0.5481777	0.1216092	4.5077001	0.0000066
provinceNorthwest Territories	0.4163847	0.4742110	0.8780578	0.3799124
provinceNova Scotia	0.4992590	0.1070345	4.6644683	0.0000031
provinceNunavut	0.6022836	0.4619960	1.3036555	0.1923511
provinceOntario	0.3497975	0.0820344	4.2640342	0.0000201
provincePrince Edward Island	0.3776642	0.2054116	1.8385734	0.0659780
provinceQuebec	0.4083120	0.0842391	4.8470613	0.0000013
provinceSaskatchewan	-0.6803602	0.1261661	-5.3925745	0.0000001
provinceYukon	-0.5216613	0.4982894	-1.0469043	0.2951437
educationCollege, CEGEP or other non-university certificate	-0.1702654	0.0682326	-2.4953666	0.0125827
educationHigh school diploma or a high school equivalency certificate	-0.2066737	0.0643732	-3.2105569	0.0013248
educationLess than high school diploma or its equivalent	-0.4696745	0.0887671	-5.2910865	0.0000001
educationUniversity certificate, diploma or degree above the bachelor's level	0.0266516	0.0723883	0.3681758	0.7127422

• Based on Table 2, the intercept term is -0.9012, showing that if the voter is 0 years old, female, living in Alberta and has a bachelor's degree, then the probability odds of voting for Trudeau is -0.9012. The p-value is almost 0, meaning that the intercept term is significant in this model.
• For the slopes of provinces, 7 out of the 12 factors are significant. Significant terms include provinces British Columbia, New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario, Quebec and Saskatchewan. To interpret, for example, the slope for the term British Columbia is 0.2171, indicating that if the voter lives in British Columbia and other variables remain constant, then the mean of probability odds is higher by 0.2171 for the voter than another voter living in Alberta. Other provinces have the similar interpretation.
• For the slopes of education, 3 out of the 4 factors are significant, including College, CEGEP or other non-university certificate, High school diploma or a high school equivalency certificate, and Less than high school diploma or its equivalent. To interpret, for example, the slope for the term College, CEGEP or other non-university certificate is -0.1703, indicating that if the voter has a degree of College, CEGEP or other non-university certificate and other variables remain constant, then the mean of probability odds is lower by 0.1703 for the voter than another voter who has a degress of bachelor. Other degrees have the similar interpretation.

Post-stratification result

Table 3: Post-stratification result of the probability of voting for Trudeau

	Probability of voting for Trudeau	Alternative Probability of voting for Trudeau
Probability	0.298384233473832	0.297687849265663

We applied the post-stratification method to determine the proportion of voters who prefer voting for Justin Trudeau and the Liberal Party using the full model and the altrnative model. It is modelled by the logistic model and the full model is using the respondents' age, sex, province and education as predictors while the alternative using province and education. According to this analysis, we estimate that the proportion of voters who are inclined to vote for the Trudeau using the full model is around 0.2984 while the proportion of voters who favoured the Trudeau using the alternative model is 0.2977.

Discussion

Summary of the study and Conclusion

In this study, my goal is to predict the 2019 Canadian election result based on the General social survey Family 2017 and 2019 Canadian Election Study - Online Survey. We build a logistic regression model based on the 2019 Canadian Election Study - Online Survey and use post-stratification on the General social survey Family 2017. The variables of my interest in this study include age, sex, province, and education.

From the full logistic regression model (with factors of age, sex, province, and education) and the alternative logistic regression model (with factors of province, and education), factors that significant are the same in both models, although the slopes of the same factors are slightly different in the two models. The factors are significant include provinces (British Columbia, New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario, Quebec and Saskatchewan) and education levels (College, CEGEP or other non-university certificate, High school diploma or a high school equivalency certificate, and Less than high school diploma or its equivalent). For example, the mean difference of voters who will vote for Trudeau between voters living in Alberta and Ontario captured by the full model is 0.3477 and 0.3498 by the alternative model. This means that if the two voters only differ in province or education, then both the full model and the alternative model may capture the mean difference in probability odds to the same extent.
From the significant variables, keeping age, sex, and education the same, we may find that voters from British Columbia, New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario, and Quebec tend to have a higher probability odds of voting for Trudeau compared to voters from Alberta. Keeping age, sex, and province the same, voters with College, CEGEP or other non-university certificate, High school diploma or a high school equivalency certificate, and Less than high school diploma or its equivalent all have a lower probability odds of voting for Trudeau.
From the probability that we predicted, the probability of voting for Trudeau are almost identical predicted by the full model and the alternative model, which is about 30%.

Discussion on the Model Results

I shall only talk about significant variables from the model.
From the previous part, it is concluded that keeping age, sex, and education the same, we may find that voters from Saskatchewan have a lower probability odds compared to voters from Alberta. This is probably because voters from Saskatchewan and Alberta want to seperate from Canada, resulting in a lower support rate from Alberta and Saskatchewan(Levinson-King, 2019).
Keeping age, sex, and province the same, voters with a bachelor's degree have a higher probability odds of voting for Trudeau. This may be due to the sights of the voters on policies. One policy that may make a difference include the Federal Carbon Tax. Voters with higher degrees tend to value more about the environment and maybe the carbon tax is more affordable for them. In contract, it is not possible that all voters with lower degrees do not care as much about the environment, but certainly to a lower extent. Imposing the carbon tax can be one of the reasons(Murphy, 2019).

Weakness and Next Steps

The study has some aspects that need to be improved in future studies.

First of all, the two datasets that I obtained are not designed specifically for this study. That is, the survey data is conducted and summarized in 2019 while the census data is generated in 2017. The census data from 2017 may not fully reflect the demographics in 2019. Meanwhile, the survey data is obtained through voluntary sample, where voters voluntarily take the survey. Sampling bias may exist in this process. Voters who respond to the survey tend to be the ones who care about politics and may not fully reflect the target population. In order to conduct a better study, it will be better if the two datasets that are used are from

the same time period. Moreover, the survey data should be obtained by stratified sampling instead in order to eliminate the sampling bias. Another thing that is worth doing is to try to get as much cases and survey results as possible such that the sample population may fully reflect the target population.

Secondly, while cleaning the data, I combined several categories of education into one category such that the survey data and census data may have corresponding education categories. There might be systematic errors occurred in the period because the corresponding education level of certain degrees, which are not commonly seen and not recorded in the code book, is ambiguous. These degrees, though not at a great portion, are eliminated from the dataset. In order to get a better dataset, the survey conducted to get the dataset (not necessarily phone call surveys) needs to be carefully designed. For example, the survey may provide strict options of the degrees level classification and explanation.

Thirdly, in the census dataset, there are no cases from the three territories while in the survey data there are. I did not remove these cases while fitting the logistics regression model. However, these variables are not useful in the prediction model because there are on cases from the three territories. In future studies, the frame population should be kept to be the same as well as the sampling method. This may help to get better predictions and more representative results.

In future studies, alternative explanatory variables other than age, sex, province, and education may also be taken to fit the model. The two datasets in this study actually do not have a lot of overlapping variables that may be fully utilized. Even the categories of four explanatory variables that I choose does not perfectly correspond to each other across datasets. By solving this problem, the data maybe more accurate to predictions. Some other interesting variables may include income levels, race/ethnicity, and so on.

In addition, one thing worth mentioning here is that in the survey data, vote choices have responses such as the Green Party, the Conservative Party. These are the options that may cause a small spread between the probabilities of getting votes for different parties. Hence, even though the probability of voting for Trudeau predicted is not high in this study, other parties may even do not get as many votes as Trudeau does. But probability of voting for other parties is beyond the scope of this study and may be conducted using the same methods in this study.

References

1. Alexander, R., & Caetano, S. (2020, October 7). Gss_cleaning.R. Retrieved December 8, 2020.
2. BBC. (2019, October 22). Canada election: Trudeau's Liberals win but lose majority. BBC. Retrieved December 8, 2020. https://www.bbc.com/news/world-us-canada-50134640 Helmenstine, A. M. (2020, July 1). Random Error vs. Systematic Error. ThoughtCo.. https://www.thoughtco.com/random-vs-systematic-error-4175358
3. Levinson-King, R. (2019, October 11). Wexit: Why some Albertans want to separate from Canada. BBC. https://www.bbc.com/news/world-us-canada-49899113 Little, R. J. (1993). Post-Stratification: A Modeler's Perspective Abstract. Journal of the American Statistical Association, 88(423), 1001-1012. doi:10.1080/01621459.1993.10476368
4. Murphy, J. (2019, October 22). Justin Trudeau: The good news - and bad - for Canada's PM. BBC. https://www.bbc.com/news/world-us-canada-50130391
5. N/A. (2017). 2017 General Social Survey: Families Cycle 31 Public Use Microdata File PUMF (Rep.). Retrieved October 19, 2020, from Statistics Canada website: https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_Codebook.pdf
6. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
7. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
8. Statistics Canada. (2017). General social survey on Family (cycle 31), 2017 - Canadian general social surveys (GSS). Retrieved December 8, 2020, from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm
9. Statistics Canada. (2020, April). General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User???s Guide (Rep.). Retrieved 2020, from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf
10. Statistics Canada. (n.d.). Table 8 Abbreviations and codes for provinces and territories, 2011 Census. Statistics Canada. https://www150.statcan.gc.ca/n1/pub/92-195-x/2011001/geo/prov/tbl/tbl8-eng.htm
11. 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
12. Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, "2019 Canadian Election Study - Online Survey Technical Report and Codebook.pdf", 2019 Canadian Election Study - Online Survey, https://doi.org/10.7910/DVN/DUS88V/HRZ21G, Harvard Dataverse, V
13. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Appendix I: GitHub Repo Link

https://github.com/Yao-He3/Canadian-Election-2019