Analyzing the effects of full voting rate on 2019 Canadian Federal Election – The Conservatives Will Win

WingYan Lee 1004840381

21/12/2020

Code and data supporting this analysis is available at:https://github.com/YvonnewyLee/STA304FinalProject

Abstract

In this study, I aimed to investigate the effects of full turnout rate on the 2019 Canadian Federal Election. To achieve this goal, I built a logistic regression model based on the 2019 Canadian Election Study – Campaign-Period Survey(CPS) data and employed post-stratification using Canada's General Social Survey(GSS) to predict the outcome of 2019 election if all eligible Canadians have voted. After comparing to the actual results, I ended up with a conclusion that the Conservative Party will still win but the Liberal Party will have less share of the overall popular vote. This study is beneficial for campaign strategy analysts to develop appropriate approaches to increase the overall popular vote of their parties and for the general public to realize the potential effect of their voting choices.

Keywords

Key words: Multi-level regression and Post-stratification(MRP), Logistic Regression, 2019 Canadian Federal Election, Canada's General Social Survey(GSS), 2019 Canadian Election Study – Campaign-Period Survey(CPS)

Introduction

Statistical analysis has been applied widely in political science. It is useful in analyzing voter attitudes for campaign messaging, predicting the most effective marketing strategy and particularly, forecasting election results. In this case, census data is powerful since it provides detailed and comprehensive information of all potential voters. Thus, using census data to predict election results is the key to getting more accurate prediction.

One important technique to forecast election results is through multi-level regression and post-stratification(MRP). This technique was originally developed by Gelman and T. Little in 1997 and was purposed for use in estimating US-state-level voter preference by Lax and Philips in 2009 [1]. In this report, I will use logistic regression model with post-stratification on census data to predict the 2019 Canadian Federal Election outcome if everyone had voted and compare it to the actual election results to analyze the importance of turnout rate.

The 2019 Canadian Federal Election, held on October 21, 2019, aimed to elect members of the 43rd Canadian Parliament. The Liberal Party, led by current Prime Minister Justin Trudeau, lost the overall popular vote to the Conservatives [15] but won enough seats to form a minority government. According to

Global News, only 65.95% of eligible Canadian voters cast a ballot, which is a 2.35% drop from the turnout in 2015 [14]. One might wonder what the results will be if everyone has voted. Hence, it would be interesting to forecast the election outcome assuming that everyone has voted and compare it with the actual results. This analysis will be beneficial for the general public to realize the potential effects of their voting choices and for campaign strategy analysts to develop approaches to increase the overall popular vote of their parties.

Two data sets, 2019 Canadian Election Study – Campaign-Period Survey(CPS) and Canada's General Social Survey(GSS), will be used to investigate how logistic regression models and post-stratification can be applied to predict how different the 2019 Canadian Federal Election outcome would have been if all eligible Canadians had voted. CPS will be used to build logistic regression model while GSS will be used as the census data for post-stratification. In the Methodology section, I describe the data and model used to perform post-stratification technique. Results of the analysis are provided in the Results section. The inference of the results, the potential weakness and future steps of this study are stated in the Discussion section.

Data

Two data sets are used in this analysis. One is the 2019 Canadian Election Study – Campaign-Period Survey(CPS) [2] which is used as the survey data for modelling. The second one is the Canada's General Social Survey(GSS) [3] which is used as the census data for post-stratification.

The 2019 Canadian Election Study – Campaign-Period Survey (CPS) containing 4021 observations of 278 variables is obtained by using cesR package in R studio [4]. This data set contains information about attitudes and opinions of Canadians during the 2019 federal election. In this data set, data is collected by telephone interviews through either wireless telephone or landline telephone. All telephone data collection is completed with Computer Assisted Telephone Interviewing (CATI)[5]. Wireless telephone respondents are found by randomly chosen from a list of randomly generated telephone numbers generated by Advanis while landline telephone respondents are found by randomly chosen from ASDE, a sample provider[5]. For the initial non-responses, researchers will try to call at a minimum of 6 attempts. If no eligible respondent is reached, researchers will randomly select a new respondent to call.

The target population of CPS is Canadian citizens 18 years of age or older who reside in one of the ten Canadian provinces (excluding the territories)[5]. In addition, the frame population is the eligible Canadians who has a telephone or a cell-phone. The sample population is all eligible Canadians who were willing to take the survey. And the sample is 4021 eligible Canadians selected. Furthermore, the sampling frame are the list of landline numbers from ASDE and the list of wireless phone numbers generated internally by Advanis [5].

A key feature of CPS is that it employs the dual-frame-with-overlap approach to determine the overlap of people who have both a wireless and a landline phone and to correct for the higher selection probability of the overlap group. The strength of CPS data set is its large number of variables which provides many interesting factors to investigate. But a disadvantage of CPS is its small sample size, which is due to difficulty in collecting data. Moreover, there is no variable in CPS data set that records the respondent's final voting choice. In the analysis, the variable which describes the party the respondent wants to vote for in CPS is used as the respondent's final vote choice. However, the respondent could make a different decision, so the analysis could be inaccurate. In addition, neither the phone survey nor the web survey will be a true census; thus, using either one of them is not representative enough as it does not record everyone's voting choice.

The Canada's General Social Survey(GSS) which contains 20602 observations of 81 variables is obtained from University of Toronto online library. GSS contains information about the living conditions and well-being of Canadians. In this data set, researchers collect data by making phone calls through the same CATI techniques used in CPS data collection and interviewing a random qualified member of the household [6]. Thus, respondents are found by randomly contact through phone calls. However, for the non-responses, researchers will try to finish the survey at first. If they still fail, the weights of responses are then adjusted for the overall representativeness.

For GSS, the target population is all non-institutionalized people who are 15 years old or older and are living in private households in the 10 provinces of Canada. Additionally, the frame population is all eligible Canadians with a telephone or a cell-phone. Furthermore, the sampling frames are the lists of telephone numbers in use (landline and cellular) available to Statistics Canada from various sources and the list of all dwellings within the 10 provinces. Besides, the sampled population is all eligible Canadians who are willing to take a questionnaire from a phone/telephone call [6]. And the sample is 20602 eligible Canadians selected.

A key feature of GSS is that researchers use Stratified Random Sampling on surveyed population with strata divided by different areas and conduct simple random sampling within each stratum without replacement[6]. In terms of advantages, GSS has a large data size with various attributes, which is beneficial for potential investigation. However, a major drawback is that GSS is used as census data in the analysis while it does not contain information of all Canadians. Hence, the respondents in this data set is not representative enough of the voting population despite of the large sample size. Another weakness is that the GSS data set is collected in 2017 while the goal of the analysis is to predict outcome of 2019 Canadian Election. Therefore, the outdated GSS data could potentially influence accuracy of the analysis.

After importing the data sets, I perform data cleaning to both CPS and GSS and extracte the variables of interest using tidyverse package in R studio [7] as well as cleaning code posted by Rohan Alexander and Sam Caetano [8]. For GSS data set, only variable sex, age, province, education, birth country and income are kept since they are the ones I am interested in. For CPS, I choose the same set of variables with one addition – voting choice, as the response variable which is required for modelling.

In CPS data set, there are two similar variables income and income range with income being a specific amount and income range being a distinctive level. I decide to use income and create a new income range variable based on income because the division of levels of the existing income range variable is different from the one in GSS, which could cause problems in modelling. Furthermore, in order to build six logistic models for each of the six parties in the Model section, I construct six new variables to indicate respondents' voting choice based on two variables in CPS, one of which states the voting choice for determined respondents and the other one of which describes the most likely voting choice for undecided voters.

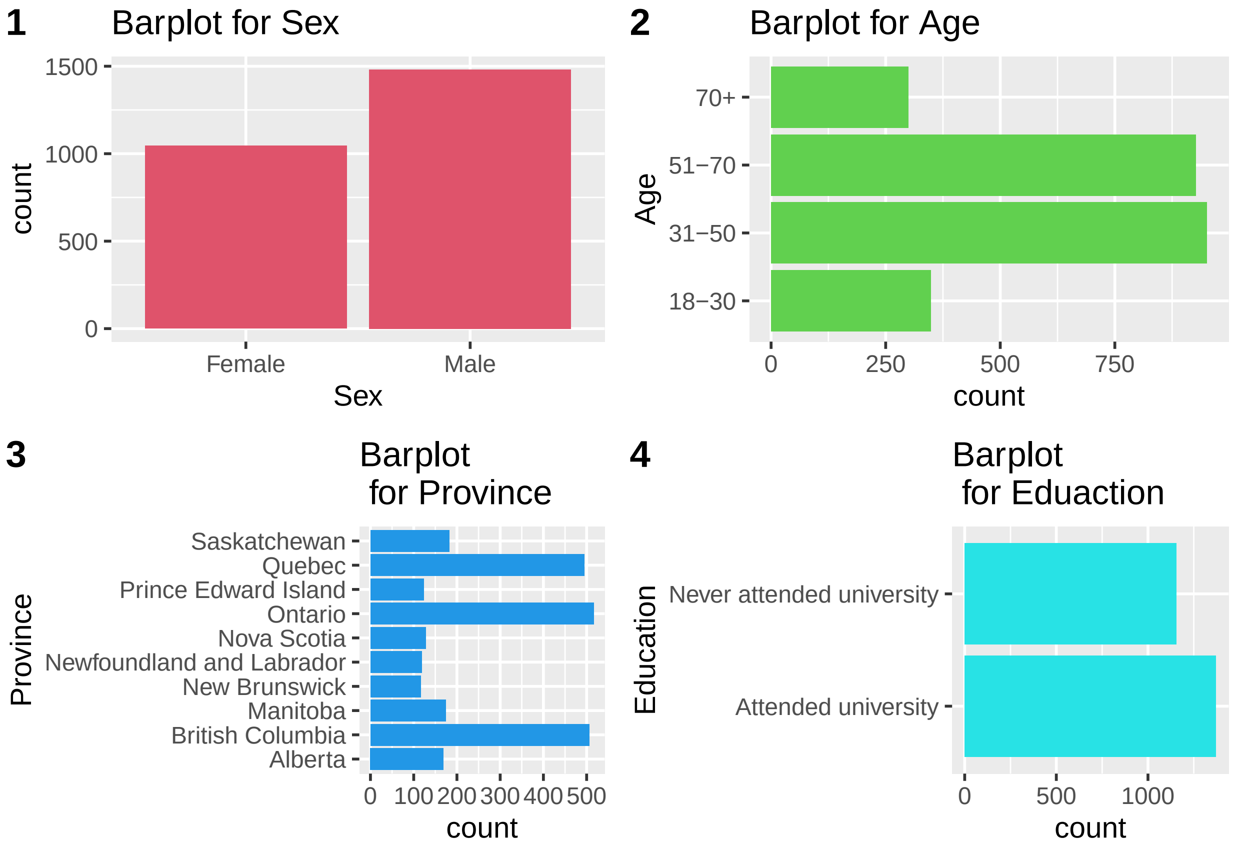

Figure 1: Plots for survey data

Looking at the above plots in Figure 1, 3-5, we can gain some insights about these two datasets. Plot 1-7 illustrate the distribution of data for each of the 6 variables of interest plus a response variable in the survey data (CPS). Meanwhile, plot 8 – 13 describe the data distribution for all 6 variables of interest in the census data (GSS). Note that plot 5-13 are included in the Appendix section (1). By plot 1-7, it can be seen that most variables in CPS are not evenly distributed. For example, the respondents in survey data have a higher male ratio and a much higher ratio of born in Canada. Additionally, most respondents live in Ontario, Quebec and British Columbia and they are in the age groups of 31-50 and 51-70. This similar trend can be observed in the census data as well. For example, most respondents in census data are born in Canada and live in Ontario, Quebec and British Columbia. As for the response variable in survey data, most people vote for the Liberal and the Conservatives with the Conservatives having slightly more votes than the Liberal.

Method

In this study, I aim to investigate the difference between the actual 2019 Canadian Federal Election popular vote outcome and my prediction if all eligible Canadians have voted. Note that eligible Canadians are defined as 18 years old or older Canadian citizens who have the right to vote in this analysis.

In order to compare the difference, I build six logistic regression models each of which predicts the probability that one of the six major parties will win the popular vote in 2019 Canadian Federal Election. Then I compare the six probability and the highest one will indicate the predicted winner party in the 2019 election. Throughout the analysis, I start with the same logistic regression model using the survey data [2], employ the same kind of technique in finding the best model and perform post-stratification using the census data [3] to all six models to obtain the results.

Model Specifics

For each of the six major parties (Liberal Party of Canada, Conservative Party of Canada, Bloc Québécois, New Democratic Party, Green Party of Canada and People's Party) [9], the outcome of the model should be whether this party will win the popular vote of the 2019 election or not. Since the response is a binary variable, a frequentist logistic regression model is suitable to model the proportion of voters who will vote for this party.

To enable post-stratification afterwards, the variables of interest in survey data and census data should match. After careful comparison, only variable age, sex, education, province, birth place, income, importance of religion and marital status exist in both data sets. However, marital status is not used because this variable comes from the post-election survey in survey data. Not using marital status means that the method and analysis I describe here could have been used to predict real-time predictions during the 2019 election [10]. In addition, importance of religion is not kept in the model since there are many missing values in this column, which would significantly decrease the number of sample size after cleaning. Thus, only the remaining 6 variables are considered for the initial full model.

In order to make the model simpler, cleaner and more concise, observations with missing values are removed and some categorical variables are regrouped into fewer but more representative groups. For example, education, birth place and income group originally have multiple levels. They are now grouped into fewer levels to be more representative and more distinguishable. Furthermore, I create age groups instead of using age as a numerical variable to allow for splitting cells based on age groups in the post-stratification step. As for the response variable in survey data, I use the variable which indicates the respondent's most likely voting choice as the predictor of his final voting choice and construct six new variables for each of the six logistic regression models. All data cleaning and wrangling steps are accomplished in RStudio[4, 7, 8].

To account for the condition that all eligible Canadians have voted, all respondents who are either less than 18 years old or do not have Canadian citizenship are removed from the survey data. Additionally, to ensure that "everyone" has voted, I keep undecided voters and use their most likely voting choice as their final voting choice. Also, respondents who decide to spoil the ballot are not removed. According to CBC, spoiled ballots are rejected and do not have strong political consequence[11]. Hence, if everyone has to vote, one still has the option of spoiling the ballot if he doesn't want to vote for any party and spoiling the ballot should also count as a form of voting. However, those observations will indicate false for all six response variables since they don't vote for any party.

For the model selection process, the initial full model starts off with 6 variables of interest as mentioned above. Then, stepwise selection is used according to both AIC and BIC criterion. Both criterions penalizes models having a large number of predictor variables with BIC enforcing stricter penalties. After that, I compare the two models obtained and choose the one with smaller AIC. If both have similar AIC, then the model with more appropriate number of predictor variables is preferred.

For example, when building the model for Conservative Party, the model obtained using AIC includes 6 predictor variables while the one by BIC only includes 3 predictors. Their corresponding AIC values are

2888.2 and 2907.4 respectively, which is very close. In the end, I select the model using BIC criterion and save the more complex one as an alternative model. This is because a model with fewer predictors is easier to describe and understand and also reduces the difficulty for data collection process if we want to use it for prediction. On the other hand, the alternative model is useful when we want to make more accurate predictions with more predictors and more adequate information on collected data. But the alternative model could be too complex for understanding and may suffer from overfitting.

The final logistic regression model for Conservative Party is :

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{BritishColumbia} + \beta_2 x_{Manitoba} + \beta_3 x_{NewBrunswick} + \beta_4 x_{NewfoundlandandLabrador} + \beta_5 x_{NovaScotia} \\ + \beta_6 x_{Ontario} + \beta_7 x_{PrinceEdwardIsland} + \beta_8 x_{Quebec} + \beta_9 x_{Saskatchewan} + \beta_{10} x_{NeverAttendedUniversity} + \beta_{11} x_{Male} \tag{1}$$

where $log(\frac{p}{1-p})$ represents the log odds of the proportion of voters who will vote for the Conservative. In other words, it is the log ratio of probability of voting for Conservative to probability of not voting for Conservative. Moreover, β_0 is the intercept of the model which represents the log ratio of probability of voting for Conservative to probability of not voting for Conservative if the voter is born in Canada, has attended university and lives in Alberta. Additionally, β_1 represents how much the log ratio of probability we expect to change for a voter who lives in British Columbia ($x_{BritishColumbia} = 1$) compared to a voter lives in other provinces ($x_{BritishColumbia} = 0$), keeping all other variables fixed. The other coefficients have similar meanings as the first two described above. Note that all coefficients in equation (1) are dummy variables except for β_0. All of them indicate whether or not the voter belongs to a group with 1 meaning the voter do belong to the group and 0 meaning the opposite.

The other 5 final logistic regression models in equation (4-8) for other parties are listed in the Appendix section (2).

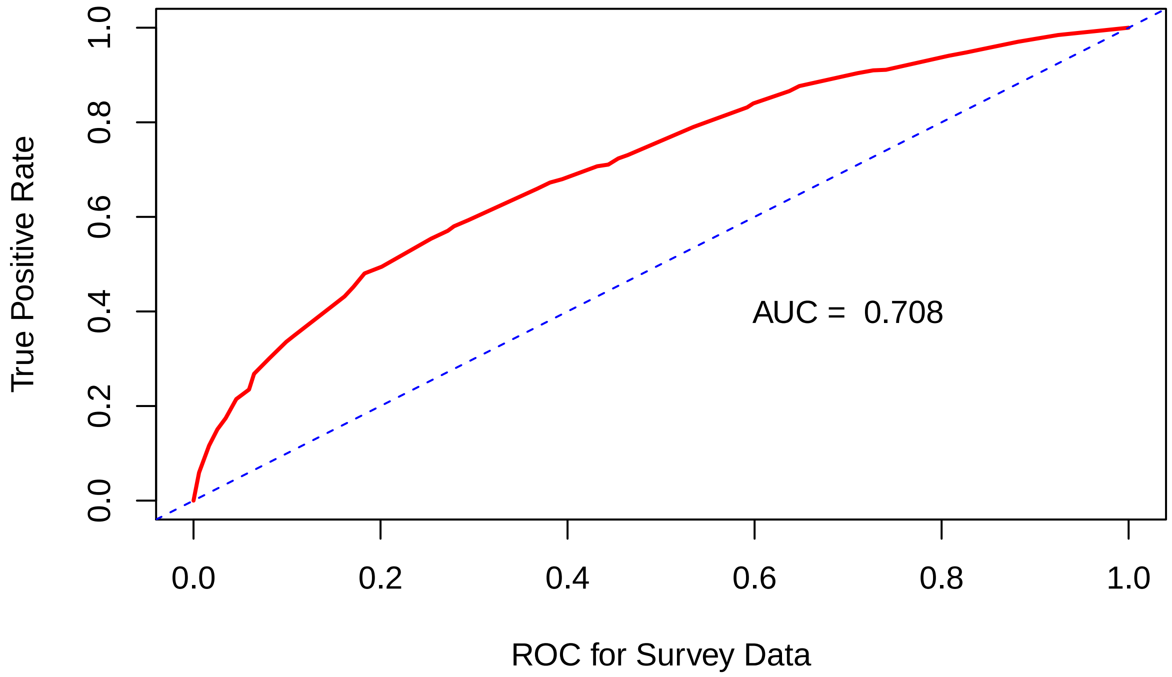

Figure 2: ROC Curve for Conservative Party

Last but not least, the performance of final model is validated by calculating the AUC value in the Receiver Operating Characteristic (ROC) curve using the pROC package in RStudio [12]. The value is expected to be far from 0.5 and closer to 1. By the ROC plot in Figure 2, the AUC value is 0.708, which shows a decent discrimination ability of the final model for Liberal Party. A similar procedure is performed for the other five final models for other parties, which could be verified by Figure 6-10 in the Appendix section (3).

Post-stratification

In order to estimate the proportion of voters who will vote for a certain party, a post-stratification analysis is employed. Firstly, I divide the census data into 811 demographic cells based on their different features - sex, age group, province, education, birth place and income group. For each demographic cell, the proportion of voters is estimated. Then each proportion estimate within each bin is weighted by the respective population size of that bin. Finally, I sum those values and divided that by the entire population size to estimate the population-level proportion of people voting for a certain party.

The formula I used to calculate the post-stratified proportion is :

$$\hat{y}^{PS} = \frac{\sum_j N_j \hat{y}_j}{\sum_j N_j} \tag{2}$$

where $\hat{y}_j$ is the estimate of proportion of voters for this party, for example Conservative Party, in the j^{th} cell, N_j is the population size of the j^{th} cell based off demographics (from census data).

The previous 6 variables of interest are all included in the cell division process because they are all likely to influence the voting outcome from a practical perspective. First of all, age has been observed as an important factor influencing political behaviors and study shows that older people tend to be more certain to vote [13]. Secondly, province is also influential since different provinces often have different economic conditions and political preferences, which could potentially affect the voters' decisions. According to the ACS - Environics survey, Canadians with lower incomes and less education express less certainty about their intention to vote in the federal election [13]. Thus, income and education are also important factors to be considered. Finally, sex and birth place influence people's self-identity, values and attitudes towards the relevant policies which concern their own benefits.

The post-stratification analysis is beneficial in this case since it can reduce the bias of representativeness in the population by accounting more for non-major groups when we use non-probability based sampling. With this technique, those underrepresented groups would contribute to the final result uniquely. Moreover, it could result in smaller variances in our predictions by capturing the inner patterns. As mentioned earlier in the Data section that the census data is not representative enough despite of its large sample size, without this technique, it would be difficult to generate accurate forecasts in this analysis.

Results

From the previous section, I have derived a frequentist logistic regression model by equation (1, 4-8) to predict the overall popular vote for each party of the 2019 Canadian Federal Election. Then a post-stratification analysis is performed on each model to estimate the proportion of voters in favor of each party.The below Table 1 lists out all significant predictor variables and its corresponding coefficient values of the model for Conservative Party.

Table 1: Coefficients Table

coefficient	predictor variable	value
β_0	intercept	0.01168
β_1	$x_{BritishColumbia}$	-1.68289
β_2	$x_{Manitoba}$	-0.76370
β_3	$x_{NewBrunswick}$	-1.16562
β_4	$x_{NewfoundlandandLabrador}$	-1.59398
β_5	$x_{NovaScotia}$	-1.66979
β_6	$x_{Ontario}$	-1.53175
β_7	$x_{PrinceEdwardIsland}$	-1.57802
β_8	x_{Quebec}	-2.28472
β_9	$x_{Saskatchewan}$	-0.29713
β_{10}	$x_{NeverAttendedUniversity}$	0.59325
β_{11}	x_{Male}	0.67794

In the above Table 1, β_0 represents that if the voter is born in Canada, has attended university and lives in Alberta, then the log ratio of probability of voting for Conservative to probability of not voting for Conservative is 0.01168. And β_1 represents that the log ratio of probability is expected to decrease by -1.68289 for a voter who lives in British Columbia ($x_{BritishColumbia} = 1$) compared to a voter lives in other provinces ($x_{BritishColumbia} = 0$), keeping all other variables fixed. Similarly, β_2 represents the expected decrease of -0.76370 in log ratio of probability for a voter who lives in Manitoba compared to a voter lives in other provinces, keeping all other variables fixed. A similar interpretation can also be made with β_3, β_4, β_5, β_6, β_7, β_8 and β_9. For β_{10}, it represents the expected increase of 0.59325 in log ratio of probability of voting for Conservative to probability of not voting for Conservation for a voter who has never attended university ($x_{NeverAttendedUniversity} = 1$) compared to someone who has attended university ($x_{NeverAttendedUniversity} = 0$). Lastly, β_{11} represents that for a male voter ($x_{Male} = 1$), we expect the log ratio of probability to increase by 0.67794 compared a female voter ($x_{Female} = 1$).

Combined with the above coefficients, the final model for Conservative Party in equation (1) can be re-expressed as the following. Using the same approach, we can obtain the final models for other parties, which are listed in the Appendix section (4).

$$\begin{aligned} log(\frac{p}{1-p}) = 0.01168 - 1.68289x_{BritishColumbia} - 0.76370x_{Manitoba} - 1.16562x_{NewBrunswick} \\ -1.59398x_{NewfoundlandandLabrador} - 1.66979x_{NovaScotia} - 1.53175x_{Ontario} + -1.57802x_{PrinceEdwardIsland} \\ -2.28472x_{Quebec} - 0.29713x_{Saskatchewan} + 0.59325x_{NeverAttendedUniversity} + 0.67794x_{Male} \end{aligned} \tag{3}$$

Based off the post-stratification analysis of the proportion of voters in favor of Conservative Party modelled by a logistic regression model which accounted for province, sex and education, I estimate that the proportion of voters in favor of voting for Conservative Party of Canada to be 0.3457622. Similarly, the proportion of voters in favor of voting for other parties can be estimated and the results are listed in the Table 2 below. In Table 2, it can be seen that the estimation of proportion of voters in favor of voting for Liberal Party of Canada to be 0.329648, which is slightly less than the proportion for Conservative Party.

Table 2: Results Table

Political Party	Proportion of voters in favor of this party
Liberal Party of Canada	0.3296480
Conservative Party of Canada	0.3457622
New Democratic Party	0.1447250
Bloc Québécois	0.0452041
Green Party of Canada	0.1015530
People's Party	0.0150300

Discussion

Summary

In this study, I access a survey data (CPS) and a census data (GSS) to develop six logistic regression models. The survey data I used is collected by random phone interviews with eligible Canadian voters from different provinces of Canada about their opinions towards the 2019 Canadian Federal Election [2] while the census data is collected using the similar approach but including non-eligible-to-vote Canadians [3].

I first standardized variable categories in both data sets and removed the voters who refuse to answer in the survey data set to avoid biased results, then establish a frequentist logistic regression model using the survey data. After that, the model is validated using AUC values. However, the survey data is non-representative since it will never be a true census. Besides, the census data is collected in 2017 and it does not truly contains demographic information of all Canadians. Thus, a post-stratification analysis is performed using the census data to reduce the bias of representativeness in the population and to get an accurate prediction of the overall popular vote for each party if all eligible Canadians have voted.

Conclusion

In conclusion, my analysis suggests that Conservative Party of Canada will win the most popular vote in 2019 Canadian Federal Election if everyone has voted. By looking at Table 3, since Conservative Party of Canada has the highest proportion of voters in favor of voting (34.58%) which is slightly more than the Liberal's, it is likely to win the overall popular vote if all eligible voters have voted. Note that voters who "don't know" which party to vote are counted as spoiling the ballot in the analysis. Hence, it is possible that they would their decision and the actual outcome could change.

Table 3: Results Table

Political Party	Proportion of voters in favor of this party
Liberal Party of Canada	0.3296480
Conservative Party of Canada	0.3457622
New Democratic Party	0.1447250
Bloc Québécois	0.0452041
Green Party of Canada	0.1015530
People's Party	0.0150300

In comparison, the actual outcome of the 2019 Canadian Federal Election with a 65.95% turnout rate [14] is that Conservative Party also won the popular vote with a 34.34% vote while the Liberal Party has a 33.12% popular vote [15] which is bigger than my prediction. Meanwhile, the actual results for New Democratic Party (15.98%) and People's Party (1.62%) is similar to the prediction. However, if everyone has voted, the proportion of votes supporting the Green Party of Canada is likely to increase while the overall popular vote for Bloc Québécois will decrease.

To be concluded, under the assumption of full turnout rate, the Conservative Party is still likely to win but the Liberal Party will have a smaller share of popular vote. Besides, different parties will experience different changes in popular vote. Therefore, if a party's popular vote increases as turnout rate increases, it needs to encourage voters to vote in order to have more popular vote.

Weakness

There are several drawbacks that needs to be aware of in this analysis: bias from data sets, inappropriate way of data cleaning and inaccuracy of prediction.

The first one is the bias resulted from my choices of data sets. I choose the phone survey data instead of web survey data as my survey data; however, none of these will be a true census data. Hence, the model built using survey data cannot make very accurate predictions. Besides, combining the phone survey data and the web survey data is not a good idea either. Complicated coding may be required to achieve it and it is challenging to deal with the overlapping issues. Additionally, the census data used for post-stratification is outdated since it is collected in 2017 and used to make predictions of 2019 election. Although age has been adjusted in the analysis to allow voters who are eligible to vote in 2019 but not eligible in 2017, other factors including education and income may have changed dramatically over the course of two years, which could potentially affect one's political opinions.

Another weakness arises when we need to map gender in survey data to sex in census data. I choose to remove respondents who do not identify as either male or female from the sample, which is a common and easy approach. However, as described in "Using sex and gender in survey adjustment", this method would mean that the responses of non-binary individuals are not counted in the analysis where the goal is make population generalizations and predict election outcome [16]. Moreover, this structural exclusion is a form of discrimination against those respondents who have different identities. Luckily, the census data only contains one of such observation and removing it does not significantly affect the prediction. But it is still an unfair approach.

Last but not least is the inaccuracy of prediction due to several factors. One of them is removing voters who "refused" to answer the survey. For simplicity, I choose to remove the voters who select "refused" or "don't know" when they are asked about which party they want to vote for. Such approach ignores that those voters are still possible to vote, which could change the actual outcome. Another factor is that I use respondents' most likely voting choice as their final voting choice. However, it is still possible that they change their decisions at the actual election. Thus, the prediction could be inaccurate.

Next Step

For the next step, I plan to look for better alternative census data that is closer to 2019. Then I will redo the post-stratification using the new census data to check for improvement in the model. Besides, a multilevel regression model could be considered to further stress the difference between groups. A follow-up analysis can be conducted to see if there is any improvement in prediction. Furthermore, more research can be done to investigate how turnout rate influences the election and learn from the previous elections. In this way, I can refine the model and post-stratification to have more precise predictions.

References

1. Multilevel regression with post-stratification (2020, 12 22). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Multilevel_regression_with_poststratification
2. Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, "2019 Canadian Election Study - Phone Survey", https://doi.org/10.7910/DVN/8RHLG1, Harvard Dataverse, V1
3. CHASS Data Centre, Faculty of Art and Science of University of Toronto, 2020, http://dc.chass.utoronto.ca/myaccess.html
4. Paul A. Hodgetts and Rohan Alexander (2020). cesR: Access the CES Datasets a Little Easier.. R package version 0.1.0.
5. Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, "2019 Canadian Election Study - Phone Survey Technical Report.pdf", 2019 Canadian Election Study - Phone Survey, https://doi.org/10.7910/DVN/8RHLG1/1PBGR3, Harvard Dataverse, V1
6. Statistics Canada, 2020. "General Social Survey" https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf
7. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
8. Rohan Alexander, Sam Caetano (2020, 12 22). https://q.utoronto.ca/courses/184060/files/9422740?module_item_id=1867317
9. Wikipedia. 2020. "List of federal political parties in Canada" https://en.wikipedia.org/wiki/List_of_federal_political_parties_in_Canada
10. Wei Wang, D. R. (2020, 12 22). Forecasting elections with non-representative polls. Retrieved from ScienceDirect: https://www.sciencedirect.com/science/article/abs/pii/S0169207014000879
11. Mullin, M. (2020, 12 22). Here's what happens when you spoil a ballot. Retrieved from CBC: https://www.cbc.ca/news/canada/newfoundland-labrador/spoiled-ballots-1.5136461
12. Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77 http://www.biomedcentral.com/1471-2105/12/77/
13. Jedwab, J. (2020, 12 22). Intention and impact: Canadians reflect on their votes. Retrieved from Policy Options: https://policyoptions.irpp.org/magazines/north-american-integration/intention-and-impact-canadians-reflect-on-their-votes/
14. Britneff, B. (2020, 12 22). Canada election: The 2019 results by the numbers. Retrieved from Global News: https://globalnews.ca/news/6066524/canada-election-the-2019-results-by-the-numbers/
15. Wikipedia. (2020, 12 22). Retrieved from 2019 Canadian federal election: https://en.wikipedia.org/wiki/2019_Canadian_federal_election
16. Lauren Kennedy, K. K. (2020, 12 22). Cornell University. Retrieved from https://arxiv.org/abs/2009.14401

Appendix

1. Plot 5-13 for Data section

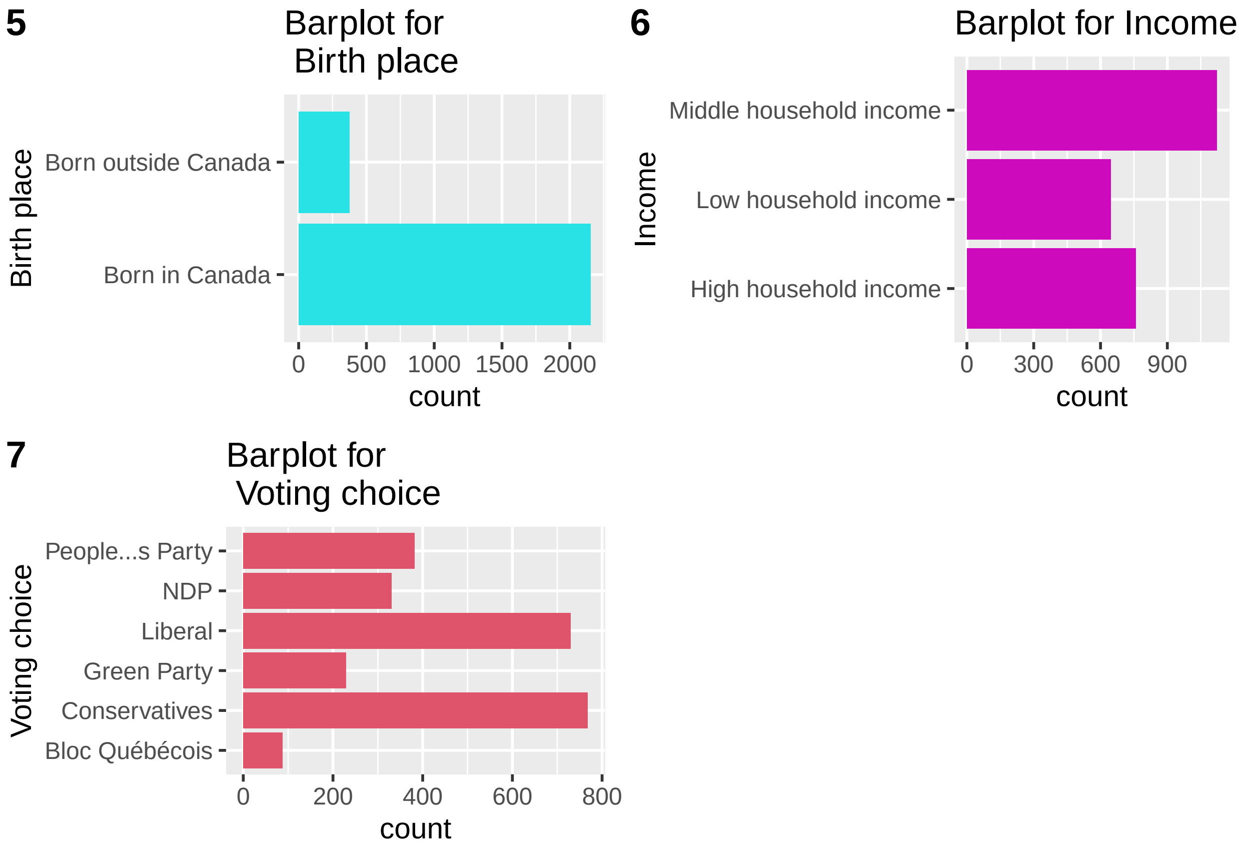

Figure 3: Plot5-7 for survey data

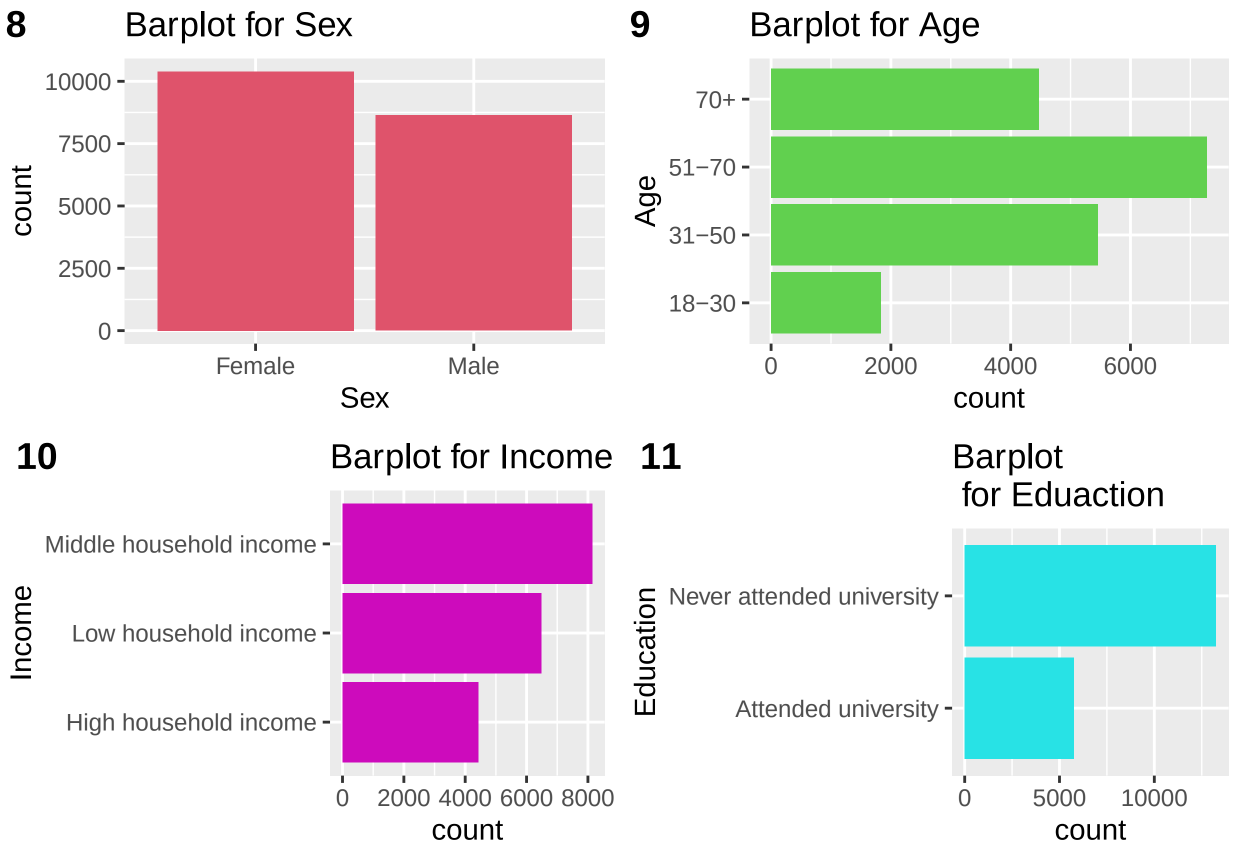

Figure 4: Plot8-11 for census data

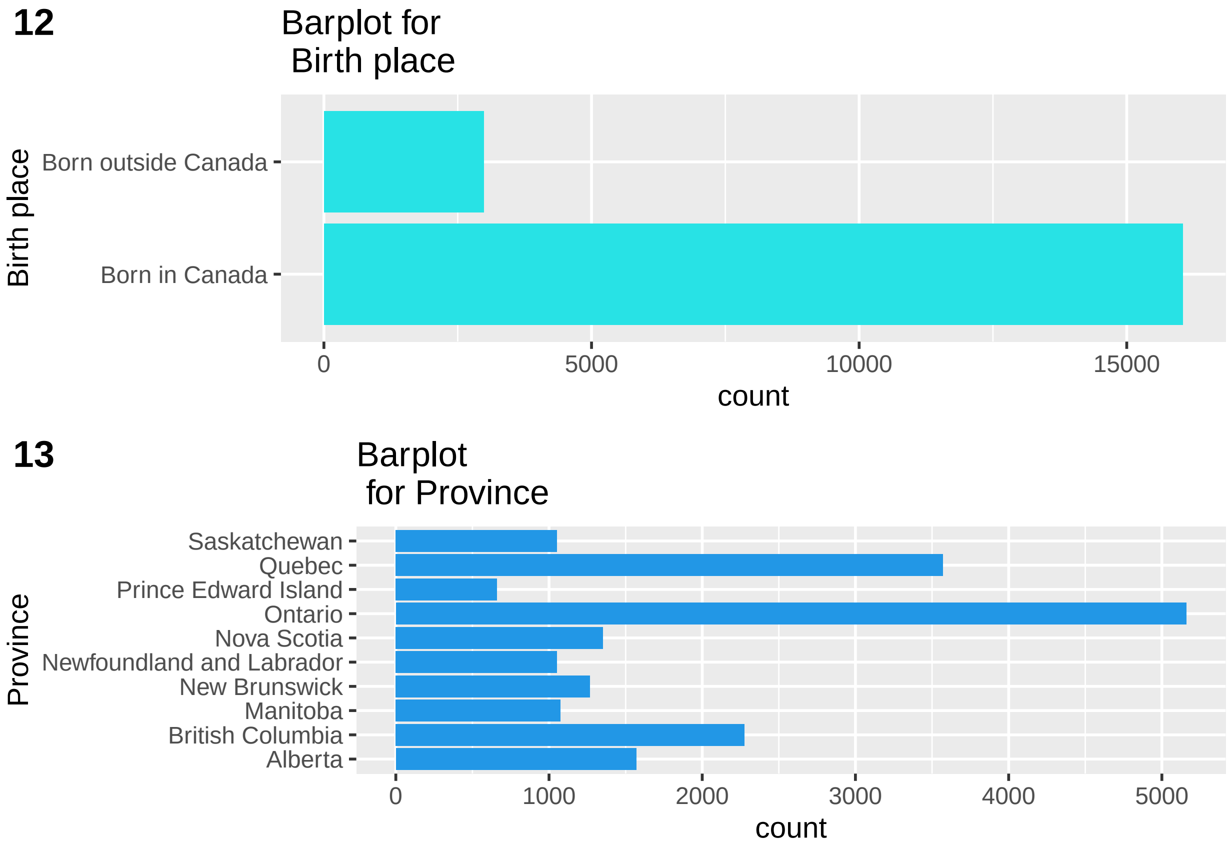

Figure 5: Plot12-13 for census data

2. 5 other Final models for Model section

The final logistic regression model for Liberal Party is :

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{BritishColumbia} + \beta_2 x_{Manitoba} + \beta_3 x_{NewBrunswick} + \beta_4 x_{NewfoundlandandLabrador} + \beta_5 x_{NovaScotia} + \beta_6 x_{Ontario} + \beta_7 x_{PrinceEdwardIsland} + \beta_8 x_{Quebec} + \beta_9 x_{Saskatchewan} + \beta_{10} x_{NeverAttendedUniversity} + \beta_{11} x_{BornoutsideCanada} \tag{4}$$

The final logistic regression model for NDP Party is :

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{Male} + \beta_2 x_{31-50} + \beta_3 x_{51-70} + \beta_4 x_{70+} + \beta_5 x_{BritishColumbia} + \beta_6 x_{Manitoba} + \beta_7 x_{NewBrunswick} + \beta_8 x_{NewfoundlandandLabrador} + \beta_9 x_{NovaScotia} + \beta_{10} x_{Ontario} + \beta_{11} x_{PrinceEdwardIsland} + \beta_{12} x_{Quebec} + \beta_{13} x_{Saskatchewan} + \beta_{14} x_{LowHouseholdIncome} + \beta_{15} x_{MiddleHouseholdIncome} \tag{5}$$

The final logistic regression model for Green Party is :

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{Male} + \beta_2 x_{BritishColumbia} + \beta_3 x_{Manitoba} + \beta_4 x_{NewBrunswick} + \beta_5 x_{NewfoundlandandLabrador} + \beta_6 x_{NovaScotia} + \beta_7 x_{Ontario} + \beta_8 x_{PrinceEdwardIsland} + \beta_9 x_{Quebec} + \beta_{10} x_{Saskatchewan} + \beta_{11} x_{NeverAttendedUniversity} + \beta_{12} x_{BornoutsideCanada} + \beta_{13} x_{LowHouseholdIncome} + \beta_{14} x_{MiddleHouseholdIncome} \tag{6}$$

The final logistic regression model for BQ Party is :

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{31-50} + \beta_2 x_{51-70} + \beta_3 x_{70+} + \beta_4 x_{BritishColumbia} + \beta_5 x_{Manitoba} + \beta_6 x_{NewBrunswick} + \beta_7 x_{NewfoundlandandLabrador} + \beta_8 x_{NovaScotia} + \beta_9 x_{Ontario} + \beta_{10} x_{PrinceEdwardIsland} + \beta_{11} x_{Quebec} + \beta_{12} x_{Saskatchewan} + \beta_{13} x_{BornoutsideCanada} \tag{7}$$

The final logistic regression model for People's Party is :

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{Male} + \beta_2 x_{BritishColumbia} + \beta_3 x_{Manitoba} + \beta_4 x_{NewBrunswick} + \beta_5 x_{NewfoundlandandLabrador} + \beta_6 x_{NovaScotia} + \beta_7 x_{Ontario} + \beta_8 x_{PrinceEdwardIsland} + \beta_9 x_{Quebec} + \beta_{10} x_{Saskatchewan} + \beta_{11} x_{NeverAttendedUniversity} \tag{8}$$

3. Figure for Model section

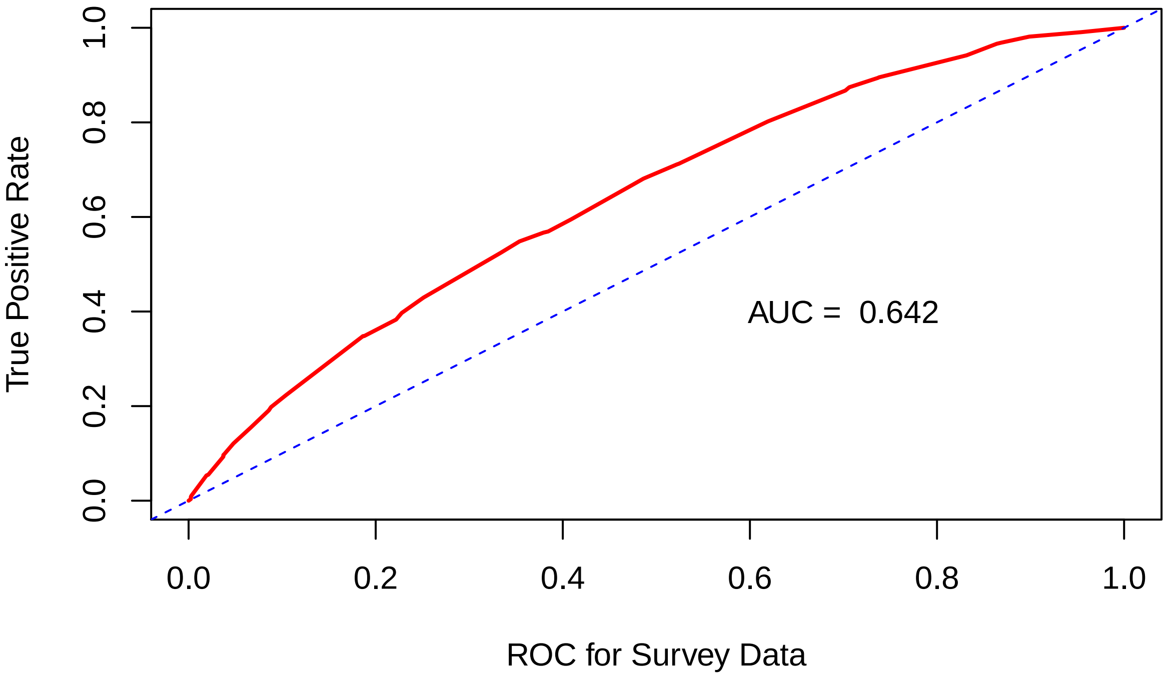

Figure 6: ROC Curve for Liberal Party

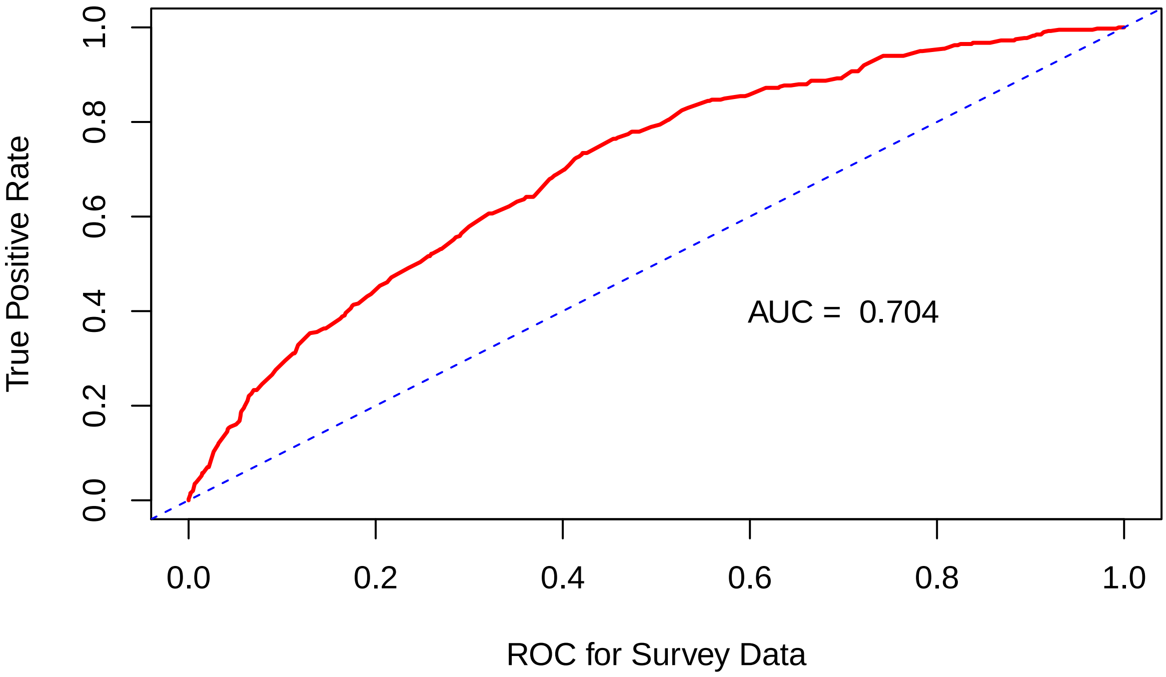

Figure 7: ROC Curve for NDP Party

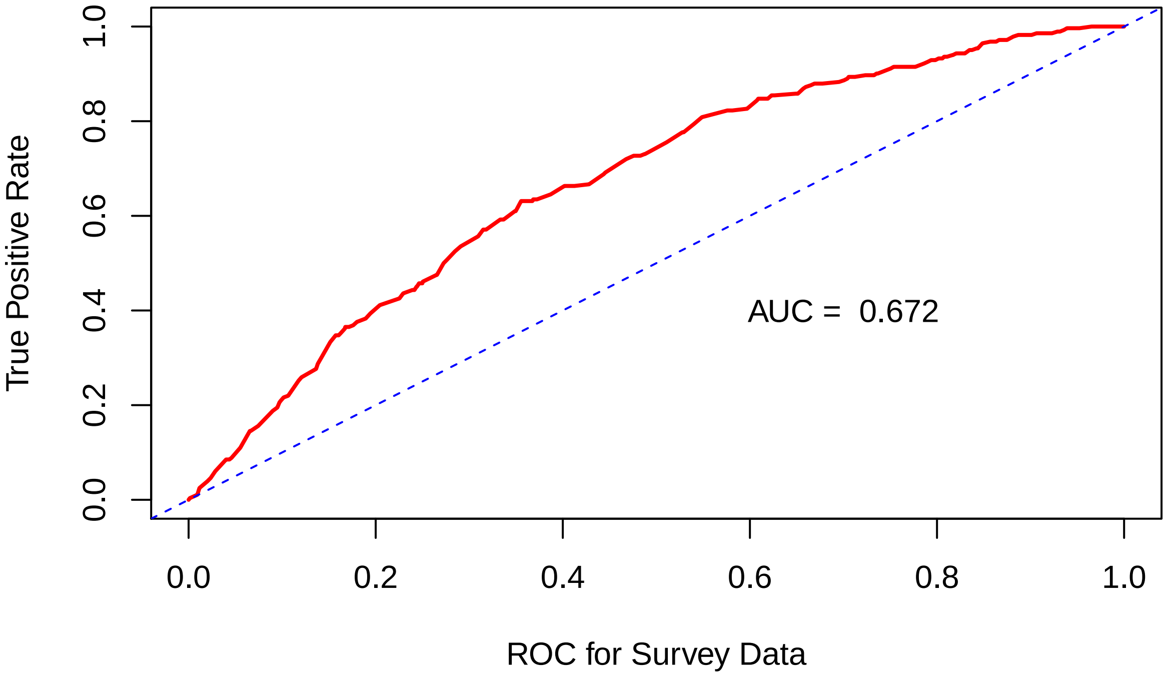

Figure 8: ROC Curve for Green Party

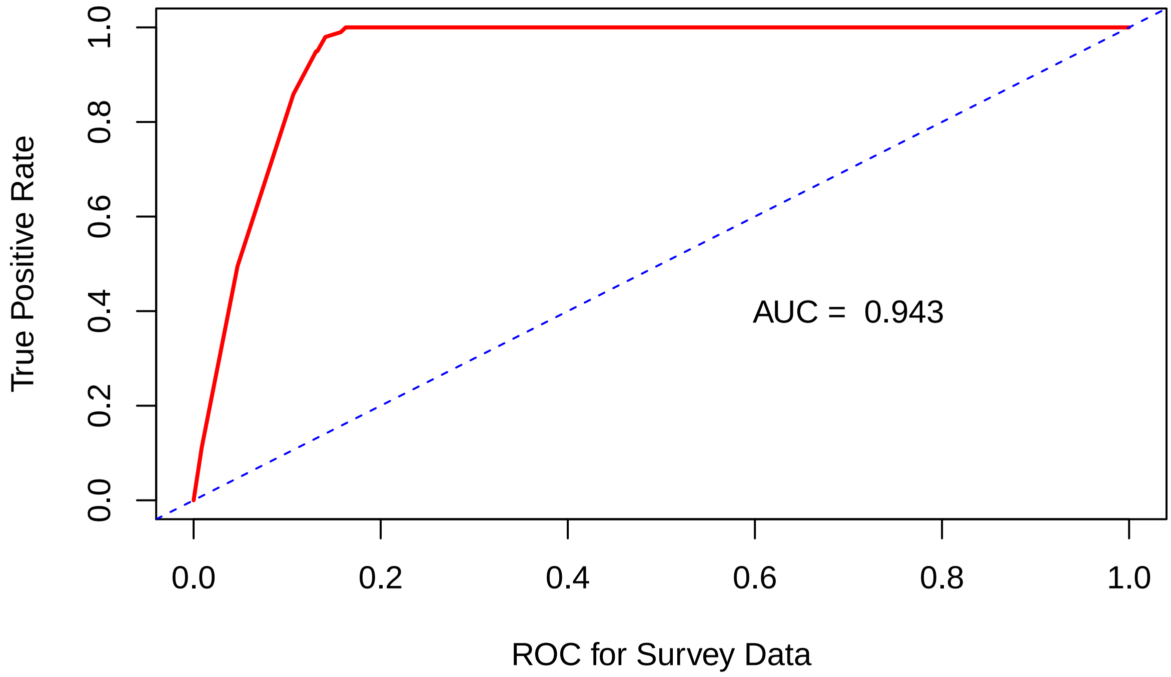

Figure 9: ROC Curve for BQ Party

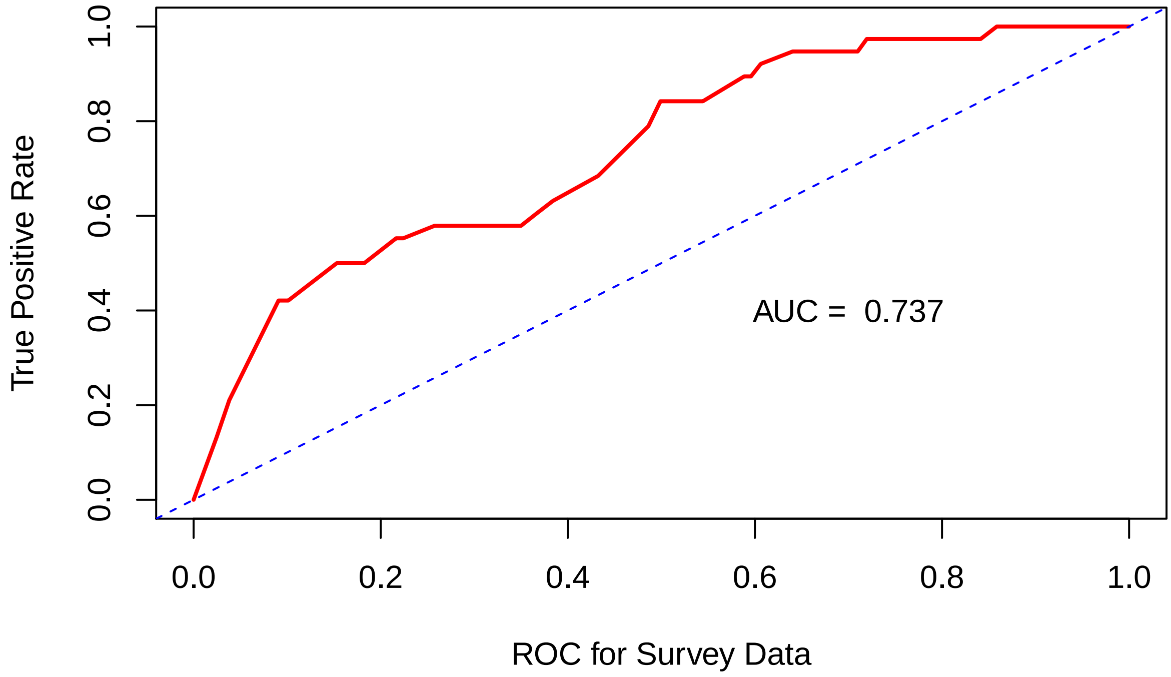

Figure 10: ROC Curve for People's Party

4. Final model with coefficients for Results section

The final logistic regression model for Liberal Party is :

$$log(\frac{p}{1-p}) = -1.51582 + 0.63471x_{BritishColumbia} + 0.75821x_{Manitoba} + 1.12743x_{NewBrunswick} \\ +1.34506x_{NewfoundlandandLabrador} + 1.35676x_{NovaScotia} + 1.33393x_{Ontario} + 1.43971x_{PrinceEdwardIsland} \\ +1.06741x_{Quebec} + 0.14982x_{Saskatchewan} - 0.46499x_{NeverAttendedUniversity} + 0.59676x_{BornoutsideCanada} \tag{9}$$

The final logistic regression model for NDP Party is :

$$log(\frac{p}{1-p}) = -1.3667 - 0.5976x_{Male} - 0.6307x_{31-50} - 1.1573x_{51-70} - 1.8739x_{70+} \\ +1.0443x_{BritishColumbia} + 0.3546x_{Manitoba} - 1.5203x_{NewBrunswick} + 1.0828x_{NewfoundlandandLabrador} \\ +0.4989x_{NovaScotia} + 0.5860x_{Ontario} - 2.993x_{PrinceEdwardIsland} + 0.1718x_{Quebec} + 0.2650x_{Saskatchewan} \\ +0.5008x_{LowHouseholdIncome} + 0.4339x_{MiddleHouseholdIncome} \tag{10}$$

The final logistic regression model for Green Party is :

$$log(\frac{p}{1-p}) = -2.911591 - 0.199985x_{Male} + 1.646875x_{BritishColumbia} + 0.563332x_{Manitoba} \\ +1.659105x_{NewBrunswick} + 0.118615x_{NewfoundlandandLabrador} + 1.075677x_{NovaScotia} + 0.776313x_{Ontario} \\ +1.446544x_{PrinceEdwardIsland} + 0.826646x_{Quebec} - 0.007498x_{Saskatchewan} - 0.385161x_{NeverAttendedUniversity} \\ -0.5548x_{BornoutsideCanada} + 0.393342x_{LowHouseholdIncome} + 0.152735x_{MiddleHouseholdIncome} \tag{11}$$

The final logistic regression model for BQ Party is :

$$log(\frac{p}{1-p}) = -22.06485 + 0.38541x_{31-50} + 0.95563x_{51-70} + 1.12800x_{70+} + 0.02929x_{BritishColumbia} - 0.11563x_{Manitoba} \\ -0.16735x_{NewBrunswick} - 0.15968x_{NewfoundlandandLabrador} - 0.13256x_{NovaScotia} - 0.02822x_{Ontario} \\ -0.17450x_{PrinceEdwardIsland} + 20.24754x_{Quebec} - 0.11538x_{Saskatchewan} - 1.48572x_{BornoutsideCanada} \tag{12}$$

The final logistic regression model for People's Party is :

$$log(\frac{p}{1-p}) = -6.1681 + 0.5071x_{Male} + 1.5277x_{BritishColumbia} + 0.0618x_{Manitoba} + 1.2692x_{NewBrunswick} \\ -14.4116x_{NewfoundlandandLabrador} + 0.3584x_{NovaScotia} + 0.1543x_{Ontario} + 1.1537x_{PrinceEdwardIsland} \\ +1.1629x_{Quebec} + 2.0572x_{Saskatchewan} + 1.0319x_{NeverAttendedUniversity} \tag{13}$$