Predicting Popular Vote for the Liberal, Conservative, and NDP of the Next Canadian Federal Election

STA304 - Assignment 2

GROUP 13: Daihao Wu, Jingyu Ye, Yihan Wang, Benjamin He

November 24, 2022

Introduction

Election and voting are crucial parts of a democratic country. They are critical links between citizen engagement and democratic representation, which provides citizens the opportunity to vote for parties that represent their ideas. In recent years, the forecasting of voting outcomes has made significant progress in various advanced industrial democracies such as the US, the UK, and France (Lewis-Beck, 2005). However, one country that has not received much attention is Canada (Bélanger & Godbout, 2010). Therefore, our analysis will focus on predicting the popular vote for the next Canadian federal election. Such prediction offers a useful lesson about elections and provides some baseline expectations beforehand. The data we used includes survey data based on 4021 cases and 278 variables from the 2019 Canada Election Study and census data from the 2017 General Social Survey. The survey data is intended to collect opinions and attitudes of the voting population during and after the 2019 federal election and the census data collects the socioeconomic characteristics of Canadians.

This paper will explore the popular vote for three political parties including Liberal, Conservative, and NDP based on voters' sex, age, education, household income, household size, and whether they are born in Canada or not. Specifically, we want to investigate which party among the three will yield the highest popular vote based on the six predictors in our model. The six predictors are chosen because (1) Age has been shown to be influencing voting decisions, where people at a younger age tend to vote more left-wing (more emphasis on freedom, equality, fraternity, rights, progress, reform, and internationalism) but later in life, they tend to vote more right-wing (more emphasize on authority, hierarchy, order, duty, tradition, reaction, and nationalism) (Kaushik, 2017; Heywood, 2015). (2) Sex differences also exist in electoral support, women are less likely to vote for the Conservative Party and more likely to vote for NDP (Anderson & Stephenson, 2010). (3) Immigrants are more stable supporters of the Liberal Party than Canadian-born voters (Anderson & Stephenson, 2010). (4) Education level also affects voting behaviors, those with higher levels of education not only are more likely to vote, but they also tend to lean toward the Liberal Party (Anderson & Stephenson, 2010). (5) Income also plays a role in voting preference: higher-income earners are more likely to vote for the Conservatives whereas low-income earners vote NDP (McGrane, 2007). (6) Household size could also alter voting behaviors as one's decision could be influenced by other family members.

The three parties are chosen because they are the top three parties that received the most votes among all parties in the 2021 Canadian election (CBC/Radio Canada, n.d.). It is hypothesized that based on the outcome of the 2021 election (CBC/Radio Canada, n.d.), we predict the model will estimate that the Conservative party will yield the highest popular vote among the three parties although the Conservative party lost the election due to winning fewer ridings than the Liberal party. Nonetheless, we recognize the limitation that our dataset does not contain information regarding which riding the person. Thus, we are predicting the popular vote share % instead of ridings won for each party which may be less influential since having the most percentage of vote share does not indicate winning the election.

Data

Data collection

Survey data collection: The survey data is based on 4021 cases and 278 variables from the 2019 Canada Election Study (Stephenson et al., 2022). This survey is intended to collect the opinions and attitudes of Canadians during and after the 2019 federal election. Citizens and permanent residents of Canada, who are 18 years of age or older, are the targeted population. Data is collected through computer-assisted telephone interviews from 2019-9-10 to 2019-10-20 (campaign period survey) and from 2019-10-22 to 2019-11-21 (post-election survey).

Census data collection: Census data comes from the 2017 General Social Survey (Government of Canada, 2019). Statistical information on the living conditions and well-being of Canadians is provided by the General Social Survey (GSS), an annual survey with rotating content. Data collection took place between February 1 and November 30, 2017. A computer-assisted telephone interviewing method was used to collect data directly from survey respondents. Non-institutionalized persons 15 years and older living in Canada's 10 provinces are the target population for the 2017 General Social Survey. Approximately 43,000 units were sampled, among whom about 34,000 invitation letters were sent to selected households across Canada. 20,000 questionnaires were expected to be completed.

Data Cleaning

The package tidyverse is used during the data cleaning process (Wickham et al., 2019).

Survey data cleaning: We preliminarily (specific predictors for each party will be decided in the model selection process) selected age, education, sex, household size, income, born in Canada or not as predictors for the voting responses for Liberal Party, Conservative Party, and New Democratic Party. Thus, these columns are filtered from the original dataset. The voting decision variable was cleaned by excluding responses such as "None of these", "Will not vote", "Will spoil ballet", "Refused", and "Don't know/undecided", since we need to make predictions based on the voting decision, we filtered the response to show which party people would vote for. Since there is only one observation of the "other" response, we exclude the category "other" from the gender variable to prevent outliers. To match the census data, the variable gender was renamed "Sex" and the responses "1" and "2" were changed to "Male" and "Female". We exclude the responses "Refused" and "Don't know" from the household income variable, since we need income information to predict voting decisions. In addition, we classified responses by income amount as "less than $25000", "$25,000 to $49,999", "$50,000 to $74,999", "$75,000 to $99,999", "$100,000 to 124,999", and "$125,000 and more". For the variable education, we classified the responses into three levels, including "High School or Under", "College or Non-University Certificate", and "Bachelor or Above". For the variable number of people in the household, the range of the number is from 1 to 6 in the census data, so we filtered the responses of the number of people in the household from 1 to 6 to match the census data. For the variable born in Canada, we classified it into two categories, if the responses show that the person was born in Canada or Quebec, then it is classified as "Born in Canada", other responses are categorized as "Born outside Canada". For the variable Age, since the highest age in the census data in the year 2017 is 80, so for the survey data in 2019, we limit the highest age to 82. The missing responses are removed from the survey data and census data in order to make the variables used in the study more informative. Moreover, since our objective is to predict the voting of the Liberal Party, Conservative Party, and New Democratic Party, each of the responses was changed into a binary response, for example, if the voting was for the Liberal Party, it was 1. Otherwise, it was 0.

Census data cleaning: For the variable age, since the age in the survey data are integers, so we rounded the age in the census data, and added two to all the ages in the census data for the reason that the census data was from 2017 but the survey data was from 2019. For the variable education, it is also categorized into three levels, for all the obtained degrees higher than Bachelor, we changed them as "Bachelor or Above", for all the obtained degree lower or equal to high school, we classified them as "High School or Under", for the obtained degree between these two levels, it is renamed as "College or Non-University Certificate". After that, we renamed the column names of the six interested predictors to match their names in the survey data. We also filtered the predictor age, education, sex, household size, income, born in Canada or not from the census data and then omit any missing values from the dataset. For the variable born in Canada, the responses

"Don't know" were removed as they caused ambiguity and confounded the effect of the responses "Born in Canada" and "Born outside Canada" in the prediction for the voting response.

Description of the important variables

Age, Education, Sex, Household Size, Income, and Born in Canada were selected as predictor variables in this study.

- Sex:

There are two options in the responses after cleaning: male and female.

- Age:

The survey asks about the age of the respondent. After data cleaning, the answer to this question is a positive integer from 18-82.

- Education:

This survey asks the respondents about their highest level of education. The responses to this question are categorized into the following three categories: "High School or Under", "College or Non-University Certificate", and "Bachelor or Above".

- Income:

In this survey question, respondents are asked to estimate their household income over the past 12 months. The answer is a positive integer. The annual household income in CAD is categorized into six categories: $<$ 25000, 25000 $\sim$ 49999, 50000 $\sim$ 74999, 75000 $\sim$ 99999, 100000 $\sim$ 124999, and $>$ 125000.

- Household size:

In this survey question, respondents are asked to estimate the number of people living in the household. The answer is a positive integer from 1 to 6.

- Born in Canada:

The survey asks about the birthplace of the respondents, after data cleaning we categorized people into two categories: born in Canada or born outside Canada.

- Response Variable: Voting for Liberal, Voting for Conservative, Voting for NDP

The survey asks which Party the respondents would vote for, we aim to predict each of the voting for the Liberal Party, Conservative Party, and New Democratic Party, the response for each Party is a binary variable.

The package "gtsummary" is used to create the summary table (Presentation-Ready Data Summary and Analytic Result Tables [R Package Gtsummary Version 1.6.2], 2022).

Table 1: **Survey data summary**

Characteristic	**N = 2,160**
Age	51(16)
Education	
Bachelor or Above	986 / 2,160 (46%)
College or Non-University Certificate	779 / 2,160 (36%)
High School or Under	395 / 2,160 (18%)
Sex	
Female	870 / 2,160 (40%)
Male	1,290 / 2,160 (60%)
Household size	
1	420 / 2,160 (19%)
2	864 / 2,160 (40%)

Characteristic	N = 2,160
3	325 / 2,160 (15%)
4	347 / 2,160 (16%)
5	149 / 2,160 (6.9%)
6	55 / 2,160 (2.5%)
Income	
$100,000 to $ 124,999	261 / 2,160 (12%)
$125,000 and more	664 / 2,160 (31%)
$25,000 to $49,999	314 / 2,160 (15%)
$50,000 to $74,999	396 / 2,160 (18%)
$75,000 to $99,999	293 / 2,160 (14%)
Less than $25,000	232 / 2,160 (11%)
Born in or outside Canada	
Born in Canada	1,839 / 2,160 (85%)
Born outside Canada	321 / 2,160 (15%)
Percentage of Voting Liberal	720 / 2,160 (33%)
Percentage of Voting Conservatives	756 / 2,160 (35%)
Percentage of Voting NDP	325 / 2,160 (15%)

From Table 1, we can see the survey data contains 2160 samples, with an average age of 51. The samples contain 46% people with a Bachelor's degree or higher, 36% people with a college degree or non-university certificate, and 18% with a high school diploma or less. The sample has a higher percentage of males (60%) than females (40%) The household size of 40% in the sample consists of two people. 31% of people earn over 125000 per year, while 11% earn less than 25000 per year. 85% of the respondents were born in Canada. There are 33% of people voting Liberal, 35% voting Conservative, and 15% voting NDP.

Table 2: **Census data summary**

Characteristic	N = 20,161
Age	54(18)
Education	
Bachelor or Above	5,572 / 20,161 (28%)
College or Non-University Certificate	6,750 / 20,161 (33%)
High School or Under	7,839 / 20,161 (39%)
Sex	
Female	10,969 / 20,161 (54%)
Male	9,192 / 20,161 (46%)
Household size	
1	5,711 / 20,161 (28%)
2	7,603 / 20,161 (38%)
3	2,818 / 20,161 (14%)
4	2,697 / 20,161 (13%)
5	966 / 20,161 (4.8%)
6	366 / 20,161 (1.8%)
Income	
$100,000 to $ 124,999	2,123 / 20,161 (11%)
$125,000 and more	4,626 / 20,161 (23%)
$25,000 to $49,999	4,253 / 20,161 (21%)
$50,000 to $74,999	3,603 / 20,161 (18%)
$75,000 to $99,999	2,857 / 20,161 (14%)
Less than $25,000	2,699 / 20,161 (13%)
Born in or outside Canada	

Characteristic	**N = 20,161**
Born in Canada	16,142 / 20,161 (80%)
Born outside Canada	4,019 / 20,161 (20%)

According to Table 2, there are in total of 20161 people in the census data, with an average age of 54. There are 28% of people in the census with a bachelor's degree or higher, 33% with a college or non-university certificate, and 39/% with a high school diploma or less. The population is more female than male (54% vs. 46%). In 38% of households, two people are living together. 23% of people earn more than 125000 per year, while 13% make less than 25000 per year. In the census data, 80 percent of people were born in Canada.

Table 3: Voting Liberal Party, Conservative Party and New Democratic Party between male and female

Sex	Number of people	Probability of voting Liberal Party	Probability of voting Conservative Party	Probability of voting New Democratic Party
Female	870	0.3597701	0.2655172	0.1954023
Male	1290	0.3155039	0.4069767	0.1201550

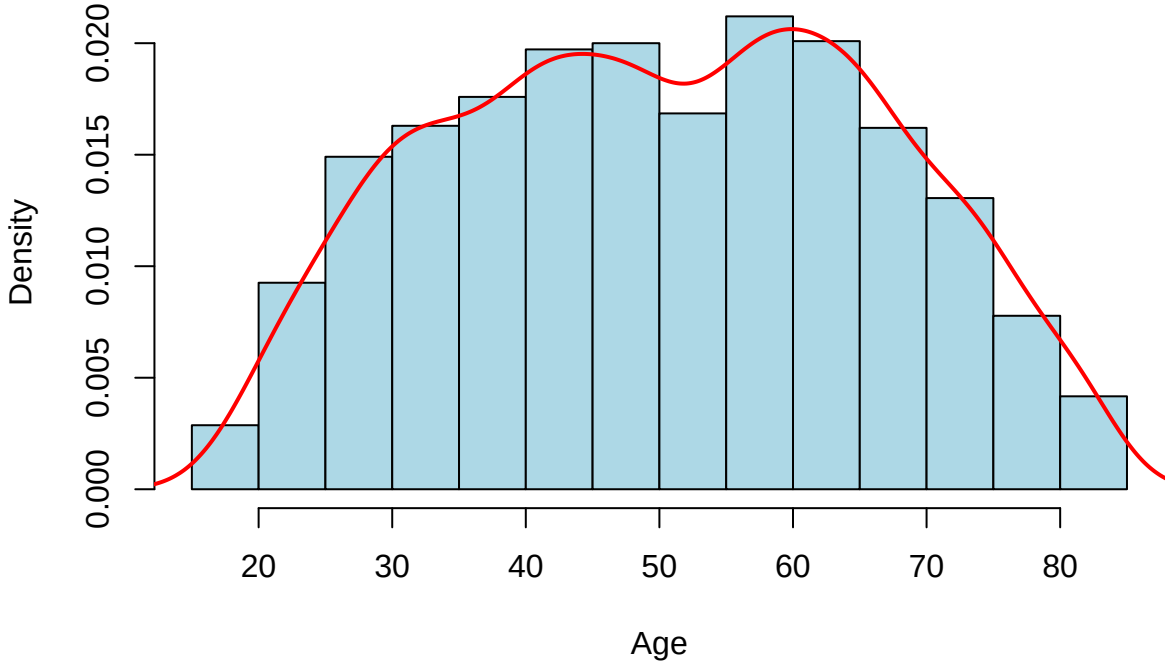
According to Table 3, there are 870 females and 1290 males in the survey data, compared to males, females have a larger probability to vote more for the Liberal Party (0.36 vs. 0.32) and New Democratic Party (0.195 vs. 0.120), and a smaller probability to vote for the Conservative Party (0.27 vs. 0.41).

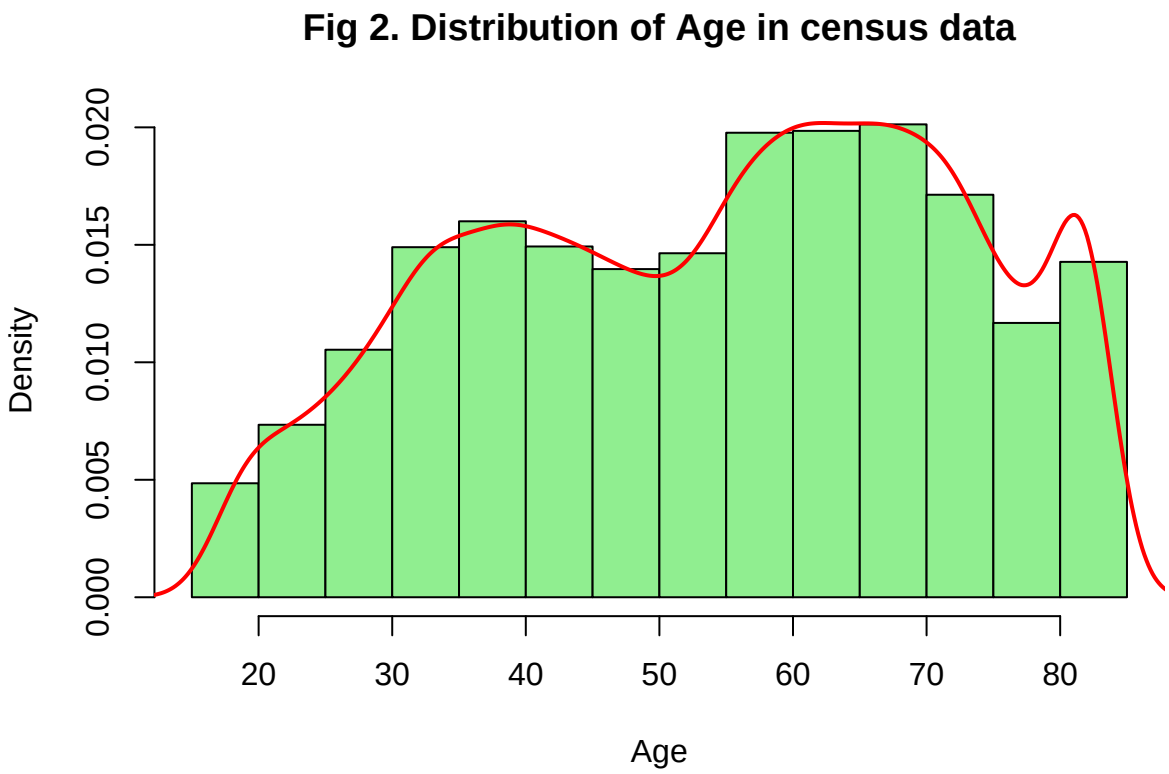
Table 4: Voting Liberal Party, Conservative Party and New Democratic Party between people born in or outside Canada

Born in Canada or not	Number of people	Probability of voting Liberal Party	Probability of voting Conservative Party	Probability of voting New Democratic Party
Born in Canada	1839	0.3099511	0.3621533	0.1489940
Born outside Canada	321	0.4672897	0.2803738	0.1588785

According to Table 4, there are 1839 people who were born in Canada and 321 people who were born outside Canada, people who were born outside Canada have a larger probability to vote for the Liberal Party than people who were born in Canada (0.47 vs. 0.31), people who were born outside Canada have a smaller probability to vote for Conservative Party than people born in Canada (0.28 vs. 0.36). The probability for both of them to vote for the New Democratic Party is approximately the same.

Fig 1. Distribution of Age in survey data





From the two histograms (Fig 1 and Fig 2) we can observe that there are more elder people in the census data (Fig 2) compared to the survey data (Fig 2). From the density curve we could see that the majority age of people in the survey data is from 35 to 70, while the majority age of people in the census data is from 50 to 80. Consequently, the difference between the distribution of predictors such as age in census data and survey data may be overcome with post-stratification.

Methods

In order to investigate which party among the three will yield the highest popular vote based on some or all of the six predictors (age, sex, income, education level, household size, and born in Canada or not) in our model, Logistic regression is helpful in computing the probability of a binary outcome, like in our case, it would be "vote" or "Do not vote". Since we are identifying the proportion of voters who would vote for a certain party, which is equivalent to finding out the probability of voting for a certain party, we could use a Logistic regression model for each of the three most popular political parties such as the Liberal, Conservative and NDP based on some or all of our interested variables: age, sex, education level, income level, whether or not the voter was born in Canada and the household size the voter belongs to (specific predictors will be decided through the model selection process). Therefore, we plan to build three logistic regression models using the survey data to predict the percentage of the popular vote for the Liberal Party, the Conservative Party, and the NDP. Then post-stratification would be applied to each of the models because it is an effective method for correcting the known differences between our survey sample data and the target populations (census data). This method partition the population into "cells" based on all combinations of different attributes in the model, and estimate the response within each cell using the survey sample data, then all of the estimates from the cells would be aggregated to a population level by re-weighting each cell by its relative proportion compared to the population. The detailed description of the model for each of the parties and how post-stratification would be implemented would be described in the following sections.

Model Specifics

The Akaike Information Criterion (AIC) is used in our model selection. AIC is a mathematical method for evaluating how well a model fits the data it was generated from. It is calculated from the number of predictors and the MLE of the model. The best-fit model according to AIC is the one that explains the greatest amount of variation in the response using the fewest possible independent variables. Therefore, since we are planning to use the six predictors that have been discussed previously, writing out all possible models would be overwhelming. Thus, we decided to compare the AIC of models as we are removing predictors because AIC works by penalizing models that use more parameters. We compute the AIC of the models as we remove one of the predictors to see if it decreases. If it decreases, we drop the removed predictor and keep trying out removing predictors based on the reduced model. We repeat the process until the AIC does not decrease. It is worth noting that AIC does not provide information about the absolute quality of the model, but only the quality relative to other models.

From the process, the three models with the lowest AIC are selected for the Liberal, Conservative, and NDP parties.

Liberal Party:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \hat{\beta}_0 + \hat{\beta}_1 x_{Age} + \hat{\beta}_2 x_{College} + \hat{\beta}_3 x_{HighschoolOrBelow} + \hat{\beta}_4 x_{Male} + \hat{\beta}_5 x_{BornOutsideCanada}$$

The model introduced is the logistic regression model for predicting the proportion of voters who would vote for the Liberal Party using the variables age, sex, education level, and whether or not the voter was born in Canada.

$\hat{p}$ represents the probability of voting for the Liberal Party.

$\hat{\beta}_0$ coefficient represents the intercept term of the model, that is the log odds of voting for the Liberal Party when the voter is at a certain age, education background, sex, and place of birth.

$\hat{\beta}_1$ coefficient represents the change in log odds of voting for the Liberal Party for every one-year increase in age.

$\hat{\beta}_2$ coefficient represents the average difference in log odds of voting for the Liberal Party between voters with a college or non-University certificate degree and voters with a bachelor or above degree with a certain age, sex, and place of birth.

$\hat{\beta}_3$ coefficient represents the average difference in log odds of voting for the Liberal Party between voters with high school or below education background and voters with a bachelor's or above degree with a certain age, sex, and place of birth.

$\hat{\beta}_4$ the average difference in log odds of voting for the Liberal Party between male and female voters with a certain age, education background, and place of birth.

$\hat{\beta}_5$ coefficient represents the average difference in log odds of voting for the Liberal Party between voters born outside Canada and voters born in Canada with a certain age, education background, and sex.

x_{Age} equals a numerical value that represents the age of the voter.

$x_{College} = 1$ if the voter has a college or non-University certificate education background, else $x_{College} = 0$.

$x_{HighschoolOrBelow} = 1$ if the voter has a highschool or below education background, else $x_{HighschoolOrBelow} = 0$.

$x_{Male} = 1$ if the voter is Male, and if the voter is Female, then $x_{Male} = 0$.

$x_{BornOutsideCanada} = 1$ if the voter is born outside of Canada, and if the voter is born in Canada, then $x_{BornOutsideCanada} = 0$

For this logistic regression model, there are four assumptions to check for:

1. Outcome is binary

Because the response variable for this model is to predict whether or not a voter would vote for the Liberal Party, so the outcome of this model would be either "vote" or "do not vote", so this assumption is satisfied.

2. linearity in the logit for continuous variables

From this logistic regression model, there is only one continuous variable, and that is "Age". To check for the linearity of the variable "Age" in this logistic regression model, we set up the Box-Tidwell Test to test whether the logit transformation is a linear function of the predictor "Age" by adding the non-linear transformation of the original predictor "Age" as an interaction term to test if such change did not improve the prediction.

From Box-Tidwell, H_0 : linearity between Age and log odds vs. H_A : non-linearity between Age and log odds.

The resulting P-value from the Box-Tidwell test for the Liberal Party model is 0.1880396, which is quite large, so we fail to reject the null hypothesis, implying that the linearity assumption seems to be satisfied for the Liberal Party model.

The Box-Tidwell test is run in R using the package "car" (Fox & Weisberg, 2019).

3. Absence of multicollinearity

To check whether or not the predictors in our sample data are highly correlated, we could compute the variance inflation factors from our dataset for this model to help us identify multicollinearity. Variance Inflation Factor measures the ratio between the variance of estimating the parameter in a model that includes multiple other predictors and the variance of a model that only includes one predictor, so the ratio interprets the severity of multicollinearity between predictors, and generally, a factor exceeds 5 would be due to existence of multicollinearity in the model (James et al., 2017).

From table 5, we can see that none of our predictors have variance inflation factors exceeding 5, which means that none of them seem to be correlated to each other in this model, so the multicollinearity assumption would be satisfied.

The Variance Inflation Factor is computed in R using the package "car" (Fox & Weisberg, 2019).

Table 5: Variance Inflation Factors of Logistic Regression Model for the Liberal Party

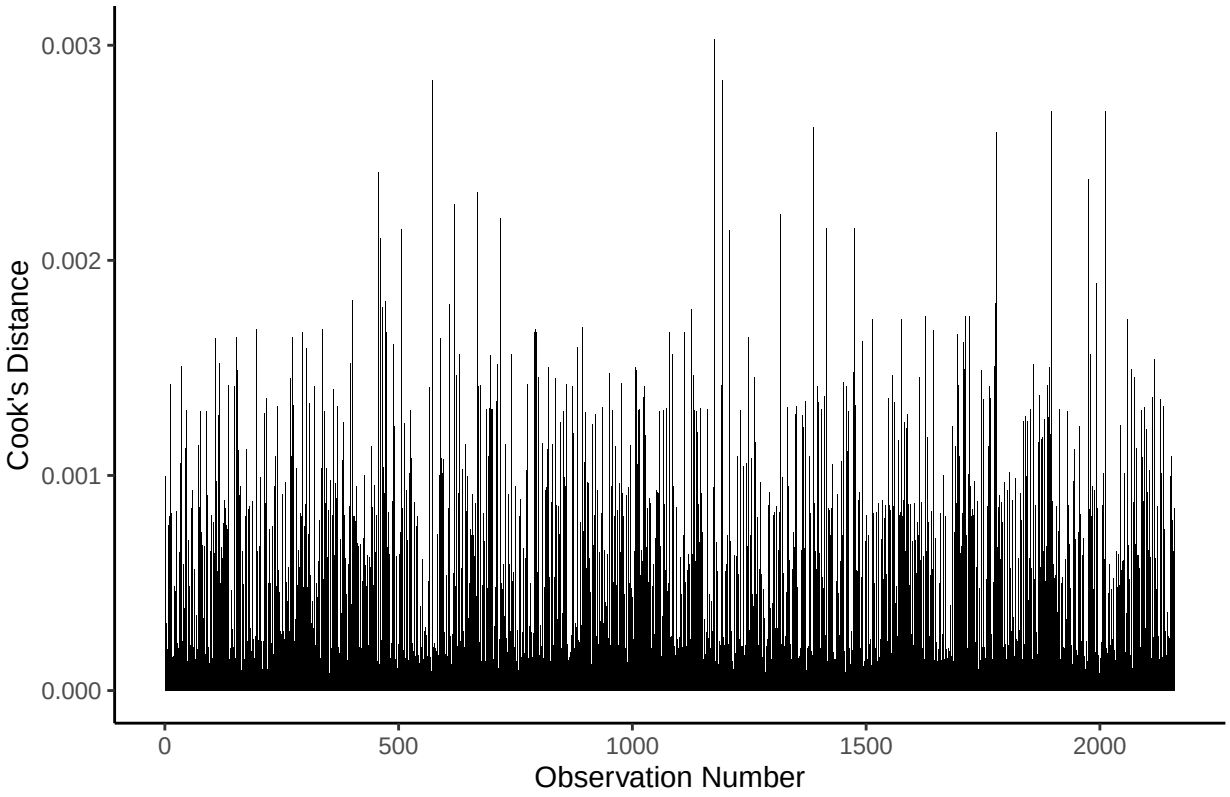
	Variance Inflation Factors	Degree of Freedom
Age	1.010081	1
Education	1.032403	2
Sex	1.009711	1
Born_Canada	1.020040	1

4. Checking influential values

To check for the influential values that could affect the quality of this model, we would examine these values by visualizing Cook's distance values for each of the models (Kassambara, 2018).

By observing Fig 3, we can see that there are a few large values that seem to be influential observations in this model, but because the difference of the Cook's distance between these observations and the majority of the observations is merely around 0.0015, so it seems relatively acceptable to keep these observations in the dataset.

Fig 3. Cook's Distance for Each Observation in Liberal Party Model



Therefore, after checking all the assumptions, we now have the logistic regression model for predicting the outcome of the Liberal Party using the survey data based on the predictor age, education background, sex, and place of birth.

Conservative Party:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \hat{\beta}_0 + \hat{\beta}_1 x_{Age} + \hat{\beta}_2 x_{College} + \hat{\beta}_3 x_{HighschoolOrBelow} + \hat{\beta}_4 x_{Male} + \hat{\beta}_5 x_{HouseholdSize} +$$

$$\hat{\beta}_6 x_{Income>125k} + \hat{\beta}_7 x_{Income=25k-50k} + \hat{\beta}_8 x_{Income=50k-75k} + \hat{\beta}_9 x_{Income=75k-100k} + \hat{\beta_{10}} x_{Income<25k} +$$

$$\hat{\beta_{11}} x_{BornOutsideCanada}$$

The model introduced is the logistic regression model for predicting the proportion of voters who would vote for the Conservative Party using the variables age, sex, education level, income level, whether or not the voter was born in Canada, and the household size the voter belongs to.

$\hat{p}$ represents the probability of voting for the Conservative Party.

$\hat{\beta}_0$ coefficient represents the intercept term of the model, that is the log odds of voting for the Conservative Party when the voter is at a certain age, education background, sex, household size, income level, and place of birth.

$\hat{\beta}_1$ coefficient represents the change in log odds of voting for the Conservative Party for every one-year increase in age.

$\hat{\beta}_2$ coefficient represents the average difference in log odds of voting for the Conservative Party between voters with college or non-University certificate degree and voters with bachelor's or above degree with a certain age, sex, household size, income level and the place of birth.

$\hat{\beta}_3$ coefficient represents the average difference in log odds of voting for the Conservative Party between voters with high school or below education background and voters with a bachelor's or above degree with a certain age, sex, household size, income level and the place of birth.

$\hat{\beta}_4$ the average difference in log odds of voting for the Conservative Party between male and female voters with a certain age, education background, household size, income level, and place of birth.

$\hat{\beta}_5$ coefficient represents the change in log odds of voting for the Conservative Party for every one unit increase in the household size.

$\hat{\beta}_6$ coefficient represents the average difference in log odds of voting for the Conservative Party between voters with household income more than $125,000$\$ and voters with household income from $100,000$\$ to $124,999$\$ with a certain age, education background, sex, household size and the place of birth.

$\hat{\beta}_7$ coefficient represents the average difference in log odds of voting for the Conservative Party between voters with household income from $25,000$\$ to $49,999$\$ and voters with household income from $100,000$\$ to $124,999$\$ with a certain age, education background, sex, household size and the place of birth.

$\hat{\beta}_8$ coefficient represents the average difference in log odds of voting for the Conservative Party between voters with household income from $50,000$\$ to $74,999$\$ and voters with household income from $100,000$\$ to $124,999$\$ with a certain age, education background, sex, household size and the place of birth.

$\hat{\beta}_9$ coefficient represents the average difference in log odds of voting for the Conservative Party between voters with household income from $75,000$\$ to $99,999$\$ and voters with household income from $100,000$\$ to $124,999$\$ with a certain age, education background, sex, household size and the place of birth.

$\hat{\beta_{10}}$ coefficient represents the average difference in log odds of voting for the Conservative Party between voters with household income less than $25,000$\$ and voters with household income from $100,000$\$ to $124,999$\$ with a certain age, education background, sex, household size and the place of birth.

$\hat{\beta_{11}}$ coefficient represents the average difference in log odds of voting for the Conservative Party between voters born outside Canada and voters born in Canada with a certain age, education background, household size, income level, and the place of birth.

x_{Age} equals a numerical value that represents the age of the voter.

$x_{College} = 1$ if the voter has a college or non-University certificate education background, else $x_{College} = 0$.

$x_{HighschoolOrBelow} = 1$ if the voter has a high school or below education background, else $x_{HighschoolOrBelow} = 0$.

$x_{Male} = 1$ if the voter is Male, and if the voter is Female, then $x_{Male} = 0$.

$x_{HouseholdSize}$ equals a numerical value that represents the number of people in the household the voter belongs to.

$x_{Income>125k} = 1$ if the voter has a household income greater than $125,000$\$, else $x_{Income>125k} = 0$

$x_{Income=25k-50k} = 1$ if the voter has a household income from $25,000$\$ to $49,999$\$, else $x_{Income=25k-50k} = 0$

$x_{Income=50k-75k} = 1$ if the voter has a household income from $50,000$\$ to $74,999$\$, else $x_{Income=50k-75k=0}$

$x_{Income=75k-100k} = 1$ if the voter has a household income from $75,000$\$ to $99,999$\$, else $x_{Income=75k-100k} = 0$

$x_{Income<25k} = 1$ if the voter has a household income less than $25,000$\$, else $x_{Income<25k} = 0$

$x_{BornOutsideCanada} = 1$ if the voter is born outside of Canada, and if the voter is born in Canada, then $x_{BornOutsideCanada} = 0$

Similarly, For this logistic regression model, there are four assumptions to check for:

1. Outcome is binary

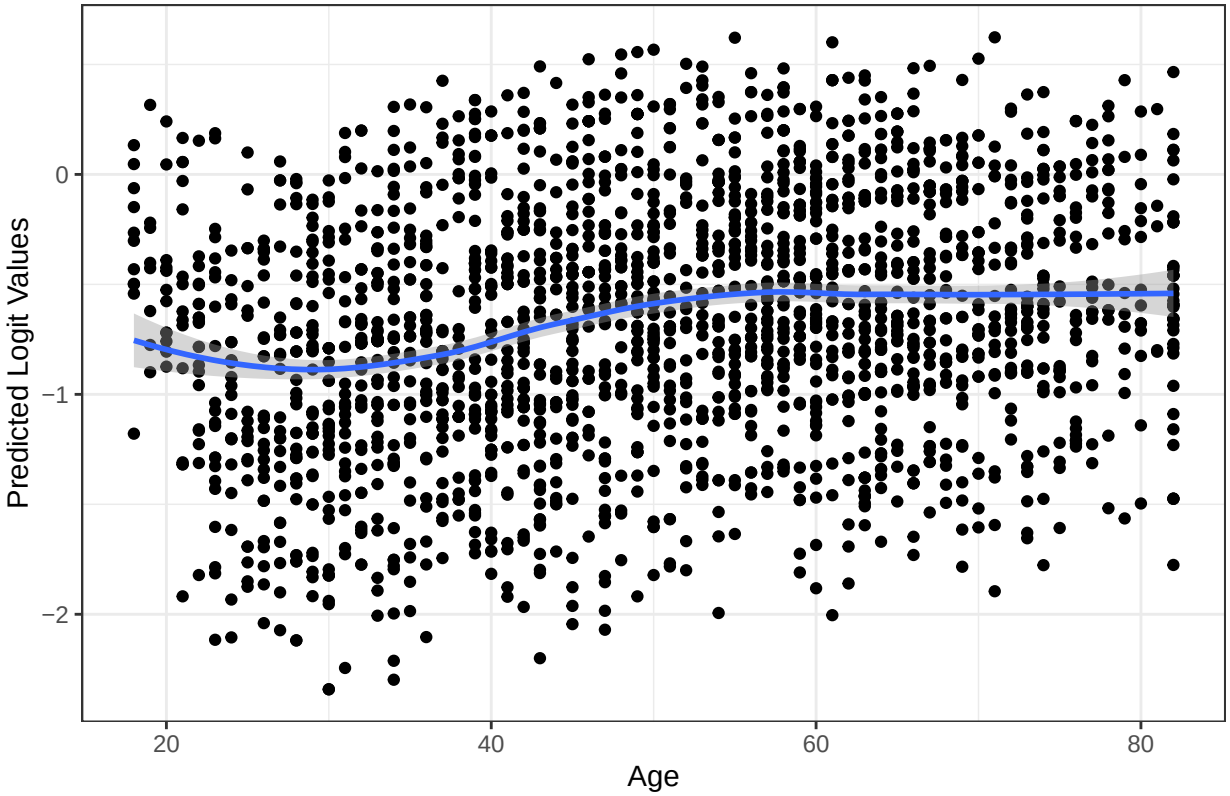
Because the response variable for this model is to predict whether or not a voter would vote for the Liberal Party, so the outcome of this model would be either "vote" or "do not vote", so this assumption is satisfied.

2. linearity in the logit for continuous variables

From this logistic regression model, there is only one continuous variable, and that is "Age". Similarly, we could have run the Box-Tidwell test to see if the linearity is satisfied like what we did for the previous model, but the Box-Tidwell could not successfully perform the test due to various possible reasons in this model.

Therefore, we could instead visually inspect the scatter plot between the predictor "Age" and the logit values predicted from the Conservative Party model to see if a linear relationship exists between them (Kassambara, 2018). From Fig 4, the blue line across the plot shows that the variable "Age" is linearly associated with the Conservative voting outcome in a logit scale, so we could say that the linearity assumption might not be violated for the Conservative Party model.

Fig 4. Predicted Logit Values from Conservative Party Model against Age



3. Absence of multicollinearity

To check whether or not the predictors in our sample data are highly correlated, we would also compute the variance inflation factors from our dataset for this model to help us identify multicollinearity (James et al., 2017).

From table 6, we can see that none of our predictors have variance inflation factors exceeding 5, which means that none of them seem to be correlated to each other in this model, so the multicollinearity assumption would be satisfied.

The Variance Inflation Factor is computed in R using the package "car" (Fox & Weisberg, 2019).

Table 6: Variance Inflation Factors of Logistic Regression Model for the Conservative Party

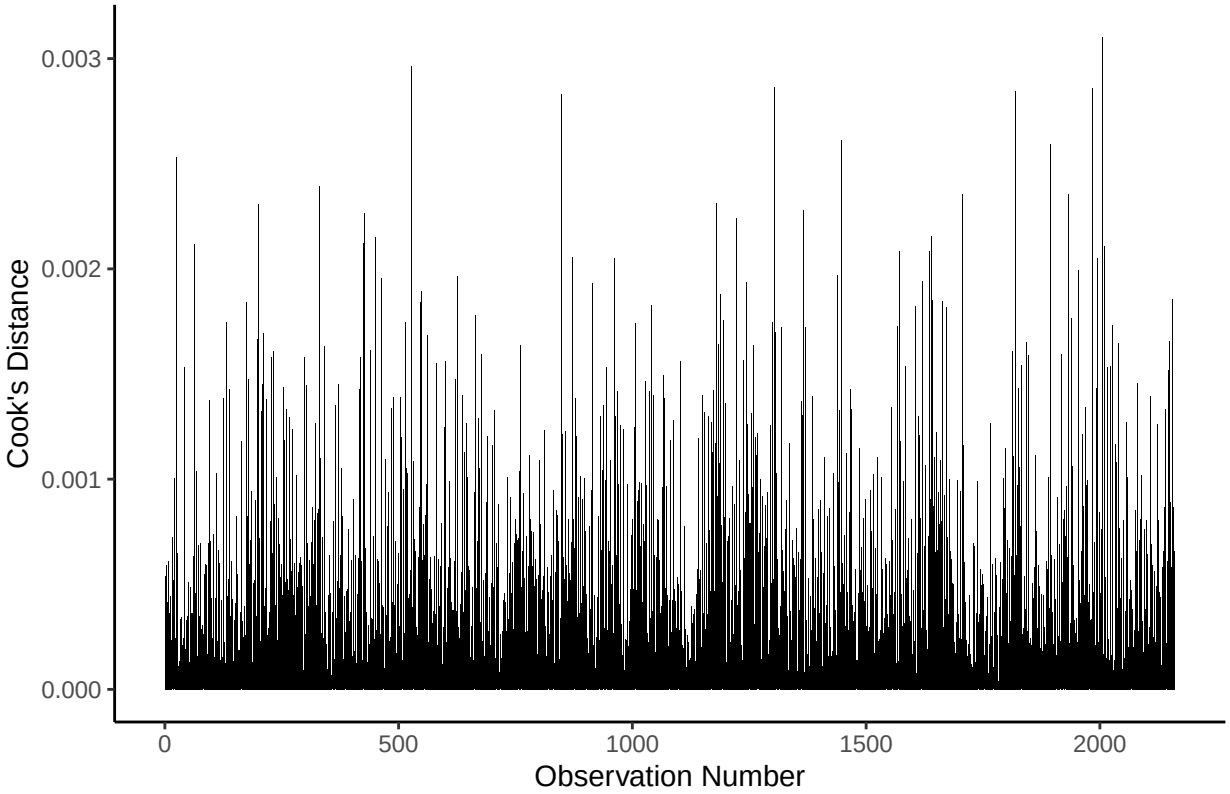
	Variance Inflation Factors	Degree of Freedom
Age	1.133719	1
Education	1.137255	2
Sex	1.019097	1
Household_Size	1.257629	1
Income	1.255898	5
Born_Canada	1.044703	1

4. Checking influential values

To check for the influential values that could affect the quality of this model, similarly, we would examine these values by visualizing Cook's distance values for each of the models (Kassambara, 2018).

By observing Fig 5, we can see that there are quite a few large values that seem to be influential observations in this model, but because the difference of the Cook's distance between these observations and the majority of the observations is still small, so it seems relatively acceptable to keep these observations in the dataset.

Fig 5. Cook's Distance for Each Observation in Conservative Party Model



Therefore, after checking all the assumptions, we now have the logistic regression model for predicting the outcome of the Conservative Party using the survey data based on the predictor age, education background, sex, household size, household income, and place of birth.

NDP:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \hat{\beta}_0 + \hat{\beta}_1 x_{Age} + \hat{\beta}_2 x_{Male} + \hat{\beta}_3 x_{Income>125k} + \hat{\beta}_4 x_{Income=25k-50k} + \hat{\beta}_5 x_{Income=50k-75k} +$$

$$\hat{\beta}_6 x_{Income=75k-100k} + \hat{\beta}_7 x_{Income<25k}$$

The model introduced is the logistic regression model for predicting the proportion of voters who would vote for the NDP using the variables age, sex, and income level.

$\hat{p}$ represents the probability of voting for the NDP.

$\hat{\beta}_0$ coefficient represents the intercept term of the model, that is the log odds of voting for the NDP when the voter is at a certain age, sex, and income level.

$\hat{\beta}_1$ coefficient represents the change in log odds of voting for the NDP for every one-year increase in age.

$\hat{\beta}_2$ the average difference in log odds of voting for the NDP between male and female voters with a certain age and income level.

$\hat{\beta}_3$ coefficient represents the average difference in log odds of voting for the NDP between voters with household income more than $125,000$\$ and voters with household income from $100,000$\$ to $124,999$\$ with a certain age and sex.

$\hat{\beta}_4$ coefficient represents the average difference in log odds of voting for the NDP between voters with household income from $25,000$\$ to $49,999$\$ and voters with household income from $100,000$\$ to $124,999$\$ with a certain age and sex.

$\hat{\beta}_5$ coefficient represents the average difference in log odds of voting for the NDP between voters with household income from $50,000$\$ to $74,999$\$ and voters with household income from $100,000$\$ to $124,999$\$ with a certain age and sex.

$\hat{\beta}_6$ coefficient represents the average difference in log odds of voting for the NDP between voters with household income from $75,000$\$ to $99,999$\$ and voters with household income from $100,000$\$ to $124,999$\$ with a certain age and sex.

$\hat{\beta}_7$ coefficient represents the average difference in log odds of voting for the NDP between voters with household income less than $25,000$\$ and voters with household income from $100,000$\$ to $124,999$\$ with a certain age and sex.

x_{Age} equals a numerical value that represents the age of the voter.

$x_{Male} = 1$ if the voter is Male, and if the voter is Female, then $x_{Male} = 0$.

$x_{Income>125k} = 1$ if the voter has a household income greater than $125,000$\$, else $x_{Income>125k} = 0$

$x_{Income=25k-50k} = 1$ if the voter has a household income from $25,000$\$ to $49,999$\$, else $x_{Income=25k-50k} = 0$

$x_{Income=50k-75k} = 1$ if the voter has a household income from $50,000$\$ to $74,999$\$, else $x_{Income=50k-75k=0}$

$x_{Income=75k-100k} = 1$ if the voter has a household income from $75,000$\$ to $99,999$\$, else $x_{Income=75k-100k} = 0$

$x_{Income<25k} = 1$ if the voter has a household income less than $25,000$\$, else $x_{Income<25k} = 0$

Also, for this logistic regression model, there are four assumptions to check for:

1. Outcome is binary

Because the response variable for this model is to predict whether or not a voter would vote for the NDP, so the outcome of this model would be either "vote" or "do not vote", so this assumption is satisfied.

2. linearity in the logit for continuous variables

From this logistic regression model, similarly, there is only one continuous variable, and that is "Age". To check for the linearity of the variable "Age" in this logistic regression model, we set up the Box-Tidwell Test like what we did in the first model to test whether the logit transformation is a linear function of the predictor "Age".

From Box-Tidwell, H_0 : linearity between Age and log odds vs. H_A : non-linearity between Age and log odds.

The resulting P-value from the Box-Tidwell test for the NDP model is 0.0240118, which shows moderate evidence to reject the null hypothesis, implying that the linearity assumption might be violated for the NDP model. Thus, this might remain as a limitation to our model for predicting the outcome for the NDP Party.

The Box-Tidwell test is run in R using the package "car" (Fox & Weisberg, 2019).

3. Absence of multicollinearity

To check whether or not the predictors in our sample data are highly correlated, we would also compute the variance inflation factors from our dataset for this model to help us identify multicollinearity (James et al., 2017).

From table 7, we can see that none of our predictors have variance inflation factors exceeding 5, which means that none of them seem to be correlated to each other in this model, so the multicollinearity assumption would be satisfied.

The Variance Inflation Factor is computed in R using the package "car" (Fox & Weisberg, 2019).

Table 7: Variance Inflation Factors of Logistic Regression Model for the NDP

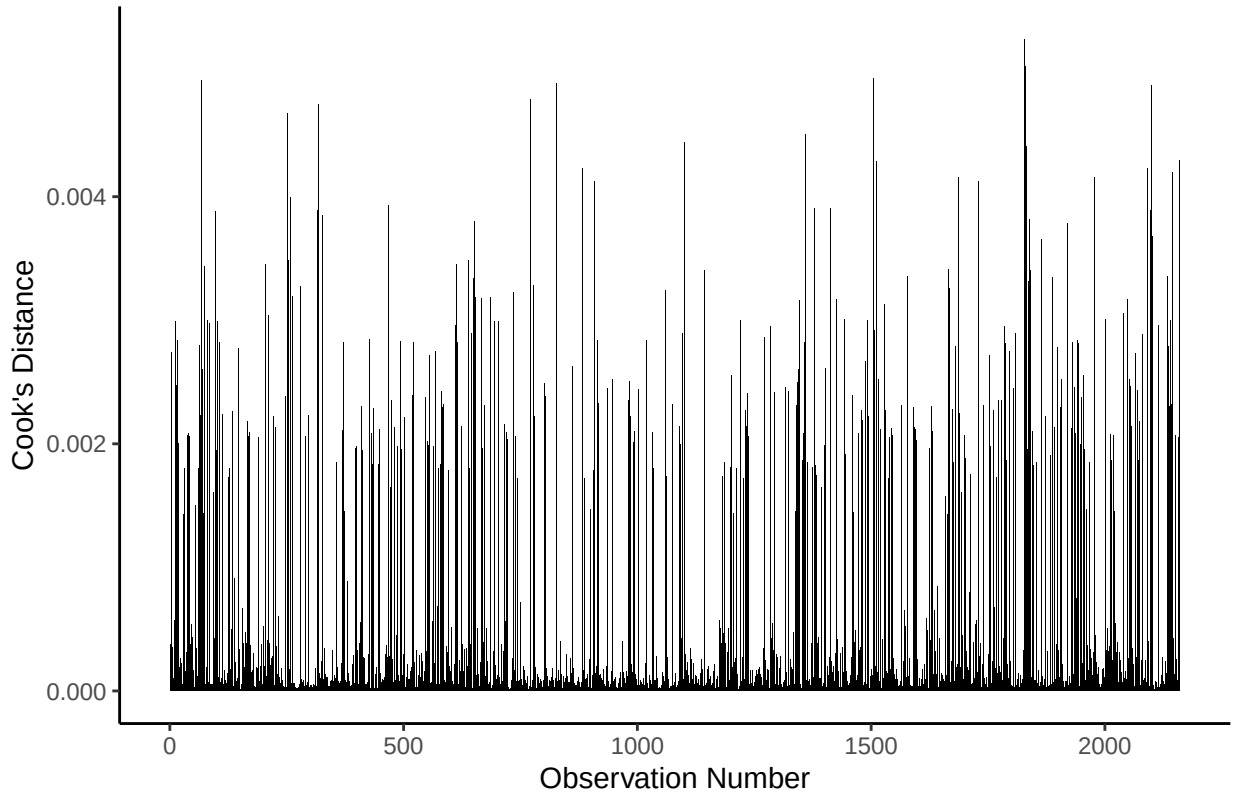
	Variance Inflation Factors	Degree of Freedom
Age	1.009671	1
Sex	1.012920	1
Income	1.010848	5

4. Checking influential values

To check for the influential values that could affect the quality of this model, we would examine these values by visualizing Cook's distance values for each of the models (Kassambara, 2018).

By observing graph Fig 6, we can see that there are many observations with large Cook's Distance, so we need to decide whether or not the Cook's distance of these observations is large enough to consider such observation as influential. According to Eberly College of Science, an observation with Cook's distance over 0.5 may be considered influential (Pardoe et al., 2018). According to our plot, there is one observation with an observation number around 1500 that exceeds 0.5, so this observation may or may not be influential to affect the quality of this model, so this could possibly be a limitation to our model.

Fig 6. Cook's Distance for Each Observation in NDP Model



Therefore, after checking all the assumptions, we now have the logistic regression model for predicting the outcome of the NDP using the survey data based on the predictors age, sex, and household income level.

Post-Stratification

In order to estimate the proportion of voters who would vote for the Liberal Party, Conservative Party, and the NDP, I would apply the post-stratification on each of the logistic models for the Liberal Party,

Conservative Party, and the NDP.

The Procedure of Post-Stratification for the Liberal Party:

1. We would partition the population using census data into "cells" based on different ages, sex, education level, and birthplace. age, sex, education level, household income level, household size, and birthplace for the Conservative Party;).
2. We would estimate the proportion of voters in each cell using the Liberal Party model built from the previous section.
3. We would then weight these estimate from each cell by its relative population size and we add all of these weighted estimates together and divide it by the population size to get the post-stratified estimate of the proportion of voters who would vote for the Liberal Party.

The Procedure of Post-Stratification for the Conservative Party:

1. We would partition the population using census data into "cells" based on different ages, sex, education level, household income level, household size, and birthplace for the Conservative Party.
2. We would estimate the proportion of voters in each cell using the Conservative Party model built from the previous section.
3. We would then weigh these estimate from each cell by its relative population size and we add all of these weighted estimates together and divide it by the population size to get the post-stratified estimate of the proportion of voters who would vote for the Conservative Party.

The Procedure of Post-Stratification for the NDP:

1. We would partition the population using census data into "cells" based on different ages, sex, and household income level for the NDP.
2. We would estimate the proportion of voters in each cell using the NDP model built from the previous section.
3. We would then weigh these estimate from each cell by its relative population size and we add all of these weighted estimates together and divide it by the population size to get the post-stratified estimate of the proportion of voters who would vote for the NDP.

$$\hat{y}^{PS} = \frac{\sum_{j=1}^{J} N_j \hat{y}_j}{\sum_{j=1}^{J} N_j}$$

$\hat{y}^{PS}$ indicates the outcome of interest, which is the estimate of the proportion of voters who would vote for the Liberal Party, Conservative Party, and NDP.

N_j represents the size of the j_{th} cell in the population.

$\hat{y}_j$ represents the estimate of the outcome in j_{th} cell from our logistic models.

J represents the total number of cells from all combinations of our predictors: Age, Education Level, Sex, Household Income Level, Household Size, and Place of Birth.

Therefore, by following the above procedures, we could arrive at the post-stratified estimates of the proportion of voters who would vote for the Liberal Party, Conservative Party, and NDP.

All analysis for this report was programmed using `R version 4.0.2`.

Results

By using 2025 Election prediction model, we have yielded the following results:

Table 8: Calculated Percentage of Votes for Each Party

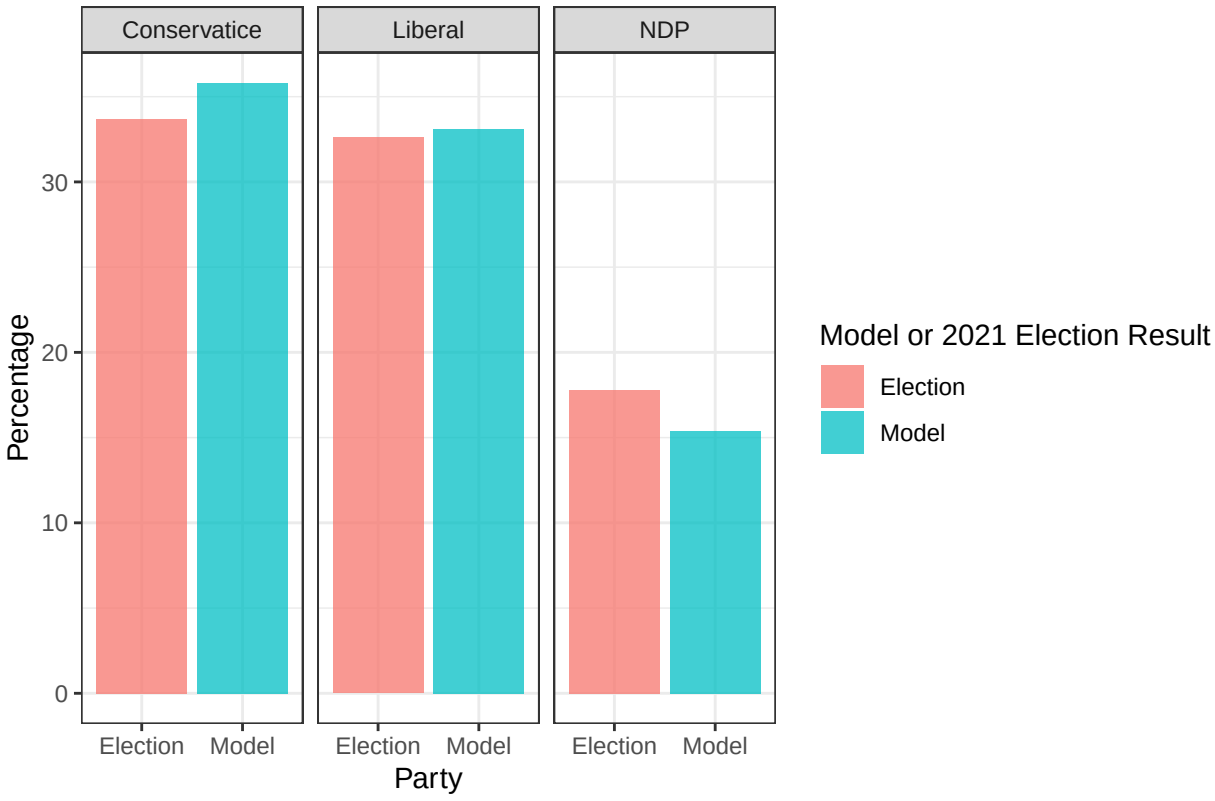
Liberal	Conservative	NDP
33.1	35.8	15.4

According to table 8, our model predicts the 2025 Election results of 33.1, 35.8 and 15.4, which corresponds to the Liberal Party, the Conservative Party and the NDP. These results indicate that out of all parties and votes, our model estimates the Liberal Party will have 33.1% of all votes, the Conservative Party will have 35.8% of all votes, and the NDP will have 15.4% of all votes. Our model predicts that the Conservative Party will have the highest votes, the Liberal Party has the second highest votes while the NDP has the least votes out of the three predicted parties.

Table 9: 2021 Canadian Election Results

Liberal	Conservative	NDP
32.6	33.7	17.8

Fig 7. Comparison of Our Prediction and 2021 Election Result



To assess whether our result seems reasonable or not, the official 2021 election data from the Canadian Government (CBC/Radio Canada, n.d.) is used for comparison. According to the official 2021 election (CBC/Radio Canada, n.d.), the Conservative Party won 33.70% of total votes, the Liberal Party won 32.60% of total votes and the NDP won 17.8%. The above bar plot (Figure 7) shows the results between our prediction model and the 2021 Election results, where the x-axis represent different parties and the y-axis represents the percentage of total votes; the bar in red represents the 2021 Election results, while the bar in blue represents our model's 2025 prediction results. Our model predicts a very similar percentage with error

margins of only $0.5 \sim 2.1$ percentage points. The 2021 results also share the same trend as our model – the Conservative Party has the highest votes, the Liberal Party has the second highest votes while the NDP has the lowest votes out of all three.

Although our goal is to predict 2025 election results, the data we used to construct this model are from before 2020. Therefore, the 2021 election result provides a reference on relatively how well our model might perform in predicting the 2025 election. However, if a more updated and more comprehensive survey and census data is given, we could arrive at a more accurate prediction for the 2025 election.

Conclusions

To restate our hypothesis: Based on the outcome of the 2021 election (CBC/Radio Canada, n.d.), we predict the Conservative party will yield the highest popular vote among the three parties although the Conservative party lost the election due to winning fewer ridings than the Liberal party. To predict the popular votes for the Liberal, Conservative, and NDP, we constructed a logistic regression model for each of these parties using the survey data. Utilizing these logistic regression models, we further used the post-stratification technique and the census data to obtain weighted estimates for the proportion of votes for each of the parties. Our model concludes the following results for Canada's 2025 Election: the Conservative party will win the most votes 35.8% of total votes, the Liberal Party will win 33.1% of total votes and the NDP will win 15.4% of total votes. These results are in alignment with our hypothesis: the Conservative party yields the highest popular vote among the three parties.

Nonetheless, we recognize that several limitations exist in our analysis. (1) The predictors we chose to predict the voting result may not fully explain all the influential factors of the election. Although predictors are selected for each model through the AIC, further analysis could integrate other varieties of model selection algorithms that check more thoroughly when deciding predictors. (2) Although we have checked the assumptions for all three models constructed, some of the assumptions for our models are shown with moderate evidence that they may be violated, so the quality and prediction capability of our models might be somewhat compromised. In future analyses, we would look into methods that could help alleviate such possible violations and improve our models. (3) Our dataset is from 2019 (survey data) and 2017 (census data). Thus, our forecasts are made far before the election day (tentatively 2025) and if voters' opinion changes significantly throughout the years, such predictions may not be representative and accurately reflect voters' decisions. Consequently, future analyses could implement the model on newer datasets to produce a more updated prediction for the next federal election.

Bibliography

Anderson, C. D., & Stephenson, L. B. (Eds.). (2010). *Voting behaviour in Canada.* UBC Press.

Andrew Heywood, Key Concepts in Politics and International Relations (2d ed.: Palgrave Macmillan, 2015), p. 119.

Bélanger, É., & Godbout, J.-F. (2010). Forecasting Canadian Federal Elections. *PS: Political Science and Politics*, 43(4), 691–699. http://www.jstor.org/stable/40927037

CBC/Radio Canada. (n.d.). *Federal election 2021 live results.* CBCnews. Retrieved November 28, 2022, from https://newsinteractives.cbc.ca/elections/federal/2021/results/

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression.* SAGE.

Government of Canada, S. C. (2019, February 6). *General Social Survey - Family (GSS).* Surveys and statistical programs. Retrieved November 30, 2022, from https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=335816

James, G., Daniela, W., Trevor, H., & Robert, T. (2017). *An introduction to statistical learning: With application in R* (8th ed.). Springer.

Kassambara, A. (2018). *Machine learning essentials: Practical guide in R.* CreateSpace Independent Publishing Platform.

Kaushik, S. (2017). *Voting Behaviour by Age and Generation: A Study of Canadian Elections from 1965 to 2015* (Doctoral dissertation, Queen's University).

Lewis-Beck, M. S. (2005). Election Forecasting: Principles and Practice. *The British Journal of Politics and International Relations*, 7(2), 145–164. https://doi.org/10.1111/j.1467-856X.2005.00178.x

McGrane, David. "Socio-Economic Determinants of Voting Behaviour in Canadian Provincial Elections from 1988 to 2006." *Presentation, Canadian Political Science Association Annual Conference, Saskatoon, Saskatchewan, May.* Vol. 30. 2007.

Pardoe, I., Simon , L., & Young, D. (n.d.). 9.5 - *identifying influential data points.* 9.5 - Identifying Influential Data Points | STAT 462. Retrieved December 1, 2022, from https://online.stat.psu.edu/stat462/node/173/

Presentation-Ready Data Summary and Analytic Result Tables [R package gtsummary version 1.6.2]. (2022). https://CRAN.R-project.org/package=gtsummary

Stephenson, L. B., Harell, A., Rubenson, D., & Loewen, P. J. (2022, November 22). *2019 Canadian Election Study (CES) - phone survey.* Harvard Dataverse. Retrieved December 1, 2022, from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2F8RHLG1

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686