Model Prediction for 2025 Canadian Federal Election Liberal Party Vote Share

STA304 - Assignment 3

Group 20: Yutong Lu-1005738356 & Zichen Gong-1005682469

November 5, 2021

Introduction

The data used in this report is from Canadian Election Study (2019) and General Social Survey (2017), which contains the phone survey information about Canadian voter preferences and demographics, as well as Canadian census data, respectively.

We aim to predict the overall popular choice among voters in the next Canadian federal election in 2025. This is very important because, through the phone survey, we may learn voter perceptions and attitudes towards the election, which are usually not observable or not available to the administrations [1]. Based on this data, we can get a better idea of Canadian voters' election dynamics and political behaviours. Also, the Canadian parliamentary system is based on the Westminster system, a type of democracy that stems from British tradition [24]. According to Bélanger and Godbout [5], the successful predictions in Westminster-style parliamentary systems in other countries such as the United Kingdom may infer that we may develop a reliable model in terms of election prediction, since in the Canadian political system there are many important similarities to those countries. This demonstrates that we may expect our model to have the predictive ability to forecast the potential outcomes in a future election to some extent. Moreover, past prediction shows that the incumbent party candidate does better during times with strong economic growth and high ratings of the current president [9]. As a result, another significance of this prediction model is that we could learn from the factors involved in election prediction and gain some insights about under what political and economic conditions a party may gain or lose public favor, resulting in better or worse performance in the election. Therefore, we may build a prediction model to forecast the popular vote in the upcoming 2025 Canadian federal election based on the historical phone survey data and census data.

Many factors may be related to individual political behaviours. Bittner and Goodyear-Grant [6] argued that gender identity structures the political attitudes and that the extent an individual identifies with their gender identity has consistent effects on their political behaviours. On the other hand, Nissanov [14] analyzed Israeli voters by income groups in four parliamentary elections and reported that higher-income groups are less loyal to parties than lower-income groups. Education may also play a part in political attitudes, where Osborne and Sibley [15] were able to show that openness consistently predicts the political attitudes for individuals with higher education. In terms of ageing, Peterson et al. [17] found that Liberals are more likely to become Conservatives under circumstances where political views do change over a lifetime. Moreover, geographical factors may also be related to political behaviours. According to Scala and Johnson [20], their regression model showed a significant relationship between the urban-rural continuum and the voting pattern in the U.S. presidential elections.

As a result, the research question of this report is that what impacts age, gender, province, education, and total household income can have on the Liberal Party vote share, and how we can use them to predict the popular vote in 2025 Canadian federal election.

Based on the literature above, we hypothesize that individuals from different age groups have different political attitudes. Moreover, we hypothesize that there is a significant relationship between age, gender, province, education, total household income, and the proportion of votes for the Liberal Party.

There are 8 sections in this report, including the introduction and bibliography. In the Data section, we will introduce the information about data context, data collection process and our data cleaning process. We will also introduce the important variables in our report and demonstrate their features with numerical and plot summaries. Then, we will introduce our model construction, method for model selection in the Method section and our post-stratification process in the Post-stratification section. Then, we will report our results in the Results section. Finally, we will address any limitations, future works and discussion in the Conclusion section. Any plots that are relevant to our report will be displayed in the Appendix.

Data

Data introduction, context and data collection process

The first data set used in this report is a survey conducted by Canadian Election Study in 2019 [21]. It is a data set composed of information from a phone survey during the campaign period and after the federal election in 2019 to gather Canadians' attitudes towards the election. In the raw data, there are 4021 observations and 278 variables in this data set.

During the data collection process, this survey was conducted by Advanis Inc. and covered all provinces in Canada [22]. The target respondents were Canadian permanent residents and citizens that are 18 years old and older. Two periods of data collection were conducted. The first one was during the pre-election campaign, which was between September 10th, 2019, and October 20th, 2019. The second post-election data collection period was from October 22nd, 2019 and November 21st, 2019. This survey used the computer assisted telephone interviewing method (CATI), where interviewers contacted the respondents via telephone and followed a guided questionnaire on a computer [26]. Among all samples collected using CATI, 66% of the sample were wireless telephone numbers, and the remaining 34% were landline telephone numbers [22]. Timing and the number of call-backs were monitored to achieve a higher response rate, aiming to obtain a more representative sample [22].

The second data set in this report is a General Social Survey on families, cycle 31 from 2017 [7]. Data are provided by Statistics Canada and are retrieved from the online database maintained by Computing in the Humanities and Social Sciences (CHASS) at the University of Toronto.

The data collection process of the 2017 General Social Survey was via computer-assisted telephone interviews, where all the interviews took place in the regional offices of Statistics Canada [7]. The household member was randomly selected to complete the interview and was encouraged to respond if they refused at first [7]. In the situations where the timing was not convenient for an interview, or no one responded to the call, alternative appointments or numerous call-backs were made [7]. During the 2017 General Social Survey, the overall response rate was 52.4% [7].

The 2019 phone survey data set from Canadian Election Study is obtained using `R` package `cesR` [11] by searching the data set name [10].

Data cleaning process

In the data cleaning process, we first selected the variables citizenship status, age, sex, province, education, and family income in the census data. After inspection, we found out that the variable type of age in our data set was character, and we changed the variable type of age to numeric because the variable age is numerical in nature. To keep the eligible voting population of the census data, we filtered the observations with age greater or equal to 18 and have a citizenship status that is either "by birth" or "by naturalization". Also, to obtain a complete case data set without missing values in any of the variables, we omitted all the observations with missing values in the census data set.

Then for the survey data set, we first selected the desirable variables corresponding to the questions of interest: question 1, age, question 3, question 4, question 10, question 11, question 61, and question 69. For clarity, we then renamed the question variables to their content, which are citizenship, age, gender, province, certainty, partisanship, highest education, and income. These variables are all categorical variables with levels indicated by numbers, so we then created new variables where all the numbered levels were mapped into their actual category names.

For the variable citizenship, we created two levels, "Yes" and "No" to denote whether an individual is a Canadian citizen.

In the variable gender, we created three levels "Male", "Female", and "Other".

In the province variable, we converted all the numbers from 1 to 13 into "Newfoundland and Labrador", "Prince Edward Island", "Nova Scotia", "New Brunswick", "Quebec", "Ontario", "Manitoba", "Saskatchewan", "Alberta", "British Columbia", "Northwest Territories", "Yukon", and "Nunavut".

Then for the variable certainty, we created 5 levels "Certain", "Likely", "Unlikely", "Certain not to vote", and "Already voted in advanced poll" to denote how certain an individual would like to vote.

The variable partisanship was created to have 6 categories, which are "Liberal", "Conservatives", "NDP", "Bloc Québécois", "Green Party", and "People's Party".

Finally, a categorical variable highest education was created with 11 categories denoting different levels of educations, including "No schooling", "Some elementary school", "Completed elementary school", "Some secondary/high school", "Completed secondary/high school", "Some technical, community college, CEGEP, College Classique", "Completed technical, community college, CEGEP, College Classique", "Some university", "Bachelor's degree", "Master's degree", and "Profession degree or doctorate".

For all the variables discussed above, all non-informative values, such as "Don't know", "Refused" or "Skipped", were interpreted and categorized as missing values. For the numerical variable income, if the recorded category was -8 or -9, then the observation with either one of these two categories of incomes was categorized as having a missing value for income.

The age variable in the original survey data set was character as well, so we then changed the type of age to numerical for future manipulations. Because we wanted to keep only the individuals who will be eligible to vote and willing to vote in the 2025 federal election, we filtered the data set to only have the observations that are Canadian citizens, age greater or equal to 18, and have a voting certainty that is not "Certain not to vote". Then, because we already set all the non-informative values to missing values, we filtered the data set again to only keep the observations with no missing values in any of the variables discussed above.

In the report, we are interested in the proportion of votes for the Liberal Party. Therefore, depending on whether the value of the variable partisanship is "Liberal" or not, we created a new variable named **Liberal votes** to be a categorical variable that can take either "Yes" or "No" as a value.

Because we wanted to use the census data to help us reweigh the survey data results, we then created a categorical variable **age status** in both data sets. The variable factors in both data sets after mapping are the same. It is a variable that categorizes the age variable in either data set by turning the numerical age into age groups. For age below or equal to 20, we categorized it as "20 or less". For age greater than 20 and smaller than or equal to 30, it is categorized as "21 to 30". For age greater than 30 and smaller than or equal to 40, we categorized it as "31 to 40". For age greater than 40 and smaller than or equal to 50, we categorized it as "41 to 50". For age greater than 50 and smaller than or equal to 60, we categorized it as "51 to 60". For age greater than 60, we categorized it as "above 60".

We wanted to map the variable **gender** in the survey data to the variable sex in the census data. After an inspection, we found out that after the cleaning above, in the survey data, there are 1302 individuals identified as male, 875 individuals identified as female, and only one individual identified as non-binary. Note that in the original, raw, uncleaned survey data set, there are no observations that refused or skipped the question, and there is only one individual identified as non-binary, or "Other", as their gender. Because there is only one observation in both cleaned and raw data sets, we decided to impute the only non-binary

observation as female. This is because we assumed the oppression faced by individuals identified as non-binary is at least to the same extent as the oppression faced by individuals identified as female [13]. By this approach, we recognized that the category "Female" in our study is rather a representation of "non-male". In the census data set, there are two categories, "Male" and "Female" for the variable sex. Because of the very low proportion of non-binary individuals in the survey data, we decided to create a new variable gender in the census data set, mapping sex directly to gender.

We also recognized that the census and survey data sets have different categories for **education** levels, so we wanted to create two education variables in the survey and census data sets respectively that have the same categories. Firstly, in the survey data set, we created a new variable named education with six categories. Individuals who completed some or all the elementary school are categorized into "Less than high school diploma". Individuals who completed some or all the secondary or high school are categorized into "High school diploma". Individuals who completed some university are categorized as "University diploma below the Bachelor's level". Individuals who completed some or all technical, community college, CEGEP, and College Classique" are categorized as "College, CEGEP or other non-university diploma". Observations with a bachelor's degree are categorized as "Bachelor's degree". Finally, individuals with a master's degree, professional degree or doctorate are categorized as "University degree above the bachelor's level". Similarly, in the census data set, we created another education variable with the same six categories, depending on the original education level in the census. Individuals with less than high school diploma or its equivalent are categorized as "Less than high school diploma". Individuals with a high school diploma or a high school equivalency certificate are categorized into "High school diploma". Observations with university certificate or diploma below the bachelor's level are categorized as "University diploma below the bachelor's level". Individuals with trade certificate or diploma, college, CEGEP or other non-university certificate or diploma are categorized as "College, CEGEP or other non-university diploma". Observations with a bachelor's degree (e.g. B.A., B.Sc., LL.B.) are categorized as "Bachelor's degree". Finally, people with university certificate, diploma or degree above the bachelor's degree are categorized as "University degree above the bachelor's level".

Finally, we wanted to use the different **family income** levels in the census data set to categorize the numerical income variable in the survey data set. Note that in the raw survey data set, the income variable represents the total household income in 2018 to the nearest thousand dollars. Also, during the data collection, any income values less than 500 were recorded as 0, and any income values between 500 and 999 were recorded as 1,000. We created a categorical variable family income in the survey data set that can take different income levels as values, depending on the numerical income. For individuals with income less than 25,000, they are categorized as "Less than $25,000". Individuals with income higher than or equal to 25,000 and less than 50,000 are categorized as "$25,000 to $49,999". Individuals with income higher than or equal to 50,000 and less than 75,000 are categorized as "$50,000 to $74,999". Individuals with income greater than or equal to 75,000 and less than 10,0000 are categorized as "$75,000 to $99,999". Individuals with income higher than or equal to 10,0000 and less than 12,5000 are categorized as "$100,000 to $124,999". Individuals with income higher than or equal to 12,5000 are categorized as "$125,000 and more".

Lastly, because all the variables in the cleaned survey and census data set are categorical, we changed the variable type of all the variables from character to factor. We then selected the necessary variables in both data sets. For the survey data set, we selected Liberal votes, age status, gender, province, education and family income. For the census data set, we selected age status, gender, province, education and family income. In the final cleaned data sets, there are 2178 observations and 6 variables in the survey data set, and there are 18750 observations and 5 variables in the census data set.

Important variables

The important variables in the survey data set are Liberal votes, age status, gender, province, education and family income. In the census data set, the important variables are age status, gender, province, education and family income, which are similar to the survey data set. All the variables in the survey data set are factor variables. During the data cleaning process, we have already made the categories in these factor variables

the same. However, note that the survey data is from 2019 and the census data is from 2017, so there may be a difference in the year these variable values are from, otherwise they have similar meanings.

The variable **Liberal votes** can take either a value of "Yes" meaning voting for the Liberal Party, or "No" otherwise. This variable represents the political attitude of the individuals, and will be the response variable in our model. All the other variables are potential predictors in our model.

The variable **age status** has 6 levels corresponding to 6 categories of age, including "20 or less", "21 to 30", "31 to 40", "41 to 50", "51 to 60" and "above 60". Each observation can be in one of these 6 levels of age, depending on the reported age of this individual.

The variable **gender** is a binary variable that can take the value of either "Male" or "Female" depending on the observations' self-reported gender. Because we grouped the one individual in the survey data set who identified as non-binary into the "Female" category, "Female" group actually represents "non-male" as opposed to strictly the female gender [13].

The variable **province** reflects the province that a particular individual locates in. This variable has 13 levels corresponding to the 13 regions in Canada, including "Newfoundland and Labrador", "Prince Edward Island", "Nova Scotia", "New Brunswick", "Quebec", "Ontario", "Manitoba", "Saskatchewan", "Alberta", "British Columbia", "Northwest Territories", "Yukon", and "Nunavut".

The variable **education** is a factor variable that represents the individual's highest education level. It can take one of the categories including "Less than high school diploma", "High school diploma", "University diploma below the Bachelor's level", "College, CEGEP or other non-university diploma", "Bachelor's degree", and "University degree above the bachelor's level".

The last variable **family income** in the survey data set represents the total household income of the observation, including levels "Less than $25,000", "$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".

Numerical summaries

Table 1: Table with summary important survey variables

Variable	Category	Survey Proportion	Census Proportion
Liberal votes	No	66.4	-
	Yes	33.6	-
Age status	20 or less	1.4	1.3
	21 to 30	12.2	9.3
	31 to 40	16.9	15.2
	41 to 50	20.2	14.3
	51 to 60	18.9	19.1
	above 60	30.4	40.8
Gender	Female	40.2	54.7
	Male	59.8	45.3
Province	Alberta	7.0	8.3
	British Columbia	20.7	12.0
	Manitoba	7.2	5.7
	New Brunswick	4.5	6.7
	Newfoundland and Labrador	4.5	5.5
	Nova Scotia	5.1	7.2
	Ontario	21.3	27.0
	Prince Edward Island	4.8	3.5
	Quebec	17.9	18.8
	Saskatchewan	7.1	5.5
Education	Bachelor's degree	28.1	18.4
	College, CEGEP or other non-university diplom	a 27.1	31.0
	High school diploma	17.5	24.5
	Less than high school diploma	1.0	13.7
	University degree above the bachelor's level	17.4	8.8
	University diploma below the bachelor's level	8.8	3.6
Family income	$125,000 and more	14.0	22.9
	$100,000 to $124,999	0.0	10.7
	$75,000 to $99,999	0.0	14.3
	$50,000 to $74,999	19.2	18.0
	$25,000 to $49,999	23.0	21.1
	Less than $25,000	43.7	13.0
Total cases		2178	18750

Information in the Table 1 is obtained using `R` package `expss` [8].

Table 1 is a summary of all the variables in the survey and census data set. For each variable, Table 1 reports the proportion of each category, and compares the proportion in the survey and census population. This comparison is based on the fact that we have made all the categories of these variables the same in the two data sets.

From the table, we can see that the "No" category in our response variable Liberal votes is 66.4%, which is approximately the double as the proportion of "Yes", which is 33.6%. This means that 1/3 of the surveyed population wishes to vote for the Liberal Party, which may be used to compare with our final predicted result.

Also, even though the proportion of people identified with the female gender in the census data is higher than the survey data, the two distributions both display a proportion of approximately 50% and 50% for the male and female gender.

In the distributions for province, 20.7% of the individuals in surveyed population are in the British Columbia, but the proportion of individuals in British Columbia is only 12.0% of the census population, resulting in a relative large difference in the province distributions of these two populations.

One noticeable difference between the two data sets' proportion is in the variable education. We can see that the proportion of individuals with less than a high school diploma in the surveyed population is only 1.0%, but in the census data set this proportion is 13.7%, which is much higher. Also, the surveyed population have 28.1% people with bachelor's degrees, whereas only 18.4% of the census population have a bachelor's degree, indicating the education distribution differences between these two data sets. Both data sets do not have individuals from provinces Northwest Territories, Yukon and Nunavut. This may either because they do not take part in the census or the survey process, or they are not easily available in terms of political or demographic surveys.

Another noticeable difference between the survey and census proportion is in the family income variable. In the survey data set, there is no individual from the income levels of \$75,000 to \$99,999 and \$100,000 to \$124,999, whereas in the census population, people who are in either of the \$75,000 to \$99,999 and \$100,000 to \$124,999 categories take up $14.3 + 10.7 = 25\%$, or $1/4$ of the total population. Also, we can see that the proportion of people with less than \$25,000 household income is 43.7% in the survey data set but 13.0% in the census data set, resulting in a difference of 30.7%. This means that the distributions for total household income in the surveyed and census population are very different.

Plot summaries

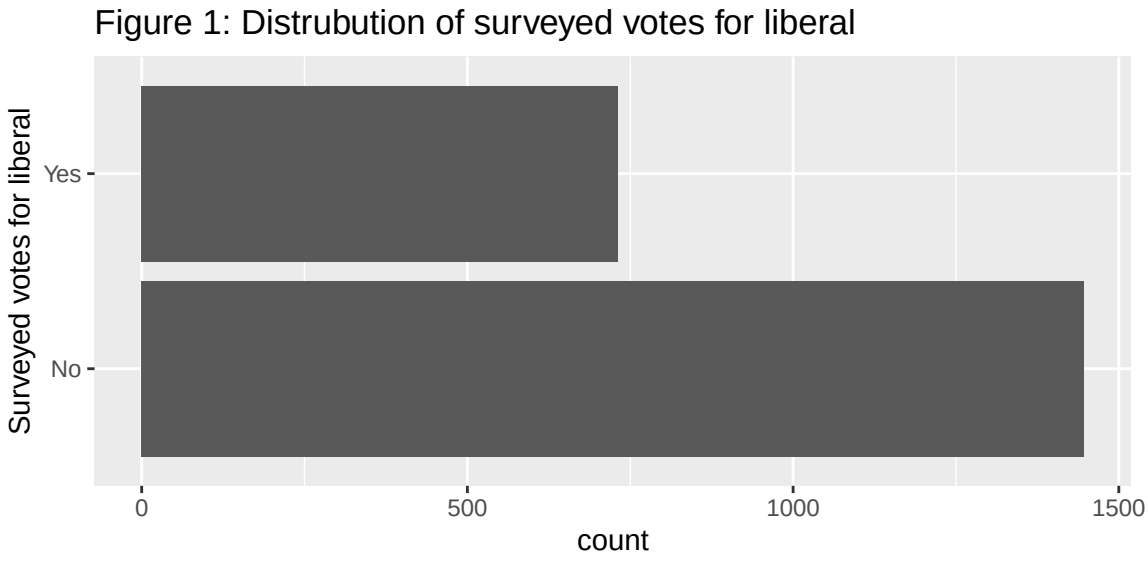

Figure 1 is a bar plot for the counts of whether an observation votes for the Liberal Party in the survey data. From the plot, we can see that the number of individuals that do not vote for the Liberal Party is approximately the double as the amount of people who vote for Liberals. It reflects the proportion of individuals who vote for the Liberal Party, which is 0.336.

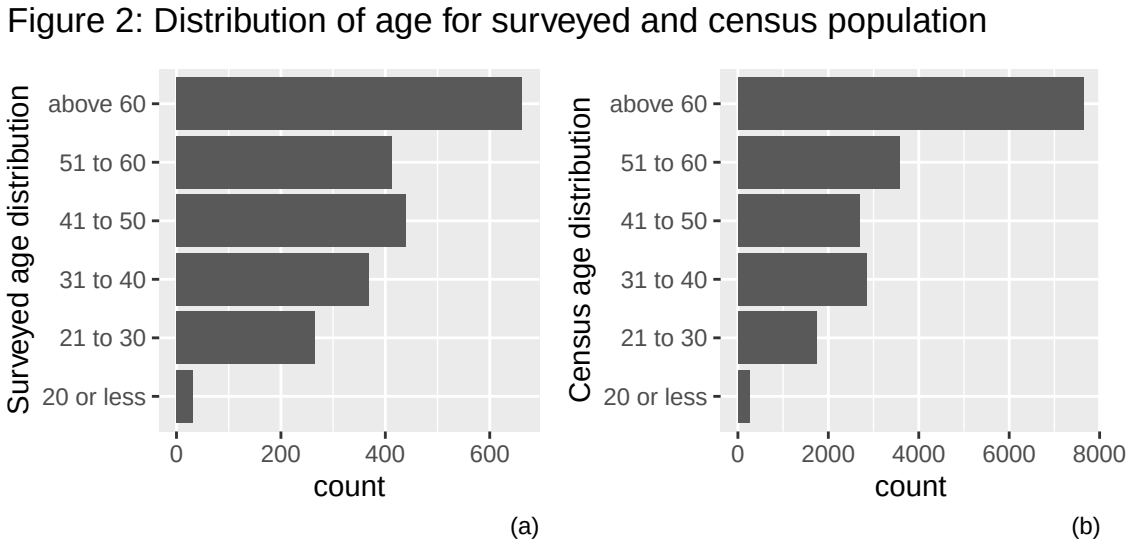

Figure 2 are bar plots of (a) age groups in the surveyed population and (b) census population. We can see from the plot (a) that the largest age group in the survey is age above 60, and the group with least people is age below 20. This is consistent with the target respondents of the phone survey, which are Canadian citizens above the age of 18 that are eligible to vote. Similarly, the plot (b) for census data shows that the

largest age group is age above 60 and the smallest group is for people aged 20 or less. However, we can see that people above 60 in the census population takes a higher proportion than the surveyed proportion, resulting in a non-similar distributions of age in these two populations and a potential re-weight later in the analysis.

Figure 3: Distribution of province for surveyed and census population

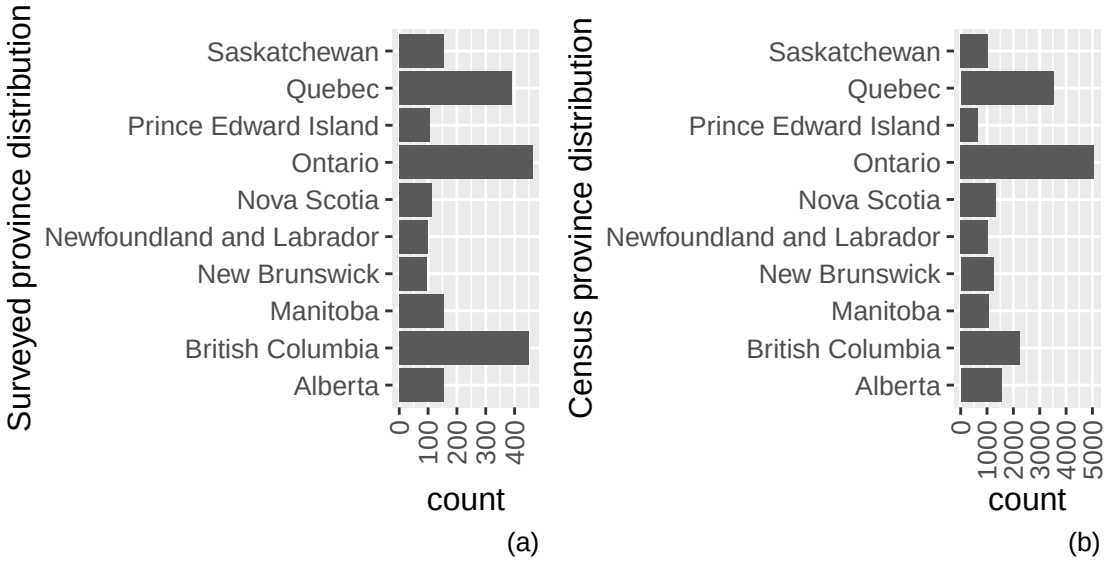

Figure 3 are bar plots of the distributions of provinces (a) in the surveyed population and (b) census population. We can see from the plot (a) that the province with greatest count in the survey is Ontario, followed by British Columbia and Quebec. Similarly, the plot (b) for census data shows that the province with the greatest number of respondents is Ontario. However, in the census population, the second largest province group is Quebec, and the third largest is British Columbia. This indicates that the distributions of province for surveyed and census population are not similar, resulting in a potential re-weight later in the analysis.

Figure 4: Distribution of education for surveyed and census population

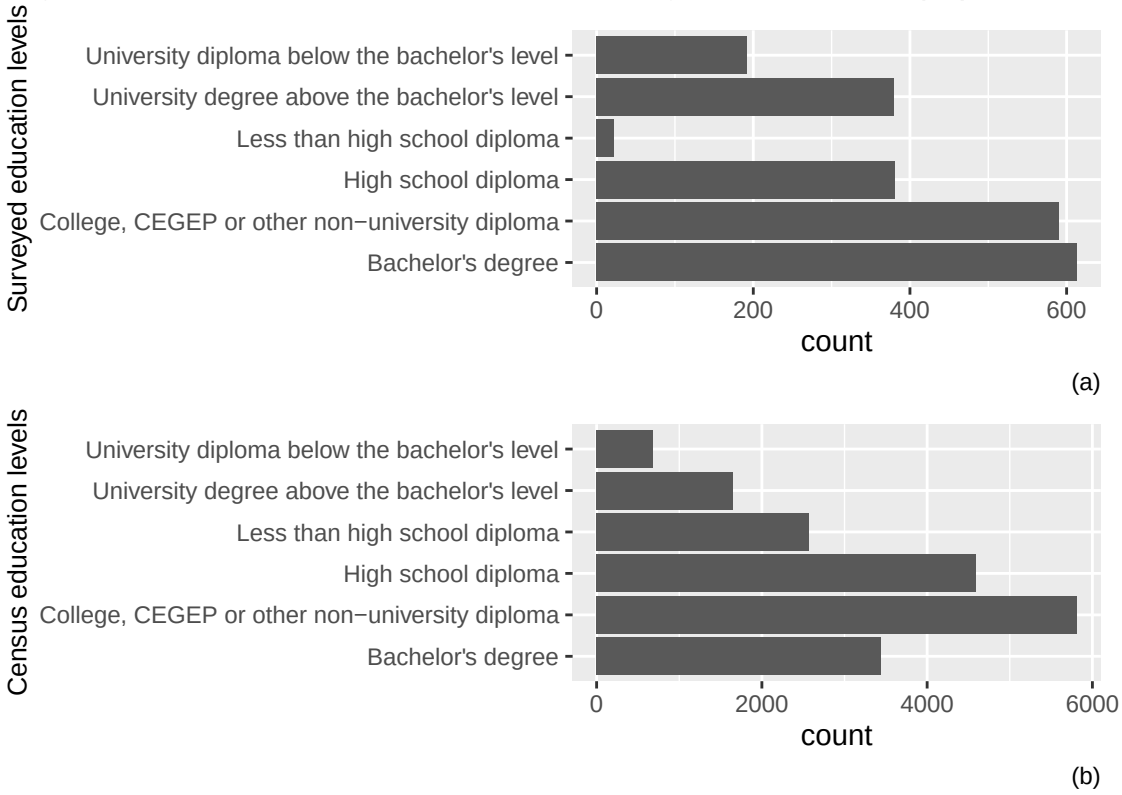

Figure 4 are bar plots of the distributions of education levels (a) in the surveyed population and (b) census population. We can see from the plot (a) that the group of people with bachelor's degree is the largest in the surveyed population, and the number of people with college, CEGEP or other non-university diploma is the second most. Conversely, the plot (b) for census data shows that the number of people with college, CEGEP or other non-university diploma is the greatest, followed by the number of people with a high school diploma. The group of people with bachelor's degrees is only the third largest group in the census data set. Also, the large difference in the length of the bars for people with less than high school diploma in the surveyed and census population corresponds to the large proportion difference in the numerical summary. As a result, we can see that the distribution of education for surveyed and census population have some large discrepancies, which may lead to a potential re-weight of results.

In this sections of this report, the plots are created using `R` packages `tidyverse` [27], `patchwork` [16], and `stringr`[28].

Methods

We will be using **Multilevel Regression (MR) model** and **Multilevel Regression Post-stratification (MRP)** for model construction. **Akaike's Information Criterion (AIC)** and **Bayesian Information Criterion (BIC)** will be used for model selection. *P-Value* will be used for a brief determination of statistical significance.

Model Specifics

The Simple Linear Regression (SLR) model is a one level model, denoted in the form

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$$

which follows Normal distribution with mean 0 and variance σ^2, per individual observation i. Here y_i represents the response variable regressed on a constant β_0 and an independent variable x_i with slope β_1.

We use the simplest case as a example, a two level MR for our reports. Based on our prior knowledge about SLR, the level one (individual level) model is denoted as

$$y_{ij} = \beta_0 + \beta_{1j} x_{ij} + \epsilon_{ij},\ \epsilon_{ij} \sim N(0, \sigma_e^2)$$

which follows Normal distribution with mean 0 and variance σ_e^2.

Here y_{ij} is the response variable regressed on a constant β_{0j} and an independent variable x_{ij} with slope β_{ij}. The level two (group level) information is represented by

$$\beta_{0j} = \alpha + a_j,\ a_j \sim N(0, \sigma_a^2)$$

$$\beta_{1j} = \beta + b_j,\ b_j \sim N(0, \sigma_b^2)$$

per individual observation i and group j.

There are three types of MR models. We use this model when we find out that the distribution of data points clusters into groups.

The first is **random intercept**. Each group trend has the same rate of change but their intercepts are different. For this type, intercepts vary while slopes are fixed. This model is denoted as

$$y_{ij} = (\alpha + a_j) + \beta_1 x_{ij} + \epsilon_{ij},\ \epsilon_{ij} \sim N(0.\sigma_e^2),\ a_j \sim N(0, \sigma_a^2)$$

α stands for the baseline intercept, while a_j represents the impact of group level variable and follows $N(0, \sigma_a^2)$ distribution. β_1 is the fixed slope for x_{ij}. ϵ_{ij} is the error term following $N(0, \sigma_e^2)$ normal distribution.

The second is **random coefficient**. Group trends cross each other but their intercepts are the same. For this type, intercepts are fixed while slopes vary. This model is denoted as

$$y_{ij} = \beta_0 + (\beta + b_j) x_{ij} + \epsilon_{ij},\ \epsilon \sim N(0, \sigma_e^2),\ b_j \sim N(0, \sigma_b^2)$$

β_0 is the fixed intercept. β stands for the baseline slope, while b_j represents the impact of group level variable and follows $N(0, \sigma_b^2)$ distribution. ϵ_{ij} is the error term following $N(0, \sigma_e^2)$ normal distribution.

The third type is a combination of the first two, which contains **both random intercept and random coefficient**. Group trends cross each other and their intercepts are different as well. Both intercepts and slopes vary. This model is denoted as

$$y_{ij} = (\alpha + a_j) + (\beta + b_j) x_{ij} + \epsilon_{ij},\ \epsilon_{ij} \sim N(0, \sigma_e^2),\ a_j \sim N(0, \sigma_a^2),\ b_j \sim N(0, \sigma_b^2)$$

α is the baseline intercept, and a_j represents the impact of group level variable on intercepts and follows $N(0, \sigma_a^2)$ distribution. β stands for the baseline slope, while b_j represents the impact of group level variable on slopes and follows $N(0, \sigma_b^2)$ distribution. ϵ_{ij} is the error term following $N(0, \sigma_e^2)$ normal distribution.

For this report, we will be using the first **random intercept** model and the **Frequentist Approach**, which means we have no prior information about the distribution of our data set. Since our response variable is a dummy variable for which people will either choose to vote or not to vote for the Liberal Party, we will be using the **Multilevel Logistic Regression model** with the function `glmer()` from `lme4`[4] package. We will choose our level two variable, or the group-level variable, based on our research question and hypothesis.

So the format of our model will be

$$\hat{y_{ij}} = (\hat{\alpha} + \hat{a_j}) + \hat{\beta}_1 x_{1j} + ... + \hat{\beta}_n x_{nj}$$

Let p be the probability of voting for the Liberal Party for one individual i from group j,

$$\hat{y_{ij}} = log(\frac{p}{1-p})$$

$(\hat{\alpha} + \hat{a_j})$ is the estimate of intercept per group j, for which $\hat{\alpha}$ is the baseline intercept and $\hat{a_j}$ is the impact on intercept by the level 2 variable. $\hat{\beta}_1 ... \hat{\beta_n}$ are coefficients on independent variables $x_{1j}...x_{nj}$, representing the estimated rate of change for the response variable when $x_{ij}^{'}s$ change. For now, $x_{1j}...x_{nj}$ are factors of four variables: gender, province, education and total household income with j being different age groups. Also, they are dummy variables that can either take the value of 1 if this individual is in this particular factor category, or 0 otherwise. We also have a reference category that comes first in the alphabetical order, which is not shown in the $x_{1j}...x_{nj}$. For example, if we have two factors in the model, then there will be two reference categories, one for each factor, that will not appear in the model. However, when a observation has $x_{1j}...x_{nj} = 0$, it means that this particular observation is in both of the reference categories for two different factors. The final version of our model after model selection in this case would be in the form

$$\begin{aligned}
\hat{y_{ij}} = (\hat{\alpha} + \hat{\alpha_j}) &+ \hat{\beta}_1 \text{educationCollege, CEGEP or other non-university diploma} \\
&+ \hat{\beta}_2 \text{educationHigh school diploma} + \hat{\beta}_3 \text{educationLess than high school diploma} \\
&+ \hat{\beta}_4 \text{educationUniversity degree above the bachelor's level} \\
&+ \hat{\beta}_5 \text{educationUniversity diploma below the bachelor's level} \\
&+ \hat{\beta}_6 \text{provinceBritish Columbia} + \hat{\beta}_7 \text{provinceManitoba} + \hat{\beta}_8 \text{provinceNew Brunswic} \\
&+ \hat{\beta}_9 \text{provinceNewfoundland and Labrador} + \hat{\beta_{10}} \text{provinceNova Scotia} \\
&+ \hat{\beta_{11}} \text{provinceOntario} + \hat{\beta_{12}} \text{provincePrince Edward Island} + \hat{\beta_{13}} \text{provinceQuebec} \\
&+ \hat{\beta_{14}} \text{provinceSaskatchewan}
\end{aligned}$$

The **AIC** value measures the goodness of fit of the model and balances with a penalty term so that the model is not over complicated [19]. We can say that this measure contains two part: goodness and complexity.

A large log-likelihood value or a small negative log-likelihood is for measuring the goodness of fit.

The Maximum Likelihood Principle is to choose the parameter(s) of interest in a way that the data is most likely given a data set $x_1, x_2, ..., x_n$. For a data set $x_1, x_2, ..., x_n$ which are realizations of a random sample $X_1, X_2, ..., X_n$ from a distribution characterized by θ, and consider $X_i s$ are in discrete form and are independent of each other, the likelihood function is given by

$$\begin{aligned}
L(\theta) &= P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) \\
&= x_n|\theta) = p(x_1|\theta) \ \cdot \ p(x_2|\theta) \ \cdot \ ... \ \cdot \ p(x_n|\theta) \\
&= \prod_{i=1}^{n} p(x_i|\theta)
\end{aligned}$$

In this context, $p(x|\theta)$ is not equal to conditional probability. It is the notation for probability mass function, which is also commonly denoted as $p_\theta(x)$. The probability density function can be expressed as $f_\theta(x)$ or $f(x|\theta)$. Probability mass functions and probability density functions are for discrete random variables and

continuous random variables, respectively. The maximum likelihood estimate of the unknown parameter(s) θ is the realization $t = h(x_1, x_2, ..., x_n)$ that maximizes the likelihood function $L(\theta)$. Its corresponding random variable is called the maximum likelihood estimator of θ:

$$T = h(X_1, X_2, ..., X_n)$$

This is also denoted as $\hat{\theta}$.

For continuous random variable, let $\epsilon > 0$ be a random fixed and small number. We have $\hat{\theta}$ that maximizes the probability:

$$\begin{aligned} &P(x_1 - \epsilon \leq X_1 \leq x_1 + \epsilon, ..., x_n - \epsilon \leq X_n \leq x_n + \epsilon) \\ &= P(x_1 - \epsilon \leq X_1 \leq x_1 + \epsilon) \ \cdot \ ... \ \cdot \ P(x_n - \epsilon \leq X_n \leq x_n + \epsilon) \\ &\approx f(x_1|\theta) \ \cdot \ ... \ \cdot \ f(x_n|\theta) \ \cdot \ (2\epsilon)^n \end{aligned}$$

Since the final factor $(2\epsilon)^n$ does not affect $\hat{\theta}$, we have the likelihood function for a data set of continuous random variables:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

In most situations, log-likelihood function $l(\theta)$ is easier to use.

$$\begin{aligned} l(\theta) &= log(L(\theta)) \\ &= log(f(x_1, x_2, ..., x_n|\theta)) \\ &= \sum_{i=1}^{n} log(f(x_i|\theta)) \text{ if the } x_i^{'}s \text{ are independent} \end{aligned}$$

The reason is $y = log(x)$ is a monotonically increasing function. The value that maximizes the log-likelihood function will also maximizes the likelihood function.

Besides goodness measured by log-likelihood value, a penalty is implemented on the number of parameters being estimated in the goodness step.

The AIC is takes the form

$$AIC = -2[log(L(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p, \ \hat{\sigma}^2|Y)) - (p + 2)]$$

Our preferred model is the one with the smallest AIC. If the model fits better, the log-likelihood value will be larger. If there are many parameters under penalty, it takes away some of the log-likelihood. This helps ensure the trade-off that only when our model is substantially good, we can have a relatively complicated model.

The **BIC** is another criterion that was developed under the Bayesian paradigm [19]. It looks similar to AIC,

$$BIC = -2log(L(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p, \ \hat{\sigma}^2|Y)) + (p + 2)log(n)$$

As we can see that BIC model takes more severe penalty, BIC will favor simpler models compared to AIC. Similar to AIC, **the model with the smallest BIC is indicated as a better model.**

Since AIC and BIC measure different aspects for model fit and take different levels of penalty for complexity, we will use both of them at once. As AIC sometimes suggest over-fitted models while BIC models are sometimes over-simplified, there is always a trade-off between AIC and BIC. We will make our decisions based on not only AIC and BIC values, but also the emphasis of our model and our understanding of the real-world context.

Finally, after we select the model with the appropriate predictors, we will then fit our model and calculate p-values for each of the coefficients in our model. This p-value is based on the Student's T-test [23], which has a formula for the test statistic of

$$T = \frac{\hat{\beta}_i - \beta_i}{\sqrt{var[\hat{\beta}_i]}}$$

In this formula, $\hat{\beta}_i$ stands for the estimated coefficient in our model and β_i represents the β_i in the null hypothesis. The null hypothesis and alternative hypothesis in our report are

$$H_0 : \beta_i = 0$$

$$H_a : \beta_i \neq 0$$

In the null hypothesis, we have $\beta_i = 0$, which means that the coefficient for this particular variable is zero, indicating a lack of relationship between this independent variable and the response variable in the presence of other predictors. Then, we can calculate the probability of observing a value that is as or more extreme than our test statistic under the null hypothesis, which is the p-value. The formula for p-value is p-value $= P(|t_{df=n-p-1}| > T)$, where n represents the number of observations in our data and p represents the number of predictors in our model. The 1 represents the intercept of our model. Thus, the distribution here, $t_{df=n-p-1}$, is a T distribution with a degree of freedom of $n - p - 1$. If the p-value is small, then it means that it is not likely that we will observe a value as or more extreme than our test statistic under the null hypothesis, and thus our test statistic is not purely due to chance. After calculating the p-value, we will compare it with a significance level of $\alpha = 0.05$. If the p-value is smaller than 0.05, then we will reject the null hypothesis and we have evidence that there is a linear relationship between this particular predictor and the response in the presence of all other predictors.

In this section, model construction and model selection use functions from `R` packages `lme4`[4], `dplyr`[29] and `MuMIn`[3].

Post-Stratification

Post-Stratification is a commonly used statistic technique to correct difference between target population and study population. There are three steps to implement a post-stratification.

Step 1: Divide the target population into cells. For example, in this census data set we use three variables, variable age group with 6 factors, variable province with 10 factors and variable education with 6 factors, the total cells we have are $6 \cdot 10 \cdot 6 = 360$ cells. However, we do not have data point for every cell in the census dataset, so the number of cells that have data points allocated is 325 cells in total.

Step 2: Use our study population to estimate the response variable value per cell. For this report, we take our final model constructed based on the survey data to estimate the probability for voting the Liberal Party for each of the 325 cells in the census data.

Step 3: Multiply the estimated value of each cell with its relative proportion in the population, and aggregate each of these reweighed cells. This can be denoted through

$$\hat{y}^{ps} = \frac{\sum_j N_j \hat{y}_j}{\sum_j N_j}$$

y stands for the population parameter (μ, p, σ, etc.). $\hat{y}_j$ is the estimated value of parameter based on our sample data set. In this report, our sample data set is the survey data, and the estimated parameter is the probability of voting the Liberal Party. N_j is the target population size for group j. $\sum_j N_j$ is the total target population size. The overall result after this post-stratification denoted as $\hat{y}^{ps}$ is the probability of voting the Liberal Party for the census population estimated by our final model and adjusted by cell-level relative proportion of the target population.

Results

In this report, we hypothesized that different age groups will have different political attitudes, so we decided to make age group as the level 2, or the group-level variable in our multilevel logistic regression model. This is based on the assumption that as people age, many perspectives of their lives will change and these changes may lead to a change in their political attitudes. Therefore, different age groups will have a different intercept in our model.

Table 2: AIC table of different models

Education	Gender	Family income	Province	AIC
+	+	NA	+	2663.522
+	NA	NA	+	2663.627
+	NA	+	+	2667.414
+	+	+	+	2667.463
NA	+	NA	+	2690.951
NA	NA	NA	+	2692.222
NA	+	+	+	2693.967
NA	NA	+	+	2695.096
+	+	NA	NA	2737.098
+	NA	NA	NA	2737.558
+	+	+	NA	2740.390
+	NA	+	NA	2740.608
NA	+	NA	NA	2774.805
NA	NA	NA	NA	2776.875
NA	+	+	NA	2777.218
NA	NA	+	NA	2779.047

Table 3: BIC table of different models

Education	Gender	Family income	Province	BIC
+	NA	NA	+	2754.605
NA	NA	NA	+	2754.770
NA	+	NA	+	2759.185
+	+	NA	+	2760.187
NA	NA	+	+	2774.702
+	NA	+	+	2775.451
+	NA	NA	NA	2777.361
NA	+	+	+	2779.260
+	+	+	+	2781.186
+	+	NA	NA	2782.588
NA	NA	NA	NA	2788.247
NA	+	NA	NA	2791.864
+	NA	+	NA	2797.469
+	+	+	NA	2802.938
NA	NA	+	NA	2807.478
NA	+	+	NA	2811.335

With our level 2 variable age group, Table 2 and 3 report the AIC and BIC values for all possible models with our four potential predictors, including education, gender, family income and province, and the group-level

variable age group. The symbol "+" means that this variable is a predictor in the model, whereas "NA" means this particular variable is not included in the model. Both AIC and BIC tables are in ascending order, so models at the top of tables are better models. If we take a look at the AIC table, the best model is the model using independent variables indicating education, gender and province. The second best model is the model with independent variables indicating education and province. If we take a look the the BIC table, the best model is the model using education and province. The second best model is the model with one independent variable indicating province. Since the model with education and province appears to be the second best of AIC and the best of BIC, this is the final model we use after model selection. This model will both fit well and is not over complicated.

Table 4: Model coefficient esitmates

	Estimate	p-Value
(Intercept)	-1.5693694	0.0000000
educationCollege, CEGEP or other non-university diploma	-0.5708203	0.0000101
educationHigh school diploma	-0.4576259	0.0019109
educationLess than high school diploma	-0.3671895	0.4429284
educationUniversity degree above the bachelor's level	0.1830522	0.1801557
educationUniversity diploma below the bachelor's level	-0.2492431	0.1702178
provinceBritish Columbia	0.7538836	0.0032315
provinceManitoba	0.8236366	0.0050060
provinceNew Brunswick	1.2269272	0.0000984
provinceNewfoundland and Labrador	1.4865892	0.0000019
provinceNova Scotia	1.3493150	0.0000096
provinceOntario	1.4517618	0.0000000
provincePrince Edward Island	1.4372743	0.0000030
provinceQuebec	1.2608320	0.0000008
provinceSaskatchewan	0.0079959	0.9805664

The model output is indicated by the "Model coefficient estimates" table above. The p-value for the intercept is extremely small, meaning that this average intercept value is statistically significant and our choice of choosing age group as the level two random intercept variable is appropriate. When all other independent variables are equal to zero, the $\hat{y_{ij}} = log(\frac{p}{1-p}) = -1.5693694$.

For the categorical variable indicating education, the categories "College, CEGEP or other non-university diploma" and "High school diploma" are statistically significant with p values smaller than 0.01. This shows that when the highest level of education for one person is college, CEGEP or other non-university diploma, the model output $\hat{y_{ij}}$ decreases by 0.5708203 with all other predictors remain constant. When the highest level of education is high school diploma, the model output decreases by 0.4576259, when all other predictors are the same. Other categories "Less than high school diploma", "University degree above the bachelor's level" and "University diploma below the bachelor's level" all have p-values larger than 0.05, so we will not consider them as statistically significant. When the highest level of education is bachelor's degree, which is the reference category, values of other factors are all equal to zero and the model output will not change. This result could be explained by the report of Osborne and Sibley [15], where they reported a difference in the openness of political attitudes for individuals with higher and lower education.
For the categorical variable indicating province, except for the factor Saskatchewan with a very large p value, other factors are all statistically significant with p values larger than 0.05. As a result, when one person is from British Columbia, the model output increases by 0.7538836, holding all other predictors constant. When it is Manitoba, the output increases by 0.8236366, when all other predictor values are constant. When the province is New Brunswick, the output increases by 1.2269272, holding all other predictors constant. When the province is Newfoundland and Labrador, the output increases by 1.4865892, when all other predictors are constant. When the province is Nova Scotia, the output increases by 1.3493150, holding other predictors constant. When the province is Ontario, the output increases by 1.4517618, when all other

predictors are constant. When the province is Prince Edward Island, the output increases by 1.4372743, holding all other predictors the same. When the province is Quebec, the output increases by 1.2608320, when all other predictors are the constant. If the province is Alberta, which is the reference category so that it does not appear in the model, then values of other factors are all equal to zero and this will not cause the model output to change, holding all other predictors constant. Note that some of the provinces other than the reference province Alberta, such as Yukon, do not show up in our model, and this is because in the surveyed population there are no observations from these provinces. Overall, variables related to province are generally statistically significant. This result makes sense because different provinces have different levels of urbanization, and based on the report of Scala and Johnson [20], there tends to be significant relationship between the urban-rural regional pattern and and the popular vote in presidential elections.

Therefore, we failed to reject part of our hypothesis that there is a significant relationship between the predictors, province and education, and the proportion of votes for the Liberal Party.

Table 5: Estimated intercept for different age groups

Age group cell	Intercept
20 or less	-1.617806
21 to 30	-1.657328
31 to 40	-1.540764
41 to 50	-1.671215
51 to 60	-1.597707
above 60	-1.326983

Table 5 reports the different estimated intercepts for different age groups. Since we use random intercept MR model, we have different models for data points in different age groups. Each model has the same coefficients on independent variables, but intercepts are different. For data points under age group 20 or less, the intercept is -1.617806. If data points are under age group 21 to 30, the intercept is -1.657328. For the age group 31 to 40, the intercept will be -1.540764. For data points under age group 41 to 50, the intercept is -1.671215. For data points under age group above 60, the intercept is -1.326983. This group contains overall 6 cells. Its random effects have a standard deviation of 0.1536. From the results, we do not see a monotonic increase or decrease in the intercept as the age group becomes older. However, the different intercepts for different age groups make sense based on the report by Peterson et al. [17], where they reported that under certain circumstances, an individual's political view may change as they age. As a result, we fail to reject part of our hypothesis where we hypothesized that different age groups will have different political attitudes.

In post stratification, we use age group, province and education from census data as our cells and create 325 different groups. Although age group variable with 6 factors, province with 10 variables and education with 6 variables should create 360 groups, some groups do not have data points allocated, so the meaningful number of groups is 325.

Then we use our final model to predict the probability of voting for the Liberal Party in each of the 325 groups.

At last, we adjust our model (constructed using the survey data) outputs with sizes of cell groups from census data by its relative proportion in census. This gives our final result $\hat{y}^{ps} = 0.3336495$. This shows that our predicted probability of people voting for the Liberal Party in the 2025 election is approximately 33.36%, which is slightly lower than the original 33.6% in the surveyed proportion. This means that the Liberal Party is predicted to have over one third of the votes in the 2025 federal election, which is a large amount of vote considering that all the other parties will have to split the rest of the two thirds of the votes. This result makes sense because it is approximately the same as the proportion of votes for the Liberal Party in the 2021 federal election, which is 32.6% [25].

In this section, model construction and model selection use functions from `R` packages `tidyverse`[27], `lme4`[4], `dplyr`[29] and `kableExtra`[30].

Conclusions

In this report, our research question is what effects age, gender, province, education, and total yearly family income can have on the proportion of votes for Liberal Party, and how we can predict the popular vote in 2025 Canadian federal election using these factors. Because we are interested in the predictors mentioned above, we hypothesized that individuals from different age groups have different political attitudes. Simultaneously, based on this group-level variable of age group, there is a significant linear relationship between the predictors gender, province, education, total household income, and the response, which is the proportion of votes for Liberal party.

In the Method section, we introduced the approach of our model construction and selection. Our response of interest "Liberal votes" is a binary variable that can take either value "Yes" or "No" and we do not have any prior information about the data set, we decided to use frequentist multilevel logistic regression to construct our model. In the multilevel regression approach, we will be using the random intercept model, where we assume that each group has a different intercept but the same slope. Based on our hypothesis, we used age group as the group-level variable, where we assumed that different age groups have different intercepts in their corresponding models. After selecting the type of model, we used AIC as a measure for model goodness of fit and BIC as a measure of model simplicity to compare the different candidate models with various numbers of predictors and different AIC and BIC values [19]. Ideally, we wanted to select a model with the smallest AIC and BIC values. However, we recognized the trade-off between the goodness of fit and simplicity of the model, so we would select the model considering both criteria. After model selection, we fitted our model and calculated p-values for each coefficient using Student's T-test [23] to evaluate whether the predictors in our model are significantly related to our response, which is whether to vote for the Liberals or not.

Then in the Post-stratification section, we divided the census population into cells, calculated the response variable predicted by our model, multiplied the predicted value by the proportion of these cells in the population, and then added these reweighed values together to get a final proportion of votes for the Liberals.

Based on our hypothesis, we decided to use age groups as the level 2 variable in our multilevel logistic regression model. In the results, we first reported the AIC and BIC values for the candidate models with the group-level variable age group and potential predictors, including education, gender, family income and province. The model with predictors education and province has the second least AIC and the least BIC values, so we selected this model to be in our report. After fitting the model, we found that different age groups have different intercepts in their models, which means that we failed to reject our hypothesis that different age groups have different political attitudes. Also, we found that 8 out of 9 dummy variables for province categories are significantly related to our response, and 2 out of 5 dummy variables for education categories are significant, by comparing the corresponding p-values and a significance level of 0.05. This means that we have evidence that each of the two variables, province and education, is significantly related to the response variable in the presence of other predictors in the model. Therefore, we failed to reject the part of our hypothesis that province and education are significantly related to the response, votes for the Liberal Party.

After post-stratification, our model predicted that the final proportion of votes for the Liberal Party is 33.36% in the 2025 Canadian federal election. This means that using our model, we predicted that over one third of the population will vote for the Liberal Party in the upcoming election, which is a considerable number of votes as the other five parties will split the remaining two thirds of the votes. Also, this is relatively consistent with the 32.6% of the popular vote that the Liberal Party got in the 2021 federal election [25].

All analysis for this report was programmed using `R version 4.1.1` [18].

Limitations

Firstly, there is a time difference between the census and survey data sets. The survey data set with the political attitudes of the observations is from 2019, whereas the census data set is from 2017. The potential predictors in our data set, including age status, province, education, family income, and the identified gender,

are all prone to change with time. Also, because we are building the model on historical data that is in 2019 and trying to predict the results in 2025, there is a long period between the data and the future election as well. According to Jennings et al. [12], the closer forecast to the election, the more accurate it tends to be. Therefore, it is possible that our model may produce inaccurate predictions due to the long time difference between the data collection and the actual election in 2025.

On the other hand, the survey data was targeted at individuals aged 18 and above, and we only kept the individuals over 18 years old in our report. However, there will be more eligible voters that were not 18 years old when the survey or the census took place. As a result, the reproducibility of our model may not apply to the eligible voter population in the federal election in 2025.

Another limitation is that in the raw survey and census data sets, there are different categories for the education variable. During the data cleaning process, we arbitrarily combined and redefined the different education levels in the survey and census data sets so that they could have the variable education with the same categories. However, it is possible that we will produce different results if we have categorized the various education levels in a different way. As a result, this lack of education level standard may impact the predictive ability of our model.

Future works

In the future research, we can try to use different level 2, or group-level variables in our model, and compare their predicted results. In this report, we used age status as a group-level variable, but we can try to use province, education level, or other predictors as the level 2 variable in the future work. In this report, we use random intercept approach. In the future, we can try random coefficient or combine both random intercept and random coefficient in our model.

Moreover, in this report, we only focused on whether the respondents would like to vote for the Liberal Party, without considering the particular candidate of the party that year. There might be some features related to the candidates themselves that will likely to affect the popularity of the party. Based on the report by Armstrong and Graefe [2], it is possible to use the biographical index of the candidates to forecast the results of U.S. presidential election. As a result, we can consider the features of candidates as potential predictors in the future research.

Discussion

In this report, we predicted that the proportion of votes for the Liberals in the 2025 federal election is 33.36%. However, capturing the popular vote does not necessarily mean winning the final federal election. According to Tahirali and Hahn [25], the most popular vote may not be directly related to the final result if the party with a popular vote fails to win the greatest number of riding. Eventually, the parties need to transfer their popularity and support measured in the popularity of votes into the number of riding they won. Based on our model, people in different age groups tend to have different political attitudes, and education and province have a significant relationship with the political attitudes of the individuals. Therefore, by studying certain features such as the demographics in a region, the parties running the election may get a rough idea about their vote popularity in that specific region, and they may carefully plan out their strategies to not only get a higher proportion of popular vote but also lead more ridings in the electoral districts.

Bibliography

1. Alvarez, R. M., Adams-Cohen, N., Kim, S. S., & Li, Y. (2020). *Securing American Elections: How Data-Driven Election Monitoring Can Improve Our Democracy.* Cambridge: Cambridge University Press. doi:10.1017/9781108887359

2. Armstrong, J. S., & Graefe, A. (2010, January 5). *Predicting Elections from Biographical Information about Candidates.* https://faculty.wharton.upenn.edu/wp-content/uploads/2012/04/PollyBio58.pdf. (Last Accessed: October 29, 2021)

3. Bartoń, K. (2020). *MuMIn: Multi-Model Inference.* R package version 1.43.17.

4. Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01.

5. Bélanger, É., & Godbout, J. F. (2010). Forecasting Canadian Federal Elections. *PS, Political Science & Politics, 43*(4), 691–699. https://doi.org/10.1017/S1049096510001113

6. Bittner, A., & Goodyear-Grant, E. (2017). Digging Deeper into the Gender Gap: Gender Salience as a Moderating Factor in Political Attitudes. *Canadian Journal of Political Science, 50*(2), 559–578. https://doi.org/10.1017/S0008423917000270

7. Diversity and Sociocultural Statistics. (2020, April). *General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User's Guide.* CHASS. https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf. (Last Accessed: October 30, 2021)

8. Demin, G. (2020). *expss: Tables, Labels and Some Useful Functions from Spreadsheets and 'SPSS' Statistics.* R package version 0.10.7. https://gdemin.github.io/expss/

9. Gelman, A., Hullman, J., Wlezien, C., & Morris, G. E. (2020). Information, incentives, and goals in election forecasts. *Judgment and Decision Making, 15*(5), 863–880.

10. Gibbs, A., & Stringer, A. (2021, January 20). *Chapter 16 Short tutorial on pulling data for Assignment 1.* Probability, Statistics, and Data Analysis. https://awstringer1.github.io/sta238-book/section-short-tutorial-on-pulling-data-for-assignment-1.html#section-toronto-open-data-portal. (Last Accessed: October 30, 2021)

11. Hodgetts, P.A. & Alexander, R. (2021). *cesR: Access the CES Datasets a Little Easier.* R package version 0.1.0.

12. Jennings, W., Lewis-Beck, M., & Wlezien, C. (2020). Election forecasting: Too far out? *International Journal of Forecasting, 36*(3), 949–962. https://doi.org/10.1016/j.ijforecast.2019.12.002

13. Kennedy, L., Khanna, K., Simpson, D., & Gelman, A. (2020). *Using sex and gender in survey adjustment.*

14. Nissanov, Z. (2019). Israeli political attitudes and income in the 2006-2015 elections. *Israel Affairs, 25*(4), 740–753. https://doi.org/10.1080/13537121.2019.1626100

15. Osborne, D., & Sibley, C. G. (2015). Within the Limits of Civic Training: Education Moderates the Relationship Between Openness and Political Attitudes: Education Moderates the Relationship. *Political Psychology, 36*(3), 295–313. https://doi.org/10.1111/pops.12070

16. Pedersen, T.L. (2020). *patchwork: The Composer of Plots.* https://patchwork.data-imaginist.com, https://github.com/thomasp85/patchwork.

17. Peterson, J. C., Smith, K. B., & Hibbing, J. R. (2020). Do People Really Become More Conservative as They Age? *The Journal of Politics, 82*(2), 600–611. https://doi.org/10.1086/706889

18. R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/. (Last Accessed: October 30, 2021)

19. Rossi, R., Murari, A., Gaudio, P., & Gelfusa, M. (2020) *Upgrading Model Selection Criteria with Goodness of Fit Tests for Practical Applications.* entropy. Retrieved November 2, 2021, from https://www.mdpi.com/1099-4300/22/4/447/pdf.

20. Scala, D. J., & Johnson, K. M. (2017). Political Polarization along the Rural-Urban Continuum? The Geography of the Presidential Vote, 2000–2016. *The Annals of the American Academy of Political and Social Science, 672*(1), 162–184. https://doi.org/10.1177/0002716217712696

21. Stephenson, L.B., Harell, A., Rubenson, D., Loewen, P. J. (2020). 2019 Canadian Election Study - Phone Survey, https://doi.org/10.7910/DVN/8RHLG1, Harvard Dataverse, V1, UNF:6:eyR28qaoYlHj9qwPWZmmVQ== [fileUNF]

22. Stephenson, L.B., Harell, A., Rubenson, D., Loewen, P. J. (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

23. Student. (1908). The probable error of a mean. *Biometrika*, 1–25.

24. *The Canadian Parliamentary System.* (n.d.). The Canada Guide. https://thecanadaguide.com/government/parliament/. (Last Accessed: October 29, 2021)

25. Tahirali, J., & Hahn, P. (2021, September 27). *Six charts to help you understand the 2021 federal election.* CTVNews. https://www.ctvnews.ca/politics/federal-election-2021/six-charts-to-help-you-understand-the-2021-federal-election-1.5598419. (Last Accessed: November 3, 2021)

26. *What is CATI?* (n.d.). B2B International. https://www.b2binternational.com/research/methods/faq/what-is-cati/. (Last Accessed: October 31, 2021)

27. Wickham et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686. (Last Accessed: October 30, 2021)

28. Wickham, H. (2019). *stringr: Simple, consistent wrappers for common string operations.* R package version 1.4.0. https://CRAN.R-project.org/package=stringr. (Last Accessed: October 30, 2021)

29. Wickham, H., François, R., Henry, L. & Müller, K. (2021). *dplyr: A Grammar of Data Manipulation.* https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr. (Last Accessed: October 30, 2021)

30. Zhu, H. (2021). *kableExtra: construct complex table with 'kable' and pipe syntax.* R package version 1.3.4. https://CRAN.R-project.org/package=kableExtra. (Last Accessed: October 30, 2021)

Appendix

Figure 5: Distribution of gender for surveyed population

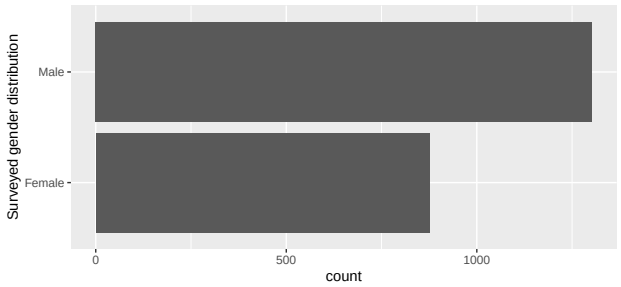

Figure 6: Distribution of gender for census population

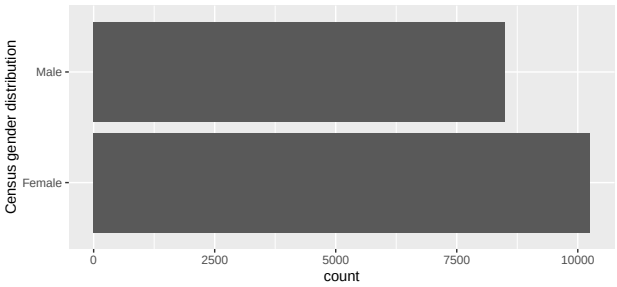

Figure 5 and 6 are bar plots for gender in the surveyed population and census population. From the two bar plots, we can see that in the surveyed population, the male proportion is higher than the female proportion by a visible amount. On the other hand, the number of individuals identified as female is greater than the number of people identified as male, but the difference in the gender proportion in census data appears to be smaller than that in the survey data.

Figure 7: Distribution of total income per family for surveyed population

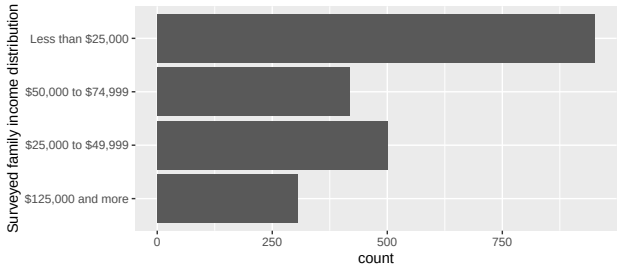

Figure 8: Distribution of total income per family for census population

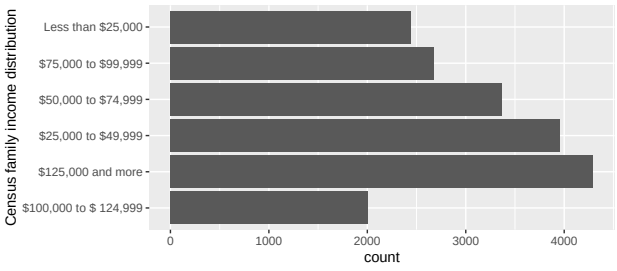

Figure 7 and 8 are bar plots for total household income in the surveyed population and census population. From the two figure we can see that the distribution of income per family in the surveyed population has two fewer bins than the census population. This is consistent with the numerical summary where we find that the proportions of individuals with income "$75,000 to $99,999" and "$100,000 to $124,999" are zero in the surveyed population. Besides the missing bins, the proportion of other categories, such as "$125,000 and more", is also very different in the two data sets. Therefore, we can see from the two figures that the distributions for total family income in the surveyed and census population are very different.