The Prediction of People's Voting to the Liberal Party in the Next Federal Election of Canada

STA304 - Assignment 2

Group 55: Xinle Cui, Guanhao Dong, Penny Hu

November 24, 2022

Introduction

Since Canada's founding in 1867, Canada has held 44 elections. Free and fair elections are a significant characteristic of a healthy democracy. Citizens 18 or older are responsible for voting for a candidate representing a party, and the number of votes is an essential factor in winning. Once in office, the winning candidate will represent the entire Canadians for the next couple of years. He/She will also lead the government to keep Canada's communities safe, grow the economy, provide social welfare, and so on. Therefore, voting is essential for all citizens to ensure their voices are heard in Canada.

Interestingly, the voting result often depends on different factors, including age, sex, income, province, etc. Glenn Norval and Michael (1968) state that older citizens are more likely to vote, and younger people show an increasing insensitivity to politics [1]. This is also demonstrated in survey data, that only 10 percent of people aged 30 or younger completed the survey.

Besides, some economists believe that since people care about their after-tax income, lowering the income tax will increase the employment rate and grows the economy. However, some believe an increase in income tax will simultaneously improve overall social welfare [2]. Since different Parties hold different political ideas, the income also becomes an important factor in voting.

Moreover, according to research by The State University of New Jersey (n.d.), the turnout gap between females and males grew slightly larger with each presidential election in the USA [3]. Thus, gender also plays a crucial role in the result of elections. Lastly, since the distribution of votes to different political parties varies significantly from province to province, we would also like to include this factor in our study.

Based on the above research, our team decided to predict the likelihood of people voting for the Liberal Party in the next Canadian federal election in 2025 using survey and census data. We would use age, sex, income, and province as predictors. Since Voting for the Liberal Party is a binary variable (1 means voting for Liberal Party, and 0 means not), we plan to use binary logistic regression to build the model. Due to the varies between survey and census data, we will also imply the post-stratification method to adjust the estimates and weighted average of estimates from all possible combinations of attributes. This will be discussed later in the Method and Model section.

Among all the predictors, we expect a significant difference in provinces voting Liberal Party. From the latest election result [4], Ontario should have the highest increase in log odds of voting for the Liberal Party. And the log odds of voting for the Liberal Party should differ in each province.

Data

The survey data was collected in 2019 during the election campaign, with telephone interviews completed with 4,021 Canadian citizens [5]. Interviewers completed all Computer Assisted Telephone Interviewing

(CATI) that are located throughout Canada. A modified random digit dialing (RDD) procedure was used in the data collection process, in which the respondent was determined by using the birthday selection. The team used such a method to select phone numbers for the survey. Call attempts were made day and night on weekdays and weekends to maximize the chances of getting a completed interview from each telephone number. Once six attempts had been made for a sample record without an eligible respondent being reached, the team moved to the next randomly selected sample. However, some calls exceeded the six attempts at the end of data collection.

Lastly, the survey was designed with 77 questions. The data contains 273 variables with both numerical (such as age, income, etc.) and categorical (such as sex, province, etc.) with 4021 observations across Canada.

The census data was collected via CATI in 2017 [6]. Respondents were interviewed in their choice of official language (either English or French). All interviews took place using centralized telephone facilities in 5 of Statistics Canada's regional offices, day and night on weekdays and weekends. Respondents who first refused to participate were re-attempts up to two more times to explain the importance of the survey to encourage them to participate. Eventually, there are 81 variables with 20602 observations contained in the data set.

Generally, the data cleaning [7] process focuses on eliminating the non-voters. The survey asked "which party you are likely or already voted for" in question 11 [8], and option 8 implies that the respondent would not vote. Besides, according to Statistics Canada, 24% of eligible Canadians reported as non-voters in the 2021 federal election due to everyday life situations (too busy, disability, being out of town, etc.) [9]. These non-voters represent a huge proportion of the population. Therefore, we decided to remove those observations to increase the preciseness of the model. After filtering those non-voters, there are 85 observations removed from the survey data.

In addition, question 2 asked about the year of born, thus we used the base year (2019) minus the birth year to find out the respondent's age. Question 3 asked about the respondents' gender, which the options include "male", "female", and "other". After checking the number of respondents in the gender group and comparing it with the census data, only one respondent determined themselves as "other". It's better to remove this observation and change the variable name to "sex".

Moreover, we divided the respondent's income into six groups: "Lower-class", "Lower-middle-class", "Middle-class", "Upper-middle-class", "Upper-class", and "Top", with the cutoff line at $0, $25000, $49999, $74999, $99999, $124999 and infinite respectively, in both census and survey data. The idea of the income segment was from both Statistics Canada [10] and Global News Canada [11], which suggested the median income in Canada was 62,400 between 2019 and 2020.

Lastly, since the census data was collected in 2017, in order to match the "age" variable in the survey data, we rounded the age into integers and added two additional years (based on the year 2019). However, there are some limitations during the process. For instance, if an observation is age 30.5, it's hard to determine whether the respondent is 30 or 31 years old. Also, we removed the observation that the age is not equal to or larger than 18, which is not eligible for voting in Canada.

After the cleaning process, we have a total of 2912 observations and 5 variables in the survey data:

- age: the respondent's age (18 years old and older)
- sex: the respondent's sex
- income: the respondent's yearly income in Canadian dollars (CAD $)
- province: the province that the respondent lives in
- vote_Liberal: whether the respondent will vote for Liberal Party or not

And we have a total of 20548 observations and 4 variables in the census data: * age: the respondent's age (18 years old and older) * sex: the respondent's sex * income: the respondent's yearly income in Canadian dollars (CAD $) * province: the province that the respondent lives in

Survey

sex: A categorical variable describes the sex of the respondent. There are approximately 58 percent of Males and 42 percent of Females in the survey (Table 1).

Table 1: Summary measure for sex in survey

sex	Number of observation	Proportion
Female	1228	0.4217033
Male	1684	0.5782967

income: A categorical variable describes the income group of the respondent. About 31 percent of the respondents are in the Top income class, which their income equal to or large than $125,000. And approximately 20 percent in Middle-class (Table 2).

Table 2: Summary measure for income group in survey

income	Number of observation	Proportion
Lower class	264	0.0906593
Lower-middle class	424	0.1456044
Middle class	537	0.1844093
Upper-middle class	403	0.1383929
Upper class	392	0.1346154
Top	892	0.3063187

age: A numerical variable describes the age of the respondent. The range of their ages is from 18 to 95 years old. This means there is no outlier. Another thing important to notice is that the mean is 50.35268 while the median is 50 (Table 3). In this case, the mean is slightly larger than the median. Thus, the distribution is slightly right-skewed. The standard deviation is 16.0521 which is quite reasonable. This indicates the confidence interval that is produced later would not be too wide or too narrow. Combining the information, it's likely that the distribution of survey data is Normal, distributed around the mean of 50.35268.

Table 3: Summary measure for age in survey

median	mean	sd	min	max
50	50.35268	16.0521	18	95

vote_Liberal: A numerical variable that describes the respondent's decision to vote for Liberal Party or not. In the survey data, the percentage of respondents who voted for the Liberal Party is 24. It is a binary variable so we assigned 1 to represent the respondent who votes for the Liberal Party, and 0 represents the respondent who does not vote for the Liberal Party (Table 4).

Table 4: Summary measure for probability of voting for Liberal Party in survey

vote_Liberal	Voter in survey	probability
0	2210	0.7589286
1	702	0.2410714

province: The most important variable is the province (Table 5). James and Thomas emphasize that in the geography of the presidential election, states matter more than the region in their article [12]. Their model indicates that differences among states are crucial. And recognize that the state is an important predictor of the elections, especially the distribution of votes in the past two elections. That also makes sense in Canada, the Bloc Québécois Party won 32 seats in the 2021 election in Canada, and all the seats came from Quebec [9].

Table 5: Summary measure for province in survey

province	Number of observation	Proportion
Alberta	193	0.0662775
British Columbia	569	0.1953984
Manitoba	190	0.0652473
New Brunswick	147	0.0504808
Newfoundland and Labrador	141	0.0484203
Nova Scotia	154	0.0528846
Ontario	578	0.1984890
Prince Edward Island	148	0.0508242
Quebec	589	0.2022665
Saskatchewan	203	0.0697115

In the survey data, the respondent is more likely to come from British Columbia, Ontario, or Quebec province. They share approximately 60 percent of voters in the survey data. However, this also shows a limitation. According to Statistics Canada, the population in Ontario is as large as twice that of Quebec [7]. That to some degree affects the preciseness of the outcome.

Census

sex: A categorical variable describes the sex of the respondent. There are approximately 46 percent of Males and 54 percent of Females in the census (Table 6).

Table 6: Summary measure for sex in census

sex	Number of observation	Proportion
Female	11177	0.5439459
Male	9371	0.4560541

income: A categorical variable describes the income group of the respondent in the census. There 33 percents of the observations are in the lower income class, which there income of less than $25,000. Noticeably, the difference between survey and census is huge. In the survey, only 9 percent of the observation is in the lower-income class (Table 7).

Table 7: Summary measure for income group in census

income	Number of observation	Proportion
Lower-middle class	6173	0.3004185
Lower class	6718	0.3269418
Middle class	3896	0.1896048
Top	885	0.0430699
Upper-middle class	2030	0.0987931
Upper class	846	0.0411719

age: A numerical variable describes the age of the respondent. The range is from 18 to 82 years old in the census. This means there is no outlier. The median is slightly large than the mean. Thus, the distribution is slightly left-skewed. The standard deviation is 17.67284 which is also reasonable. Combining the information, it's likely that the distribution of census data is Normal, distributed around the mean of 54.28017 (Table 8).

Table 8: Summary measure for age in census

median	mean	sd	min	max
56	54.28017	17.67284	18	82

province: In the census data, the respondents are also from the province same as the survey data. However, the distribution of province are more reasonable than the survey data. There are 5600 observation came from Ontario shares 27 percents of the population, and 3813 observations came from Quebec shares 19 percents of the population (Table 9).

Table 9: Summary measure for province in census

province	Number of observation	Proportion
Alberta	1726	0.0839984
British Columbia	2515	0.1223963
Manitoba	1190	0.0579132
New Brunswick	1336	0.0650185
Newfoundland and Labrador	1093	0.0531925
Nova Scotia	1420	0.0691065
Ontario	5600	0.2725326
Prince Edward Island	708	0.0344559
Quebec	3813	0.1855655
Saskatchewan	1147	0.0558205

Figure 1: Histogram of respondent shares in different prov
in survey

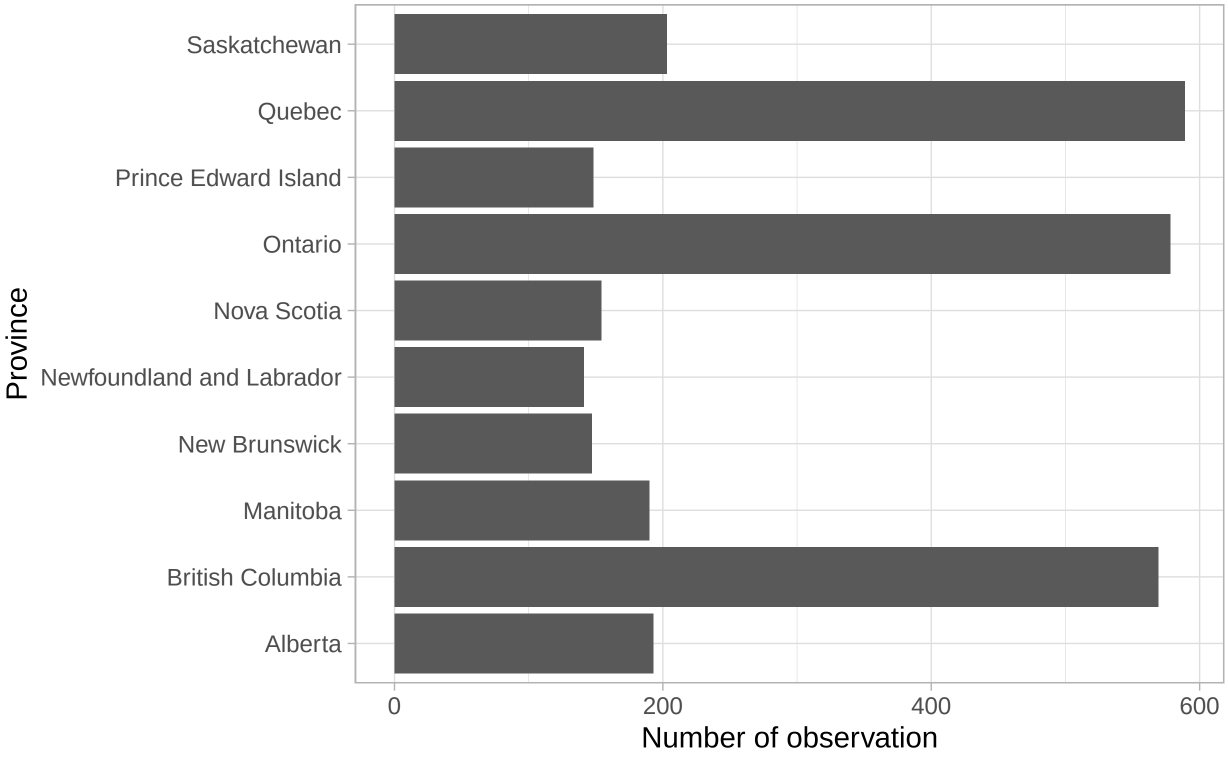
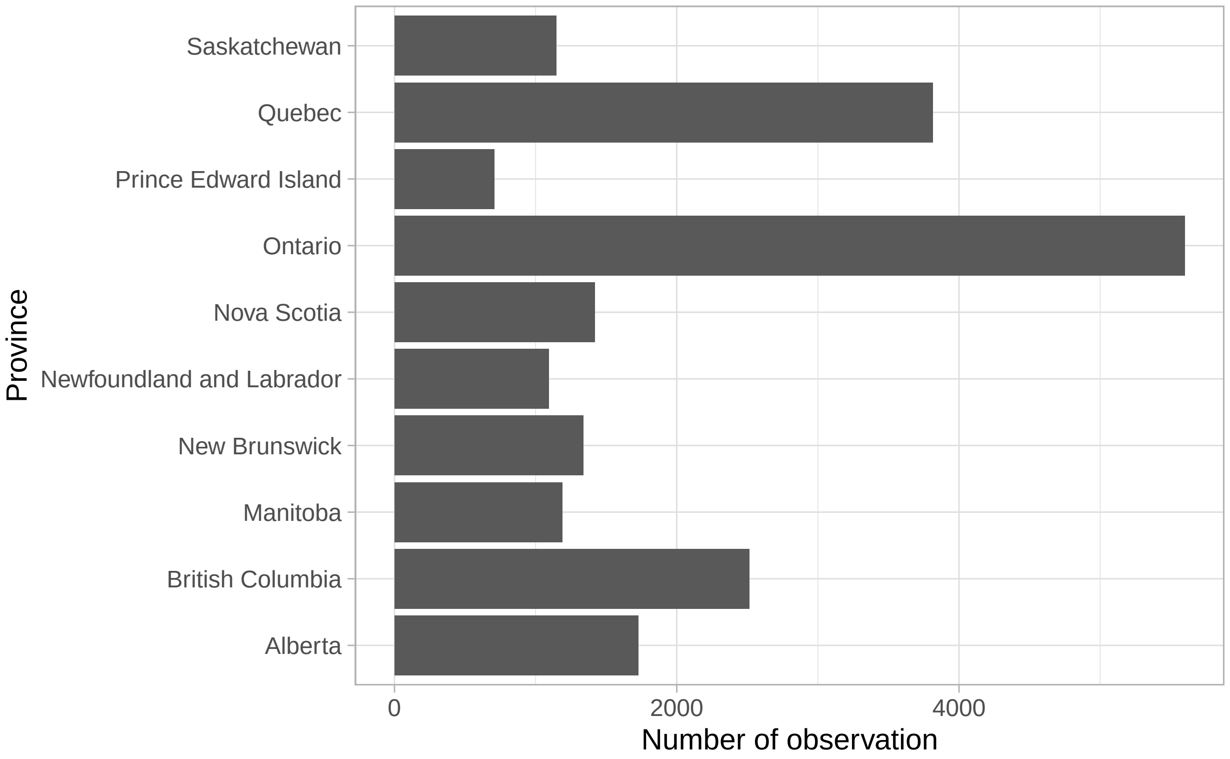


Figure 2: Histogram of respondent shares in different prov in census



To investigate the distribution of the population in provinces, we construct two histograms [10] (Figure 1 & Figure 2). In survey data, the proportion of the respondent in Quebec, Ontario and British Columbia are close. On the other hand, in census data, the proportion of the respondent differs from the survey. Quebec shares twice as than British Columbia. According to Statistics Canada, in 2021 Q3, Ontario had a population of 14,809,257, Quebec had 8,602,335, and British Columbia had 5,202,378 [7]. Ontario's population is almost three times that of British Columbia and 1.7 times that of Quebec. The results are reasonable in census data but differ in survey data. That we are going to use the post-stratification method when a simple random sample does not reflect the distribution of some known variable in the population.

Methods

We will use a logistic regression model to predict people's probability of voting for the Liberal Party according to the survey, and then apply the model to post-stratification to estimate the people's probability of voting for the Liberal Party for the next Canadian Federal election based on the population in Canada.

Logistic Regression

We use logistic regression because we want to know whether or not people will vote for the Liberal Party, which only has two outcomes, yes or no, indicated by 1 or 0, respectively. This is a binary response variable so we will use logistic regression constructed through all the variables in the cleaned survey data to predict the probability of people voting for the Liberal Party.

Prior to that, we need to convert categorical variables, sex, income and province to numerical variables for building the model. To do so, We can turn all the possible options in a certain categorical variable into new variables so we add the variable "male (x_{male})" for sex, add variables "lower-middle class ($x_{lowermiddle}$),

middle class (x_{middle}), upper-middle class ($x_{uppermiddle}$), upper class (x_{upper}), top class (x_{top})" for income and add variables "British Columbia (x_{BC}), Manitoba (x_{MB}), New Brunswick (x_{NB}), Newfoundland and Labrador (x_{NL}), Nova Scotia (x_{NS}), Ontario (x_{ON}), Prince Edward Island (x_{PE}), Quebec (x_{QC}), Saskatchewan (x_{SK})" for province. We assign the number 1 to the added variables if the respondent matches the conditions, otherwise, 0 will be assigned to the added variables. For instance, if a 24-year-old male from Ontario whose income is \$26000, then the variables used to describe her will be as follow: $x_{age} = 24$, $x_{male} = 1$, $x_{lowermiddle} = 1$ and $x_{ON} = 1$. For those who are female, the variable used to describe sex (x_{male}) will be equal to 0 because the base variable of sex is female. In the same way, for those whose income is lower class, the variables used to describe income ($x_{lowermiddle}$, x_{middle}, $x_{uppermiddle}$, x_{upper}, x_{top}) will all be equal to 0 and for those who come from Alberta, the variables used to describe province (x_{BC}, x_{MB}, x_{NB}, x_{NL}, x_{NS}, x_{ON}, x_{PE}, x_{QC}, x_{SK}) will all be equal to 0 because the base variables of income and province are lower class and Alberta, respectively. Then, we remove the original categorical variables.

Since the respondents reported their decisions that whether or not they would like to vote for the Liberal Party and they also reported their ages, sex, income and residential provinces, we are able to set up multiple predictors to obtain the probability of people voting for the Liberal Party through the logistic regression model, according to the data of respondents in the survey. However, we will first use predictors to obtain the logarithm of the odds, $\frac{p}{1-p}$, which is our response variable.

We will have the following equation to build the logistic regression model if the survey data meets the assumption of logistic regression model (independence of errors, linearity in the logit for continuous variables, absence of multicollinearity and lack of strongly influential outliers) [13]:

$$\begin{aligned} log(\frac{p}{1-p}) = & \beta_0 + \beta_1 x_{age} + \beta_2 x_{male} + \beta_3 x_{lowermiddle} + \beta_4 x_{middle} + \beta_5 x_{uppermiddle} \\ & + \beta_6 x_{upper} + \beta_7 x_{top} + \beta_8 x_{BC} + \beta_9 x_{MB} + \beta_{10} x_{NB} + \beta_{11} x_{NL} + \beta_{12} x_{NS} \\ & + \beta_{13} x_{ON} + \beta_{14} x_{PE} + \beta_{15} x_{QC} + \beta_{16} x_{SK} \end{aligned}$$

- p represents the probability of the vote for Liberal Party among the population.
- β_0 represents the intercept of the model, and is the average log of odds of voting for the Liberal Party when all variables take the value of 0.
- β_1 represents the slope of age in the model. After controlling for sex, income and province, a one unit increase in age, we expect a β_1 increase in log odds of voting for the Liberal Party.
- β_2 represents the average difference in log odds of voting for the Liberal Party between Male and Female after controlling the age, income, and province. (The base variable is Female)
- β_3 represents the average difference in log odds of voting for the Liberal Party between the income class of the lower class and the lower-middle class after controlling the age, sex and province. (The base variable is Lower-class)
- β_4 represents the average difference in log odds of voting for the Liberal Party between the income class of the lower class and the middle class after controlling the age, sex and province. (The base variable is Lower-class)
- β_5 represents the average difference in log odds of voting for the Liberal Party between the income class of the lower class and the upper-middle class after controlling the age, sex and province. (The base variable is Lower-class)
- β_6 represents the average difference in log odds of voting for the Liberal Party between the income class of the lower class and the upper class after controlling the age, sex and province. (The base variable is Lower-class)

- β_7 represents the average difference in log odds of voting for the Liberal Party between the income class of the lower class and the top class after controlling the age, sex and province. (The base variable is Lower-class)
- β_8 represents the average difference in log odds of voting for the Liberal Party between the province of Alberta and British Columbia after controlling the age, sex and income. (The base variable is Alberta)
- β_9 represents the average difference in log odds of voting for the Liberal Party between the province of Alberta and Manitoba after controlling the age, sex and income. (The base variable is Alberta)
- β_{10} represents the average difference in log odds of voting for the Liberal Party between the province of Alberta and New Brunswick after controlling the age, sex and income. (The base variable is Alberta)
- β_{11} represents the average difference in log odds of voting for the Liberal Party between the province of Alberta and Newfoundland and Labrador after controlling the age, sex and income. (The base variable is Alberta)
- β_{12} represents the average difference in log odds of voting for the Liberal Party between the province of Alberta and Nova Scotia after controlling the age, sex and income. (The base variable is Alberta)
- β_{13} represents the average difference in log odds of voting for the Liberal Party between the province of Alberta and Ontario after controlling the age, sex and income. (The base variable is Alberta)
- β_{14} represents the average difference in log odds of voting for the Liberal Party between the province of Alberta and Prince Edward Island after controlling the age, sex and income. (The base variable is Alberta)
- β_{15} represents the average difference in log odds of voting for the Liberal Party between the province of Alberta and Quebec after controlling the age, sex and income. (The base variable is Alberta)
- β_{16} represents the average difference in log odds of voting for the Liberal Party between the province of Alberta and Saskatchewan after controlling the age, sex and income. (The base variable is Alberta)

By looking at the age (x_{age}), sex (x_{male}), income group ($x_{lowermiddle}$,x_{middle}, $x_{uppermiddle}$, x_{upper}, x_{top}) and the residential province (x_{BC}, x_{MB}, x_{NB}, x_{NL}, x_{NS}, x_{ON}, x_{PE}, x_{QC}, x_{SK}) of a person, we can predict the probability of people voting for the Liberal Party of the population through the logistic regression model based on the obtained log odds.

Post-Stratification

In the survey, since the proportion of the respondents in a variable may differ from the proportion in the population, for example, Ontario's population is almost three times that of British Columbia and 1.7 times that of Quebec but the proportion of respondents live in Ontario, Quebec and British Columbia is almost the same in the survey, we need to use post-stratification to better estimate our prediction regarding the probability of people's vote for Liberal Party according to the census data.

We will create multiple cells according to the census data for the post-stratification. Each cell consists of the people in the census who have the same age, income class, sex and residential provinces. After that, we can estimate the probability of people voting for the Liberal Party in each cell by the obtained logistic regression model.

We will also obtain the number of people in each cell so we can weight the estimate of each cell for the overall estimate in the future.

Then, we have the following equation:

$$\hat{p}^{PS} = \frac{\sum N_j \widehat{p}_j}{\sum N_j}$$

- $\hat{p}^{PS}$ represents the overall estimate of the probability of people voting for the Liberal Party.
- N_j represents the number of people in the j^{th} cell.
- $\widehat{p}_j$ represents the estimate of the probability of people voting for the Liberal Party in the j^{th} cell.

We can better obtain the overall estimate of the probability of people voting for the Liberal Party according to the estimate and the number of people in each cell.

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

Results

We obtained the logistic regression model and the predictors are shown as below in the "Predictors" column (Table 10):

Table 10: Predictors of the logistic regression model

Variables	Predictors	Standard Error	Statistic	P-Value
(Intercept)	-2.9548910	0.3200228	-9.2333756	0.0000000
age	0.0103749	0.0028474	3.6437077	0.0002687
sexMale	-0.0582663	0.0899247	-0.6479458	0.5170200
incomeLower-middle class	0.3281804	0.2002157	1.6391343	0.1011853
incomeMiddle class	0.4202435	0.1917745	2.1913415	0.0284271
incomeUpper-middle class	0.6040066	0.1987926	3.0383764	0.0023786
incomeUpper class	0.3330932	0.2051824	1.6234000	0.1045039
incomeTop	0.5460884	0.1823441	2.9948227	0.0027460
provinceBritish Columbia	0.7008163	0.2494963	2.8089243	0.0049707
provinceManitoba	0.8144198	0.2856925	2.8506868	0.0043625
provinceNew Brunswick	0.8172315	0.3008250	2.7166341	0.0065949
provinceNewfoundland and Labrador	1.1363893	0.2944016	3.8599966	0.0001134
provinceNova Scotia	1.1360694	0.2899819	3.9177257	0.0000894
provinceOntario	1.3199456	0.2438845	5.4121744	0.0000001
provincePrince Edward Island	1.1690848	0.2910962	4.0161463	0.0000592
provinceQuebec	0.8799128	0.2477830	3.5511428	0.0003836
provinceSaskatchewan	-0.0730417	0.3207600	-0.2277146	0.8198681

that is, we have the logistic regression model:

$$\begin{aligned} log(\frac{p}{1-p}) = {} & -2.9548910 + 0.0103749x_{age} - 0.0582663x_{male} + 0.3281804x_{lowermiddle} + 0.4202435x_{middle} \\ & + 0.6040066x_{uppermiddle} + 0.3330932x_{upper} + 0.5460884x_{top} + 0.7008163x_{BC} + 0.8144198x_{MB} + 0.8172315x_{NB} \\ & + 1.1363893x_{NL} + 1.1360694x_{NS} + 1.3199456x_{ON} + 1.1690848x_{PE} + 0.8799128x_{QC} - 0.0730417x_{SK} \end{aligned}$$

That means:

- The intercept of the model, and the average log of odds of voting for the Liberal Party is -2.954891 when all variables take the value of 0.
- The slope of age in the model is 0.0103749 after controlling for sex, income and province, a one-unit increase in age, we expect a 0.0103749 increase in log odds of voting for the Liberal Party.

- The average difference in log odds of voting for the Liberal Party between Male and Female is -0.0582663, after controlling the age, income, and province.
- The average difference in log odds of voting for the Liberal Party between the income class of the lower class and the lower-middle class is 0.3281804, after controlling the age, sex and province.
- The average difference in log odds of voting for the Liberal Party between the income class of the lower class and the middle class is 0.4202435, after controlling the age, sex and province.
- The average difference in log odds of voting for the Liberal Party between the income class of the lower class and the upper-middle class is 0.6040066, after controlling the age, sex and province.
- The average difference in log odds of voting for the Liberal Party between the income class of the lower class and the upper class is 0.3330932, after controlling the age, sex and province.
- The average difference in log odds of voting for the Liberal Party between the income class of the lower class and the top class is 0.5460884, after controlling the age, sex and province.
- The average difference in log odds of voting for the Liberal Party between the province of Alberta and British Columbia is 0.7008163, after controlling the age, sex and income.
- The average difference in log odds of voting for the Liberal Party between the province of Alberta and Manitoba is 0.8144198, after controlling the age, sex and income.
- The average difference in log odds of voting for the Liberal Party between the province of Alberta and New Brunswick is 0.8172315, after controlling the age, sex and income.
- The average difference in log odds of voting for the Liberal Party between the province of Alberta and Newfoundland and Labrador is 1.1363893, after controlling the age, sex and income.
- The average difference in log odds of voting for the Liberal Party between the province of Alberta and Nova Scotia is 1.1360694, after controlling the age, sex and income.
- The average difference in log odds of voting for the Liberal Party between the province of Alberta and Ontario is 1.3199456, after controlling the age, sex and income.
- The average difference in log odds of voting for the Liberal Party between the province of Alberta and Prince Edward Island is 1.1690848, after controlling the age, sex and income.
- The average difference in log odds of voting for the Liberal Party between the province of Alberta and Quebec is 0.8799128, after controlling the age, sex and income.
- The average difference in log odds of voting for the Liberal Party between the province of Alberta and Saskatchewan is -0.0730417, after controlling the age, sex and income.

We will apply this model to our post-stratification. Firstly, we can divide the people in the census data into 4,994 cells, which means 20,548 people in the cleaned census data can be stratified into 4,994 types based on their sex, income and residential provinces. Then, we use the logistic regression model to estimate the probability of voting for the Liberal Party in each cell. Here are the first six of the cells we stratified (Table 11):

Table 11: First six of the cells for the post-stratification

age	sex	income	province	n	estimate
18	Female	Lower-middle class	Quebec	1	0.1736397
18	Female	Lower class	Alberta	4	0.0590702
18	Female	Lower class	British Columbia	7	0.1123134
18	Female	Lower class	Manitoba	6	0.1241481

age	sex	income	province	n	estimate
18	Female	Lower class	New Brunswick	3	0.1244541
18	Female	Lower class	Newfoundland and Labrador	4	0.1635904

Finally, we are able to calculate the overall estimate by the formula $\hat{p}^{PS} = \frac{\sum N_j \widehat{p_j}}{\sum N_j}$, which N_j represents n for each cell in the table, the number of people in each cell, and $\widehat{p_j}$ represents the estimate in the table, the probability of voting to the Liberal Party in each cell.

Therefore, we have the probability of 0.23324177051717 that people will vote for the Liberal Party in the next federal election in Canada (Table 12), which is different from the latest vote percentage of the Liberal Party, 33% [14].

Table 12: Estimated Probability of people voting to the Liberal Party in the next Federal Election of Canada

Probability
0.2332418

However, since some of the predictors have large p-values when constructing the logistic regression model such as the predictors for sex, we are hard to against the hypothesis that the predictors of these variables are equal to zero. Although most predictors are significant, especially the predictors for most provinces, this still will influence our estimate, which causes bias in predicting the probability [15].

Conclusions

This study first gives readers a general idea of how the voting process in Canadian election work by introducing its background information, the voting data collecting process, and some critical factors that may affect the voting result. More importantly, the study focuses on one of the parameters, the province, and predicts the probability of people voting for the Liberal Party in the following Canadian federal election in 2025. We hypothesize that, among all the possible predictors, the province is the most significant factor affecting the election result. We then set up a binary logistic regression model and used the post-stratification method to achieve our goal step by step.

We found out that there is a significant difference between the probability of the voting rate for the Liberal Party in different provinces. For example, the average difference in log odds of voting for the Liberal Party between the province of Alberta and Nova Scotia, Ontario, Prince Edward Island, Newfoundland and Labrador is greater than 1 and the p-values of predictors of these factors are very small, which verifies our expectation that there will be a significant difference in provinces voting for Liberal Party.

Eventually, we came up with the result that there is a probability of 0.2332418 that people will vote for the Liberal Party in the next federal election in Canada. Compared to the result from the past years of elections, such a value is a bit smaller since the percentage of Canada voted for liberal was around 30% - 33% in the last few elections. However, the result is still reasonable since the people's political stances are subject to change over time due to the Party's actual contribution to the country and whether their needs are being met.

During the prediction process, some drawbacks and limitations must be considered. Firstly, since we removed the "other" option in the question, which asks about people's gender, there are some minority groups of people that are not included in the study (e.g. the LGBTQ). Secondly, we only have a limited number of control variables. Instead, some unobserved conditions may cause bias in the result. For instance, we cannot ensure whether a Party is canvassing in a community, which may help increase the voting rate of that Party within that area. Besides, there are some missing data on some provinces in the survey, which may also lead to bias. Finally, some predictors in the result are not significant with large p-values, which may influence our estimate. All the above limitations may affect the final predicted probability.

The predicted result of the probability of people voting for the Liberal Party in the following Canadian federal election is smaller than it used to be during the past few elections, regardless of some possible biases and limitations. Thus, if the Liberal Party wants to continue to be elected in the future, it must pay more to gain people's favor and increase the country's well-being. This may be achieved by introducing more social benefits, raising the employment rate, increasing the country's welfare, maintaining national security and stability, and so on. Overall, to become a qualified ruling party in Canada, the Liberal Party has to continue to pay more attention to people and the country's actual demands and try its best to satisfy society.

Bibliography

1. Glenn, Norval D., and Michael Grimes. *Aging, Voting, and Political Interest.* American Sociological Review, vol. 33, no. 4, 1968, pp. 563–75. JSTOR. Retrieved November 20, 2022, from https://doi.org/10.2307/2092441.

2. Matsubayashi, T., & Sakaiya, S. (2021). *Income inequality and income bias in voter turnout.* European Journal of Political Economy, 66, 101966. Retrieved November 20, 2022, from https://doi.org/10.1016/j.ejpoleco.2020.101966.

3. *Gender differences in voter turnout.* Center for American Women and Politics. Retrieved November 20, 2022, from https://cawp.rutgers.edu/facts/voters/gender-differences-voter-turnout.

4. *Party Standings in the House of Commons.* Party Standings in the House of Commons - Members of Parliament - House of Commons of Canada. Retrieved November 20, 2022, from https://www.ourcommons.ca/members/en/party-standings.

5. Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020. *2019 Canadian Election Study - Phone Survey.* Retrieved November 20, 2022, from https://doi.org/10.7910/DVN/8RHLG1, V1.

6. Statistics Canada. *General Social Survey Cycle 31: Family, 2017.* Statistics Canada Open License, 2020.Retrieved November 20, 2022.

7. Government of Canada, Statistics Canada (2022, September 28). *Population Estimates, Quarterly.* Statistics Canada. Retrieved November 20, 2022, from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000901.

8. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media. Retrieved November 20, 2022

9. Statistics Canada. *The Daily — Reasons for Not Voting in the Federal Election, September ...* Retrieved November 20, 2022, from https://www150.statcan.gc.ca/n1/daily-quotidien/220216/dq220216d-eng.htm.

10. Statistics Canada. *Income Statistics by Selected Family Type, Canada, 2019 and 2020.* Retrieved November 20, 2022, from https://www150.statcan.gc.ca/n1/daily-quotidien/220323/t001a-eng.htm.

11. Alini, E. (2021, December 3) *Are You Earning a Middle-Class Income? Here's What It Takes in Canada, Based on Where You Live - National.* Global News. Retrieved November 20, 2022, from https://globalnews.ca/news/3828447/canada-middle-class-income-inequality/.

12. Campbell, James E., and Thomas E. Mann. (2016, July 28) *Forecasting the Presidential Election: What Can We Learn from the Models?.* Brookings. Retrieved November 20, 2022, from https://www.brookings.edu/articles/forecasting-the-presidential-election-what-can-we-learn-from-the-models/.

13. Stoltzfus JC. *Logistic regression: a brief primer.* Acad Emerg Med. 2011 Oct;18(10):1099-104. doi: 10.1111/j.1553-2712.2011.01185.x. PMID: 21996075..

14. Renfrew, M. (2022, November 15). *Liberal 32%, conservative 34%, NDP 19%: Léger.* Cult MTL. https://cultmtl.com/2022/11/liberal-32-conservative-34-ndp-19-leger-poll-federal-election-voting-intentions-canada/.

15. Frost, J. (2019, February 26) *Can high P-values be meaningful?.* Statistics By Jim. Retrieved November 20, 2022, from https://statisticsbyjim.com/hypothesis-testing/high-p-values/.

16. Wickham, H., et al. (2019). *Welcome to the tidyverse.* Journal of Open Source Software, 4(43), 1686. Retrieved November 20, 2022. https://doi.org/10.21105/joss.01686.

17. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html.

18. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/.