Predicting the 2023 Canadian Federal Election Using Bayesian Logistic Regression with Post-stratification

STA304 - Assignment 2

Group 7: Ju-Eun Kim, Fernando Sanchez-Avila, Petar Solaja, Kate Tong

May 28, 2021

Introduction

Background

There are a variety of methods one may use to predict election outcomes, ranging from polls, surveys, to different types of quantitative models. Election surveys are usually conducted close to the time of an election to obtain the highest degree of prediction accuracy. However, we may still be interested in predicting an upcoming election months before election surveys are even conducted. In that case, we may use quantitative models to predict an upcoming election based on data obtained from past election surveys and from census data. Such is the interest of this report. In this report, we aim to predict the overall popular vote of the 2023 Canadian federal election using Bayesian logistic regression models with post-stratification, based on survey data from the 2019 Canadian Election Study (CES) and census data from the General Social Survey (GSS). We built our regression model using the response variable 'vote choice,' and six predictor variables, including the year of birth, gender, province, education level, income, and religion of the survey respondents in CES 2019. Our model predicts the probability of a given person voting for the Liberal Party or Conservative Party depending on their age, gender, province, education level, income and religion. The subgroups in our survey sample are then post-stratified using census data from the GSS, which serves as the 'truth benchmark' for demographic characteristics of the Canadian population to correct for differences between the sample and the target population.

The key concept used in this report is the technique of post-stratification. Post-stratification is a post-survey reweighing technique used to correct for bias. Often, despite our best efforts, there will still be bias in our sample. For example, since the CES 2019 data, we obtained for our analysis is data from an online survey, it is possible that the younger population is oversampled and the older population undersampled, as the older generation does not use the internet as often as the younger generation does. In this case, we may use post-stratification to adjust the weights of undersampled and oversampled subgroups, so our sample is more representative of the true distributions in the target population. An example of post-stratification being used to correct for extreme differences between distributions in the sample and the target population is the paper by Wang et al. (2015) entitled "Forecasting elections with non-representative polls". In their paper, Wang et al. created an election forecast using daily voter intention polls conducted on the Xbox gaming platform for the 2012 US presidential election and adjusted responses with multilevel regression and post-stratification. With this method, they were able to obtain estimates that are in line with forecasts from leading poll analysts, which were based on aggregating hundreds of traditional voter polls. Evidently, the technique of post-stratification could be a low-cost-high-benefit way of predicting an election and potentially a broad range of other issues.

Relevance

This type of analysis serves many important purposes. Firstly, it provides information to politicians and political activists about a likely election outcome based on current demographics, allowing them to evaluate policy initiatives and make adjustments they feel necessary to their campaigns in order to target key demographics. Secondly, much is learned about what causes election outcomes in democratic polities as a whole as we learn about how different parameters, i.e. year of birth, gender, province, education level, income, and religion, influence the probability of a person voting for one political party over another. Finally, models predicting election outcomes satisfy our curiosity as citizens. Most of us, as citizens, want to know who is going to win in the next election and how voters belonging to different demographics tend to vote. These are pieces of information that help facilitate political discourse and are intrinsically interesting in any healthy democracy.

Terminology

A political terminology used in the Results section is the notion of electoral districts in Canada, also known as ridings. Every province of Canada is divided into geographic areas called electoral districts to provide their population with representation in Canada's legislative body. Each electoral district is represented by a seat in the House of Commons. In elections, voters registered in a particular voting district can only vote for candidates running for office in that district, and whichever party wins in a specific electoral district will win a seat in the House of Commons. The electoral quotient is the average population per electoral district. Each province's electoral quotient is calculated by dividing the province's total population by the number of seats allocated. With 121 seats, Ontario has an electoral quotient of 111,144, while Prince Edward Island has the lowest electoral quotient of 35,726. ("House of Commons of Canada", 2021)

Hypothesis

We hypothesize that the Liberal Party has a greater probability of winning the 2023 Canadian federal election due to a vast majority of opinion polls conducted by different polling firms for the 2023 Canadian federal election indicating that the Liberal Party is leading ("Opinion polling for the 44th Canadian federal election," 2021). In fact, all of the opinion polls conducted after October 2020 indicate that the Liberal Party is in the lead. Therefore, it is reasonable to hypothesize that the Liberal Party will win the 2023 Canadian federal election.

Data

Data Cleaning Process

The survey data was collected from Harvard Dataverse. The raw data was collected and will be cleaned throughout this section. It contains information about the 2019 Canadian Election Study done via the online survey. The survey included information about the attitudes of Canadians during and after the 2019 election (Stephenson et al., 2020). The data will be used to make the appropriate model to predict the election of the 2023 Canadian Election.

The census data was collected from Computing in the Humanities and Social Sciences (CHASS), the U of T library. CHASS allowed access to the General Social Survey (GSS) data obtained in 2017. The row data was downloaded into the CSV. File and cleaned using the `gss_cleaning.R` code to make the data more practical for analysis. Further cleaning will be done in this section, and the data will be further analyzed both numerically and visually.

First, both the survey and the census data will be cleaned by removing the irrelevant information. Both datasets contain lots of information that are unnecessary for the analysis. Therefore, the unnecessary columns will be deleted for efficient analysis. The selected columns will be only used for the entire study. The process is done using the `select` function in the `tidyverse` package.

Both dataset, they have the common or similar columns. The column related to age or the year of birth, gender/sex, province, income, and religion are common to both variables except for the `cps19_votechoice`, which will be the output of the predicting variable. Having similar variables, it will make the result easier to interpret.

Next, there were answers that many people did not remember or backed out from answering each question about in the survey. They will be removed while not answering the questions may be biased due to unclear explanation of each variable. However, it is possible that eliminating some answers may also cause bias. When one of the answers is missing from observation, the entire row will be deleted to minimize the non-response bias. That way, the cases observed for all questions will be only analyzed. This step applies to both datasets.

After the cleaning process, the remaining columns and rows from both datasets are essential for the complete analysis.

For the model building, the survey data will be introduced for the "Key Variable Section."

Key Variables of Survey Data

Dependent Variable Firstly, the `cps19_votechoice` is the critical respondent, which is also called the dependent variable for the objective of the analysis. The `cps19_votechoice` is the collection of responses from the question, "Which party do you think you will vote for?" which is the categorical variable (Stephenson et al., 2020). It is directly related to the main objective of the analysis. Using the `survey_data`, the final goal is to build the predicting model for the 2023 Canadian Election. The model will include various independent variables; the dependent variable will be shown as the output. Therefore, the dependent variable will be examined how it changes or gets influenced due to all other factors.

Independent Variables The independent, predictor variables are: `cps19_yob`, `cps19_gender`, `cps19_province`, `cps19_education`, `cps19_votechoice`, `cps19_income_number`, `cps19_religion`.

The variable `cps19_yob` is crucial because it determines the age of the voter, which is a numerical variable. Even though the result represents the year of birth, it indirectly indicates the age of the survey participant. Specific age groups may tend to prefer one party over the others because of their growing environment or from their experience. People tend to vote from previous experience with certain parties. Younger people and seniors may find different ideas for voting due to their expertise. Also, the younger people who do not have lots of interest about the politic will have other thoughts on the party compared to seniors. Also, if that is the case, the proportion of the participation matters. According to Elections Canada, the participation of voters aged 18 to 24 decreased by 3.2 percentage points to 53.9% in 2019 (Elections Canada, 2020). The age group between 35 to 44 saw the highest increase in voter turnout, with a rise of 2.7 percentage points to 64.6% (Elections Canada, 2020). Lastly, the voters aged 65 to 74 had the highest participation rate of 79.1% (Elections Canada, 2020). Therefore, it would be essential to investigate the relationship between the choice of the vote and the participant's year of birth regarding the age distribution of the participants.

The `cps19_gender` is also an essential variable to consider. It is a categorical variable. According to new Ipsos, there was a significant division between men's political leanings and women under 55 (Breen, 2019). For instance, the analysis resulted that women tended to lean towards the Conservative Party as they age (Breen, 2019). Therefore, gender and age are the key variables that influence the result of the election. Also, gender and age groups play a role together. For example, the two people in the same particular age group, men and women, will give very different results. Thus, gender and age are highly related to each other that need to be analyzed together.

The variable `cps19_province` is a categorical variable that will divide individuals into different regions. The response is collected from the question asking which province or territory the participant is currently living in (Stephenson et al., 2020). It will show how the vote result differs by region using a logistic regression model. In Canada, each province has a provincial government. People may be influenced by the experience with a specific party that can adjust the federal government's election result.

The `cps19_education` is a categorical variable that may influence the voting choice. It demonstrates the highest level of education that the survey participants have completed (Stephenson et al., 2020). People who have learned more about politics may lean to a particular party based on the parties' election pledge. People who are more familiar with politics and never studied politics and civics may have different perspectives. The interpretation of the evaluation of the previous party will be different, which may directly result in the difference in voting results.

The `cps19_income_number` shows the total household income before taxes for 2018 (Stephenson et al., 2020). It is a numerical variable that the income may show differences in party affiliation due to fiscal policies. People may prefer the federal government to collect less tax. On the other hand, some may like the federal government that provides more financial support, such as "CESB" or "CERB" during the financial emergency. The factor of income may make people lean towards a specific party for their own benefit.

Lastly, `cps19_religion` is a categorical variable that has a subtler influence in Canadian politics. It is because of the various laws that are based on religion. The most common two factors are same-sex marriage and abortion (CBC, 2019). People with strong religious affiliation may be motivated by their faith (CBC, 2019). For instance, the Roman Catholic church prohibits abortion because it is considered a sin (Schenker & Rabenou, 1993). As a result, people who do not believe in the Roman Catholic church and supports abortion are not likely to choose the party that does not allow abortion. However, people who believe in the Roman Catholic church will instead pick the party that supports their belief. Therefore, having religion may influence have the result of the election.

Numerical Summary

The numerical summary will be done on the age of the survey participant. It will give age distribution about the survey participant. It will provide what age group of people have participated in the election the most. The distribution of ages is vital while different generations may support other parties.

However, the number recorded is irrelevant while it recorded the birth year using the assigned number. For example, 1 indicates the people born in 1970, and 91 shows the people born in 2010. Therefore, to get the information of age, the data will be manipulated first.

The survey was conducted in 2019. Therefore, the benchmark age will be in 2019.

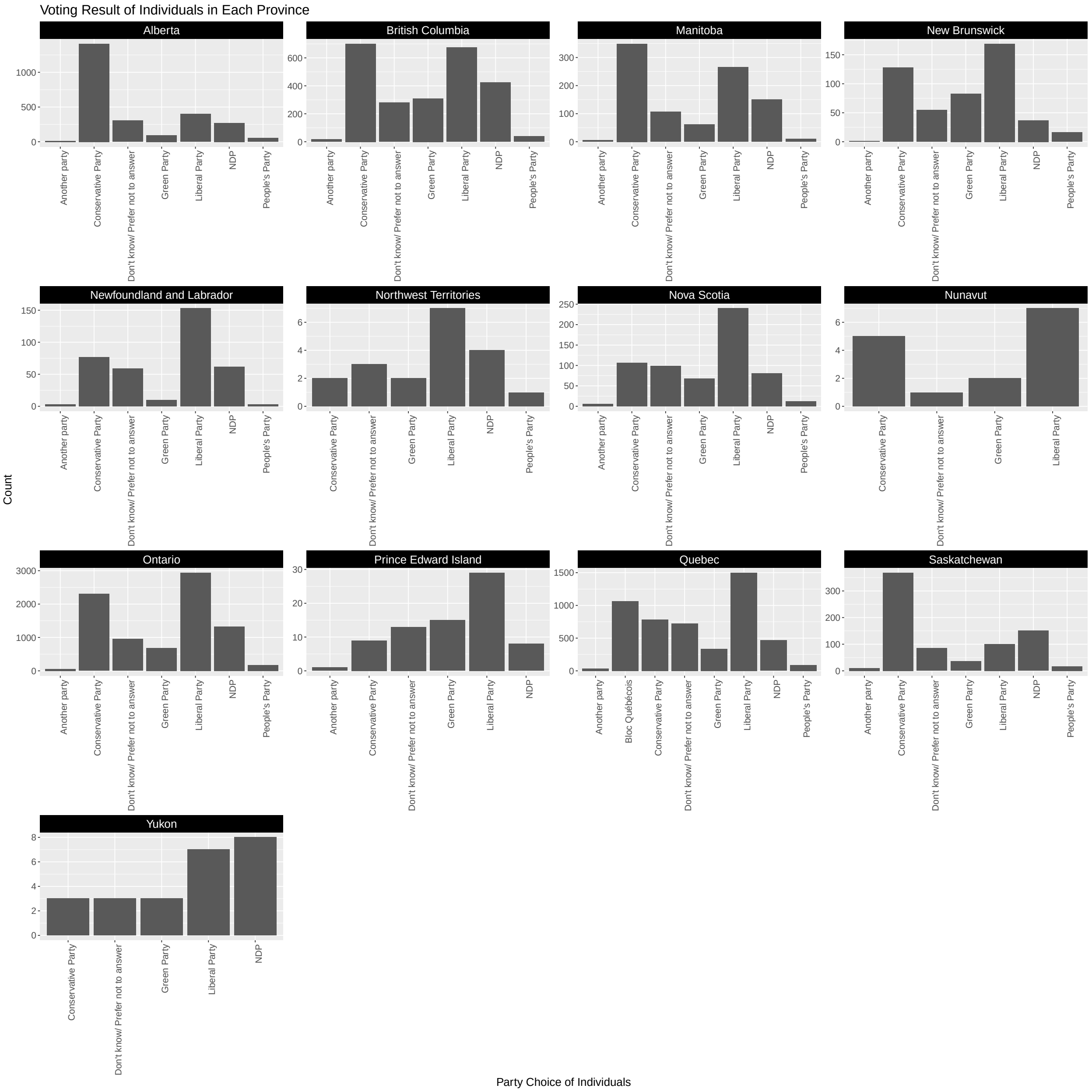
mean	median	sd	min	max
49.14567	50	16.02972	17	98

The mean and median age of the voting participants in 2019 was almost similar. It indicates the proportion of the age distribution is most likely not right or left-skewed. The maximum age who participated in the survey was 98, and the minimum age was 17. The standard deviation was quite significant, which is 16.010, that the age distribution is prominent from the mean.

Visualization

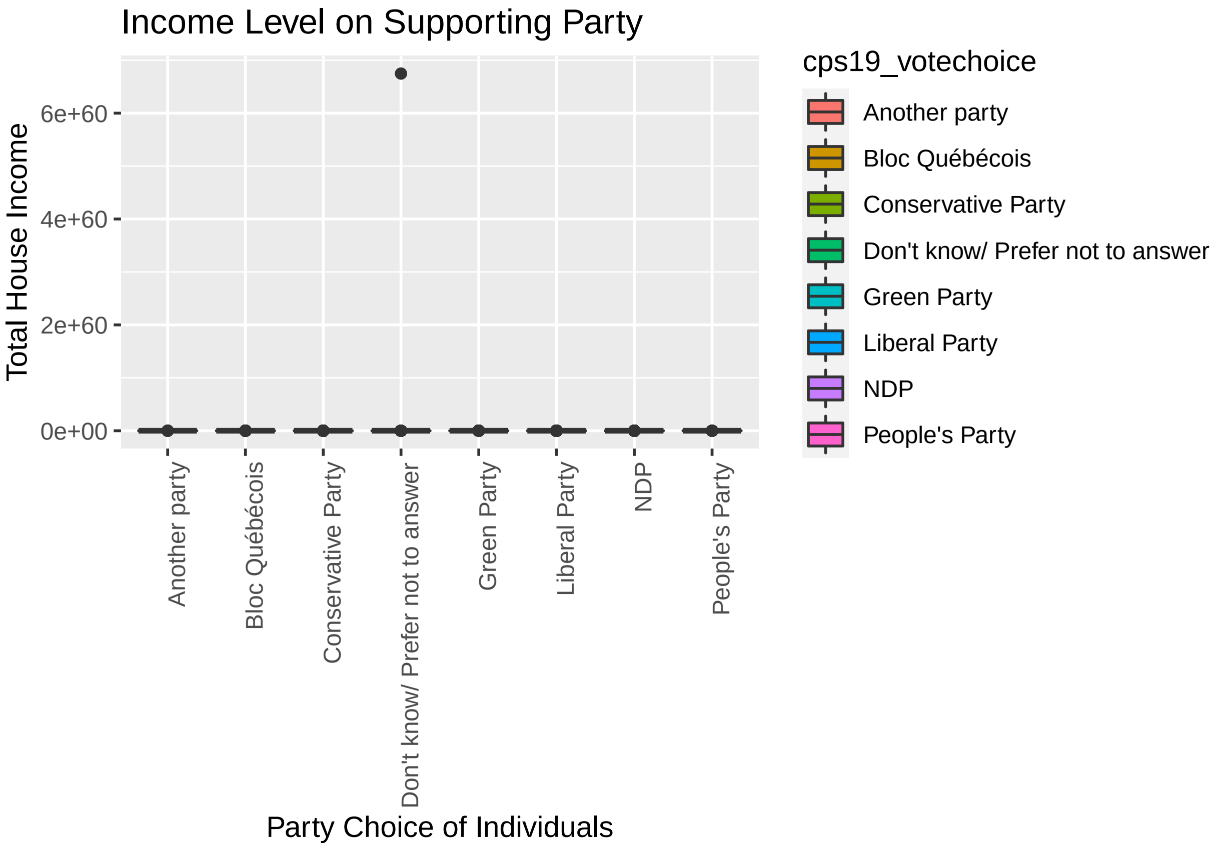
First, the `cps19_votechoice` in each region will be analyzed using the `survey_data`. The bar graph will be done but will be divided based on the province information, `cps19_province`.

The data is collected using the numbers. For instance, `14` represents Alberta, `15` represents British Columbia, and `16` represents Manitoba (Stephenson et al., 2020). The same applies to the vote choice for a party. For example, `1` represents the Liberal Party, and `2` illustrates the Conservative Party (Stephenson et al., 2020). Therefore, the data will be manipulated to make it easier for interpretation. Then, the graph will be plotted.



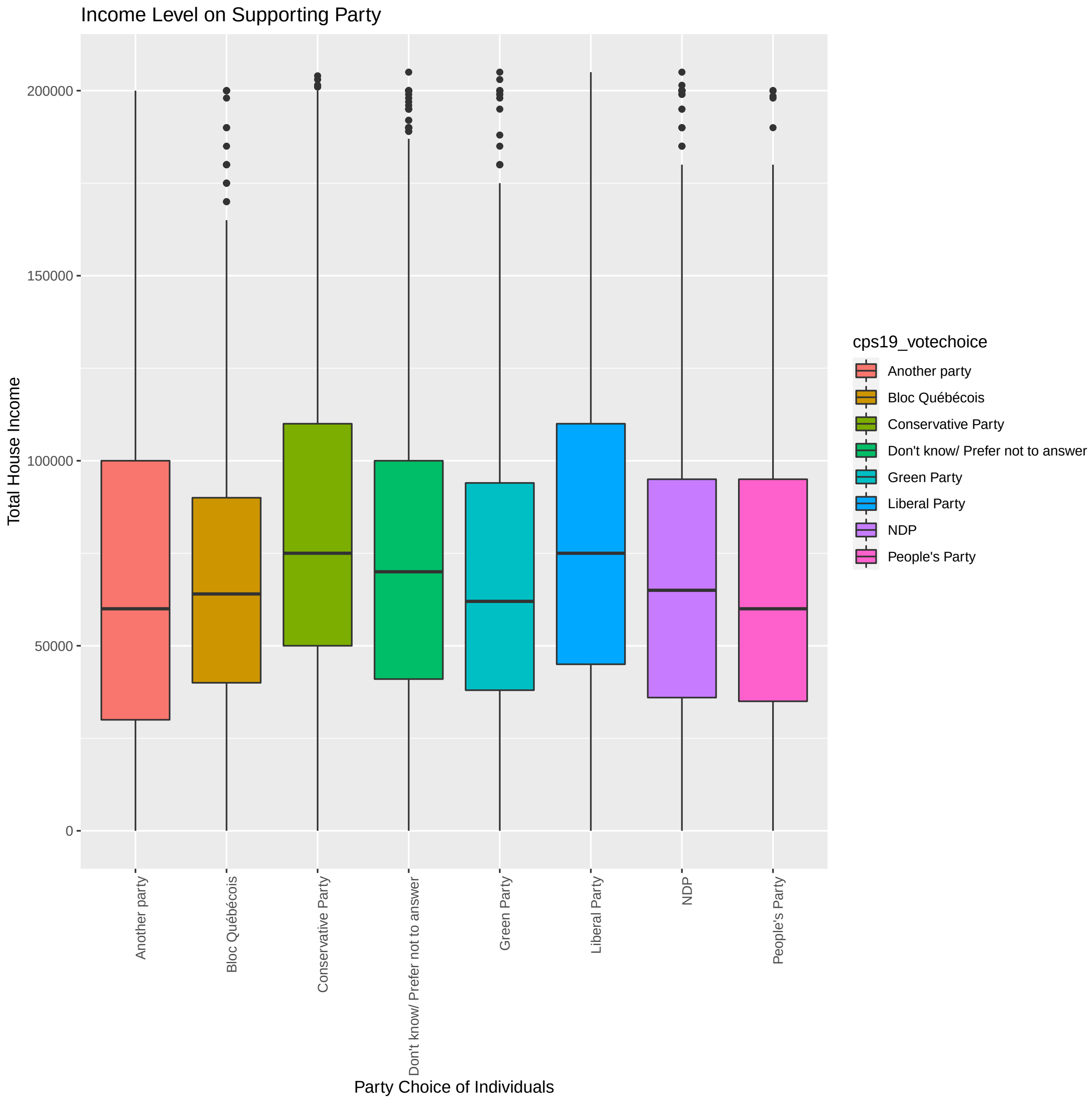
The graph demonstrated the choice of Party in each region. From the graphs, it is recognizable that each province support one or two particular parties. For example, the survey participants in Alberta tended to support Conservative Party in the most significant landslide. People who support Conservative Party in Alberta were more than half. On the other hand, provinces like British Columbia had the two most supportive parties Conservative and Liberal. Almost a similar proportion of people supported each Party. However, Yukon's survey result was different from the other provinces in that the people distribution rate was not as high. The most supporting Party and the least supporting Party's difference was not as high. Most of the provinces tended to support the Liberal or Conservative Party. However, only one region, Yukon, supported NDP the most.

The following graph will indicate the relationship between `cps19_income_number` and the `cps19_votechoice`. It will demonstrate how income influences the voting result without considering the other factors.



The graph above indicates the income level. However, it is clear that there is an extreme outlier. Due to the outlier, the result is not clear to see the distribution. Therefore, the outlier will be removed. A new graph will be generated below with the outlier.

All the outliers were removed before generating a new graph. It may include the points that were not shown in the previous graph visually. In total, 744 points were considered as outliers out of 21762 data points.

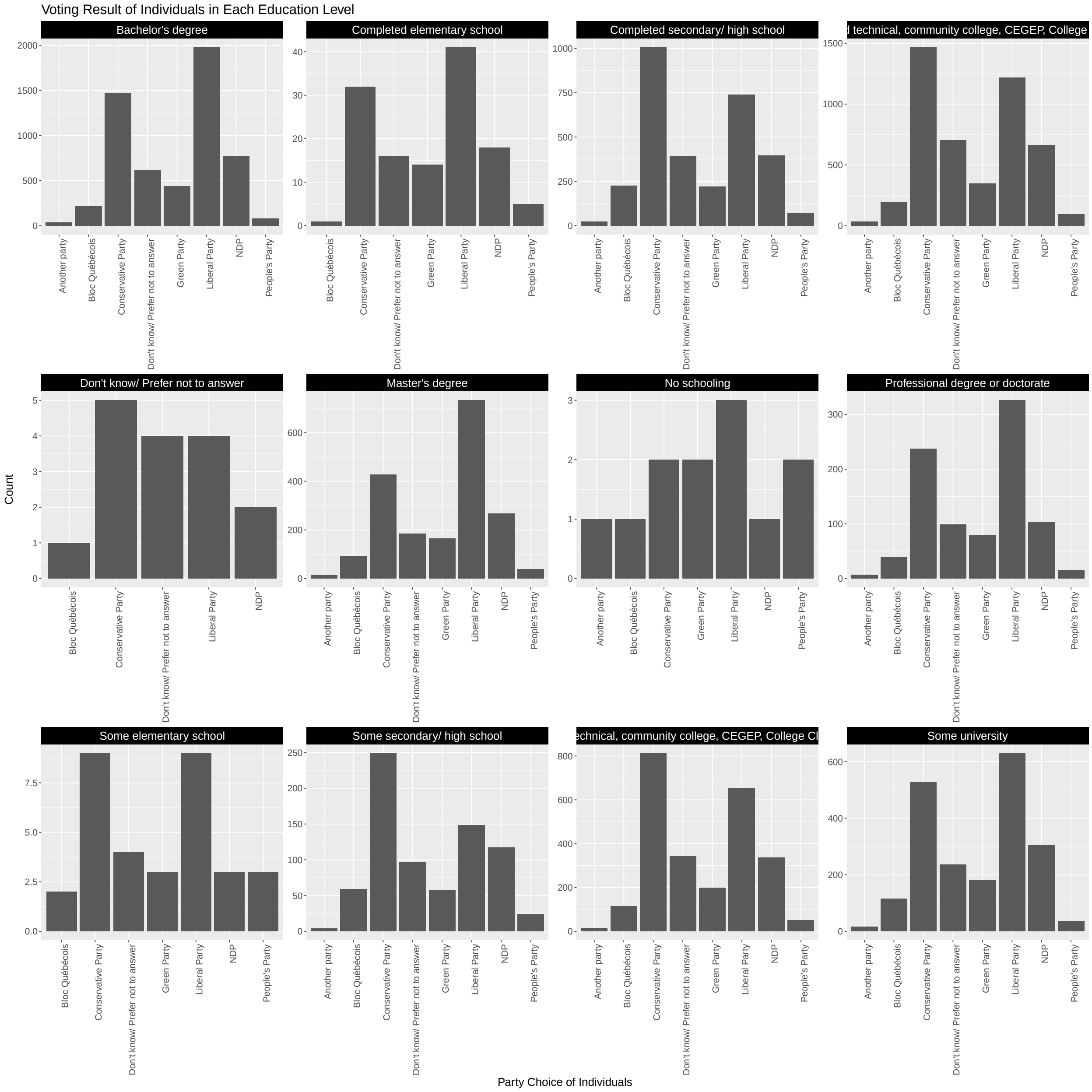


The boxplot indicates the total annual income in each household and the party choice of individuals using the new dataset without the outlier. The x-axis demonstrates the party choice of individuals. Then, the y-axis represents the total annual income in CAD. The bold middle line indicates the median value of the participants in each group. The median line was shown at a similar level in the boxplot. However, overall, the two Parties, Conservative and Liberal supporters, had the highest income. Instead, the Liberal Party had more widespread distribution of income compared to the Conservative Party. The points indicate the outliers. Even after removing the outliers in the previous step, Conservative Party had some outliers using the leftover data points. The rest of the Parties had a similar distribution in income.

The following graph will show the distribution of the `cps19_education`. Similar to the first graph with `cps19_province`, the multiple bar graphs will be generated by wrapping the same education level. Among the same education level, the Party that the individual supports are recorded for bar graphs.

Before generating a graph, the survey result will be converted into a practical format, using the actual word that describes the education level. For instance, 1 does not give any information about the education level.

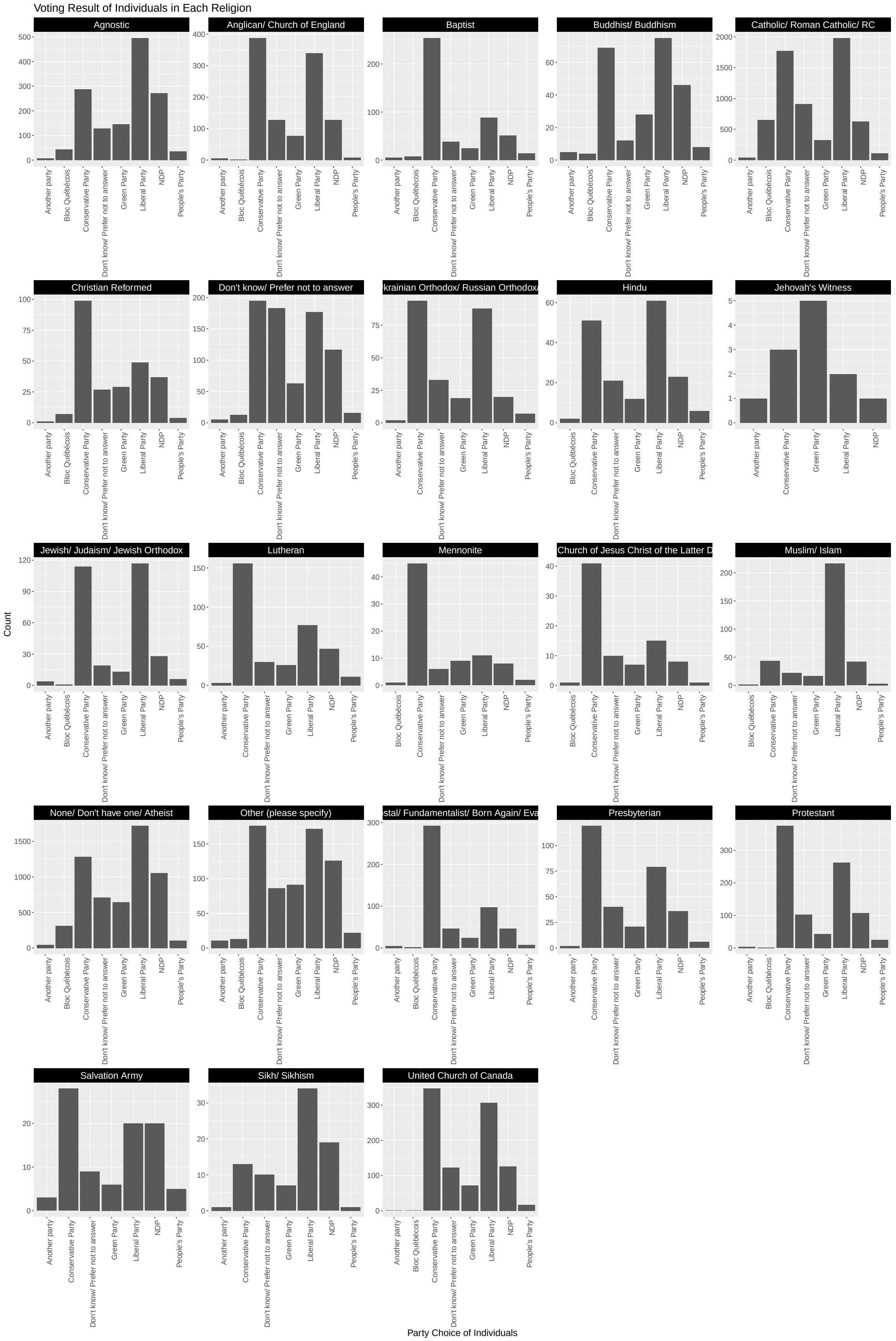
Therefore, using the guide book for the survey, it will be converted to 'No schooling' (Stephenson et al., 2020). 2 will be converted to 'Some elementary school', 3 will be converted to 'Completed elementary school', and so on (Stephenson et al., 2020).



Regarding all groups, a similar pattern was observed across most of the groups. No matter which level of education was completed, most groups supported the Liberal Party or the Conservative Party. The difference is that most groups tended to support one party over the other. Most people who completed a higher level of education, such as a professional degree or doctorate, Master's degree, Bachelor's degree tended or some university to support the Liberal Party more over the Conservative. However, the people who completed lower levels than the university did not show any pattern. Some groups tended to support Conservatives while some supported Liberal. Therefore, it will require more research to see the precise pattern of support. The new pattern was introduced in the group who did not get any schooling. They tended to support Green Party at a similar level to the Conservative Party. However, while the sample size of the people who did not get any schooling is small, a small sample size may cause sampling bias.

Lastly, bar graphs will be introduced again. The graphs will present how the religious affiliation influences the supporting Party. The Party that people support may vary due to their beliefs. The main focus these days is on abortion and same-sex marriage.

Again, the survey result will be converted into a practical format, using the actual description of religion. For example, 5 does not give any information about any religion. Therefore, using the survey guidebook will be converted to 'Jewish/ Judaism/ Jewish Orthodox' (Stephenson et al., 2020). The other numbers will be changed into an actual religion designation.


Voting Result of Individuals in Each Religion
Agnostic
Anglican/ Church of England
Baptist
Buddhist/ Buddhism
Catholic/ Roman Catholic/ RC
Christian Reformed
Don't know/ Prefer not to answer
krainian Orthodox/ Russian Orthodox/
Hindu
Jehovah's Witness
Jewish/ Judaism/ Jewish Orthodox
Lutheran
Mennonite
Church of Jesus Christ of the Latter D
Muslim/ Islam
None/ Don't have one/ Atheist
Other (please specify)
stal/ Fundamentalist/ Born Again/ Eva
Presbyterian
Protestant
Salvation Army
Sikh/ Sikhism
United Church of Canada
Count
Party Choice of Individuals
Another party
Bloc Québécois
Conservative Party
Don't know/ Prefer not to answer
Green Party
Liberal Party
NDP
People's Party

A similar pattern was observed across most of the groups. Most of the groups tended to support Liberal Party or the Conservative Party. However, some groups had more extreme results. For example, 'Baptist', 'Lutheran', or 'Mennonite' supported Conservative Party extremely, that the people who support the other Party were very minimal. Similarly, the religions such as 'Muslim/ Islam' or 'Sikh/ Sikhism' supported Liberal Party over the other parties. Some religions both Liberal Party and the Conservative Party. However, most religions preferred one Party over the other. More religions supported Conservative over Liberal. However, people who did not have any religion tended to support Liberal. While the graphs show the extreme results, the factor "religion" would be an essential point to consider to predict the future 2023 Federal Election.

Before proceeding to the next section, the `census_data` and `survey_data` were manipulated to make them have the common cells. For example, in the `survey_data`, the word that indicates male was 'A man'. However, in `census_data`, it was 'Male'. All the cell that included 'A man' in `census_data` was changed into 'Male' to avoid error in the later process. All the other columns went through the same process.

Methods

The goal of this report is to predict the popular vote of the 2023 Canadian federal election with the use of logistic modelling and post-stratification. We will create two logistic regression models, one to model the probability of a person voting for the Liberal Party and another one to model the probability of a person voting for the Conservative Party. In this case, a logistic model is appropriate since our dependent variable is going to be binary. It will tell us whether or not a person voted for the Liberal/Conservative Party with the variable only having two possible values. After creating our models we will post-stratify, allowing us to calculate the probability of each party winning the overall popular vote.

Model Specifics

Modelling the Probability of Voting for the Liberal Party

The first model we will create is the Bayesian logistic model that will predict the probability of a certain person voting for the Liberal Party, depending on their age, gender, religion, education level and the province of residence. All five variables will be treated as individual-level variables. Before creating this model, we must add another variable named `voted_liberal` to our survey data. This variable will have a value of 1 if a person voted Liberal in 2019 and a value of 0 otherwise. `voted_liberal` is the key to allowing us to create a logistic model.

Now that we have our binary dependent variable created, we can create our Bayesian model as the following:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{gender} + \beta_3 x_{religion} + \beta_4 x_{education} + \beta_5 x_{province}$$

In the model above, $log(\frac{p}{1-p})$ represents the logarithm of the odds for a certain person voting for the Liberal Party given their age, gender, religion, education, and province. Later on, in the Methods section, we will use these log odds to calculate the probability of that certain person voting Liberal in the upcoming election. On the other side of the equation, β_0 represents the model's intercept. This tells us the log odds of a person with an age of 0 and no gender, religion, education or province of residence voting for the Liberal Party. Furthermore, β_1 represents the change in log odds of voting Liberal for a one-year increase in age. Similarly, β_2, β_3, β_4, and β_5 also represent the change in log odds of voting for the Liberal Party when values for gender, religion, education and province are changed, respectively. Finally, x_{age}, x_{gender},$x_{religion}$, $x_{education}$ and $x_{education}$ are the values for a person's specific age, gender, religion, education level and province of residence, respectively.

Modelling the Probability of Voting for the Conservative Party

The next model that we will create is the Bayesian logistic model that will predict the probability of a certain person voting for the Conservative Party. We will be using the same variables as the model above:

age, gender, religion, education level and the province of residence. Similar to the Liberal model, we must add and include an extra variable named `voted_conservative` that will have a value of 1 if a person voted Conservative in 2019 and a value of 0 otherwise. This variable will once again allow us to create our logistic regression.

After creating `voted_conservative`, we can continue onto creating the model:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{gender} + \beta_3 x_{religion} + \beta_4 x_{education} + \beta_5 x_{province}$$

In this model, $log(\frac{p}{1-p})$ now represents the logarithm of the odds for a certain person voting for the Conservative Party given their age, gender, religion, education, and province. On the right-hand side, β_0 once again represents the random intercept that tells us the log odds of a person with an age of 0 and no gender, religion, education, or province voting Conservative. Similar to the Liberal model, β_1 represents the change in log odds of voting Conservative for a one-year increase in age. Additionally, β_2, β_3, β_4 and β_5 represent the change in log odds of voting for the Conservative Party when the values for gender, religion, education, and province are altered, respectively. Lastly, x_{age}, x_{gender},$x_{religion}$, $x_{education}$ and $x_{education}$ are the actual values for person's specific age, gender, religion, education level and province of residence, respectively.

Post-Stratification

After having done the regression modelling, we outsource our model to the use of post-stratification in order to provide estimates which, rather than only using raw counts in stratum, are estimates fully based on combinations of predictors. Moreover, the technique is based on adjusting a sample that is non-representative where demographic predictors characterizing the strata are present. Therefore, in simple words, it uses demographics in order to "extrapolate" how, in this case, the entire population will vote in the upcoming Canadian elections.

Mathematical Notation To estimate the proportion of voters for each party, the post-stratification formula which will be used is:

$$\hat{y}^{PS} = \frac{\Sigma N_j \hat{y}_j}{\Sigma N_j}$$

Where: $\hat{y}^{PS}$ = the estimate of the Post-Stratification, $\hat{y}_j$ = the estimate in each cell (province or territory), and N_j = the population size of the j^{th} cell (province or territory) based off demographics.

Post-stratification is highly beneficial since it adjusts the estimates (like weight averaging) from the possible combinations of all the attributes. Therefore, it corrects the model estimates for known differences between target population, sample, and population. Thus, it helps us increase the accuracy of the estimates of the model.

Post-Stratification Variables The cells in the post-stratification will be the Canadian provinces and territories, which is a total of 13. The independent variables which will be used, as mentioned under the Data section, are: `cps19_yob`, `cps19_gender`, `cps19_province`, `cps19_education`, , `cps19_income_number`, `cps19_religion`. These variables are the age of the voter, gender, province of residence, the highest level of education that the survey participants have completed, the total household income before taxes for 2018, and religion, respectively.

However, in this case, the `csp19_province` will be used in the post-stratification as the cells when predicting if the population will vote for the liberal party or the conservative party. We use the province as the cells' split in the post-stratification because it is very likely that their difference in locations can heavily influence voter election parties. This is due to the difference in industries that each province excels in (e.g. agriculture, services, natural resources, fishing, etc.). Therefore, different proposals and/or problems that each province faces can heavily influence whether the voter is satisfied with the current provincial government or not. Lastly, a provincial government that has succeeded in another province can influence other provinces to adopt a similar party if they do not have it at the moment.

With regards to the assumptions for the variables, one assumption is that the observations in the sample are independent of each other. Thus, each observation is by no means related nor influenced by others for each sample subject. This assumption is true since the sample of observations is composed of different people in it and not by the same person.

Therefore, after stating the reasons as to why we use the province as the cell split, the independent variables, and the correct assumptions of the variables, we can determine that the method is appropriate for the estimation of the entire population vote for the upcoming Canadian election.

Results

Bayesian Logistic Regression Models

Term	Estimate	Std. Error	Lower 95% CI	Upper 95% CI
Intercept	-0.8915726	0.2921967	-1.4962796	-0.3405960
Age	0.0049629	0.0009790	0.0030508	0.0068697
Male	-0.0149888	0.0309019	-0.0766016	0.0438414
Has religious affiliation	0.3110745	0.0902362	0.1344843	0.4887094
No religious affiliation	0.3142775	0.0926629	0.1379719	0.5002263
College or other non-university certificate/diploma	-0.4644935	0.0438108	-0.5510002	-0.3806032
Prefer not to answer (education)	-0.4421625	0.6085213	-1.7074012	0.6864846
High school diploma	-0.5850481	0.0486184	-0.6782033	-0.4915647
Less than a highschool diploma	-0.2317139	0.1714937	-0.5767380	0.1041416
Trade certificate or diploma	-0.4347710	0.0540579	-0.5439163	-0.3260971
University certificate/diploma below bachelors	-0.1876245	0.0561836	-0.2984479	-0.0775569
University certificate/diploma above bachelors	0.0545279	0.0482102	-0.0385231	0.1494600
Alberta	-1.0566091	0.2815437	-1.5988729	-0.4780879
British Columbia	-0.3708613	0.2773406	-0.8935854	0.1935111
Manitoba	-0.3197857	0.2850457	-0.8681482	0.2554501
New Brunswick	-0.0217024	0.2894433	-0.5915911	0.5587970
Newfoundland and Labrador	0.2964562	0.2952964	-0.2701831	0.9004224
Nova Scotia	0.1733350	0.2880479	-0.3649015	0.7578910
Ontario	-0.0364910	0.2755155	-0.5614006	0.5242422
Prince Edward Island	0.1970376	0.3695254	-0.5341232	0.9312824
Quebec	-0.2387631	0.2771255	-0.7670887	0.3277861
Saskatchewan	-1.2737710	0.2933282	-1.8370581	-0.6812359

Table 1. Summary of the regression model predicting the probability of a certain person voting for the Liberal Party.

The table above showcases the outputs of the regression model predicting the probability of a certain person voting for the Liberal Party. The slope estimates represent the change in the expected log odds of voting Liberal when values for age, gender, religion, education, and province are changed, respectively. In other words, this fitted model says that holding everything else at a fixed value, the odds of a person voting Liberal given that they live in Nova Scotia (Nova Scotia = 1) over the odds of a person voting Liberal given that they do not live in Nova Scotia (Nova Scotia = 0) is exp(0.17938) = 1.1965; and the coefficient for age says that holding everything else at a fixed value, we will see a 0.5% increase in the odds of observing a person voting Liberal for a one-unit increase in age since exp(0.004949) = 1.00496. Therefore, a positive value for the coefficient estimate entails an odds ratio greater than 1, meaning an increase in the probability of the event when we have a positive change in the independent variable; and a negative value for the coefficient estimate entails an odds ratio less than 1, meaning a decrease in the probability of the event when we have a positive change in the independent variable.

As seen in Table 1, for one year of increase in age, the odds of a person voting for the Liberal Party increases by 1.004912 (95% CI: 1.0030045 - 1.0069239). This is unexpected, as the younger generation is generally thought of as more progressive and liberal in their political views; hence, we expect to see the odds of voting Liberal decrease as age increases. This discrepancy between expectation and result may be due to a low survey participation rate among younger people, making the sample of younger people, not representative of the actual younger population in Canada due to the small sample size for younger people. Another plausible explanation is that the younger population is more likely to vote for non-major left-leaning political parties (e.g. New Democratic Party and the Green Party) than the older population. Therefore, as age increases, the likelihood of voting for a major political party (in this case, the Liberal Party) increases.

In addition, the odds of a person voting for the Liberal Party, whether that person has religious affiliation or not, increases nonetheless; this means that whether or not someone has religious affiliation is not a good predictor of whether that person will vote Liberal or not.

With regards to education, Table 1 indicates that the log odds of a person voting for the Liberal party decreases if that person has the highest level of education, anything below a University Bachelor's degree. On the other hand, the odds of a person voting for the Liberal party increases by 1.0554846 (95% CI: 0.9609816 - 1.160673) if that person has a university certificate or diploma above the bachelors level.

With regards to the province, the most notable results are Alberta and Saskatchewan. The odds of a person voting for the Liberal party decreases by 0.3478444 (95% CI: 0.206594 - 0.5850839) if that person resides in Alberta, and decreases by 0.2808316 (95% CI: 0.1636541 - 0.4906619) if that person resides in Saskatchewan. These results are unsurprising, as Alberta and Saskatchewan have always been Conservative-majority provinces in recent elections due to dominant sociopolitical and economic discourses.

Term	Estimate	Std. Error	Lower 95% CI	Upper 95% CI
Intercept	-2.4916618	0.3833262	-3.2840214	-1.7919243
Age	0.0061463	0.0010304	0.0041133	0.0081397
Male	0.5271662	0.0324334	0.4655708	0.5894878
Has religious affiliation	0.4287545	0.0894020	0.2561948	0.5997901
No religious affiliation	-0.0997745	0.0938757	-0.2808714	0.0827274
College or other non-university certificate/diploma	0.2429655	0.0465778	0.1508684	0.3357012
Prefer not to answer regarding education	-0.0121247	0.5952318	-1.2625138	1.0952391
High school diploma	0.3224585	0.0490448	0.2286964	0.4176829
Less than a highschool diploma	-0.0750514	0.1895261	-0.4505367	0.2803989
Trade certificate or diploma	0.2708442	0.0551681	0.1622625	0.3810260
University certificate/diploma below bachelors	-0.0437714	0.0615985	-0.1660172	0.0746178
University certificate/diploma above bachelors	-0.1714812	0.0569374	-0.2821877	-0.0596418
Alberta	1.8155028	0.3714373	1.1421899	2.5946606
British Columbia	0.6489573	0.3736240	-0.0215506	1.4427410
Manitoba	1.0051310	0.3765895	0.3156273	1.7969943
New Brunswick	0.4624696	0.3840270	-0.2379090	1.2538353
Newfoundland and Labrador	0.1358500	0.3940553	-0.6007989	0.9427157
Nova Scotia	-0.0466466	0.3848709	-0.7587649	0.7349648
Ontario	0.5495588	0.3706676	-0.1213525	1.3286979
Prince Edward Island	-0.6264308	0.5237648	-1.6506080	0.3850510
Quebec	-0.2069255	0.3729961	-0.8705910	0.5790805
Saskatchewan	1.4616829	0.3771000	0.7802570	2.2450540

Table 2. Summary of the regression model predicting the probability of a certain person voting for the Conservative Party.

The table above showcases the outputs of the regression model predicting the probability of a certain person voting for the Conservative Party. The slope estimates represent the change in the expected log odds of voting Conservative when values for age, gender, religion, education, and province are changed, respectively. As seen in Table 2, for one year of increase in age, the odds of a person voting for the Conservative Party increases by 1.0061186 (95% CI: 1.0041084 - 1.0081329). This is unsurprising, as the younger generation is generally thought of as more liberal and the older generation as more conservative in their political views. Hence, it is unsurprising to see the log odds of voting Conservative increase as age increases.

In addition, the odds of a person voting for the Conservative Party increases by 1.5403351 (95% CI: 1.3021282 - 1.8239418) if that person has religious affiliation. This seems plausible, as different religions often have their own dominant or affiliated sociopolitical views, and it is possible that people with religious affiliation generally find that the Conservative party's policies align best with their sociopolitical views and religious principles.

With regards to education, Table 2 indicates that the odds of a person voting for the Conservative party increases by 1.2776213 (95% CI: 1.1641602 - 1.399339) if that person has the highest level of education of a college or other non-university certificate or diploma, increase by 1.3826473 (95% CI: 1.2586 - 1.5204404) if that person has the highest level of education of a high school diploma, and increases by 1.312587 (95% CI: 1.2586 - 1.5204404) if that person has the highest level of education of a trade certificate or diploma. On the other hand, the odds of a person voting for the Conservative party decreases by 0.8442556 (95% CI 0.7563886 - 0.9442257) if that person has a university certificate or diploma above the bachelors level.

With regards to provinces, the most notable results are Alberta, Manitoba, and Saskatchewan. The odds of a person voting for the Conservative party increases by 6.2229869 (95% CI 3.2344212 - 13.5072962) if that person resides in Alberta, increases by 2.7662151 (95% CI 1.4141095 - 5.9858599) if that person resides in Manitoba, and increases by 4.3652262 (95% CI 2.2003578 - 9.3403152) if that person resides in Saskatchewan. These results are unsurprising, as Alberta, Manitoba, and Saskatchewan have predominantly been Conservative-majority provinces in recent elections due to dominant sociopolitical and economic discourses.

Post-stratification

cps19_province	lib_prob	cons_prob
Alberta	0.5315380	0.5428187
British Columbia	0.5519543	0.5526466
Manitoba	0.5247007	0.5268288
New Brunswick	0.5294012	0.5279707
Newfoundland and Labrador	0.5253493	0.5217525
Nova Scotia	0.5324636	0.5272538
Ontario	0.6211531	0.6157267
Prince Edward Island	0.5160554	0.5123753
Quebec	0.5809427	0.5711438
Saskatchewan	0.5202850	0.5277256

Table 3. Table showcasing the probabilities of a given province voting for the Liberal party versus the Conservative party.

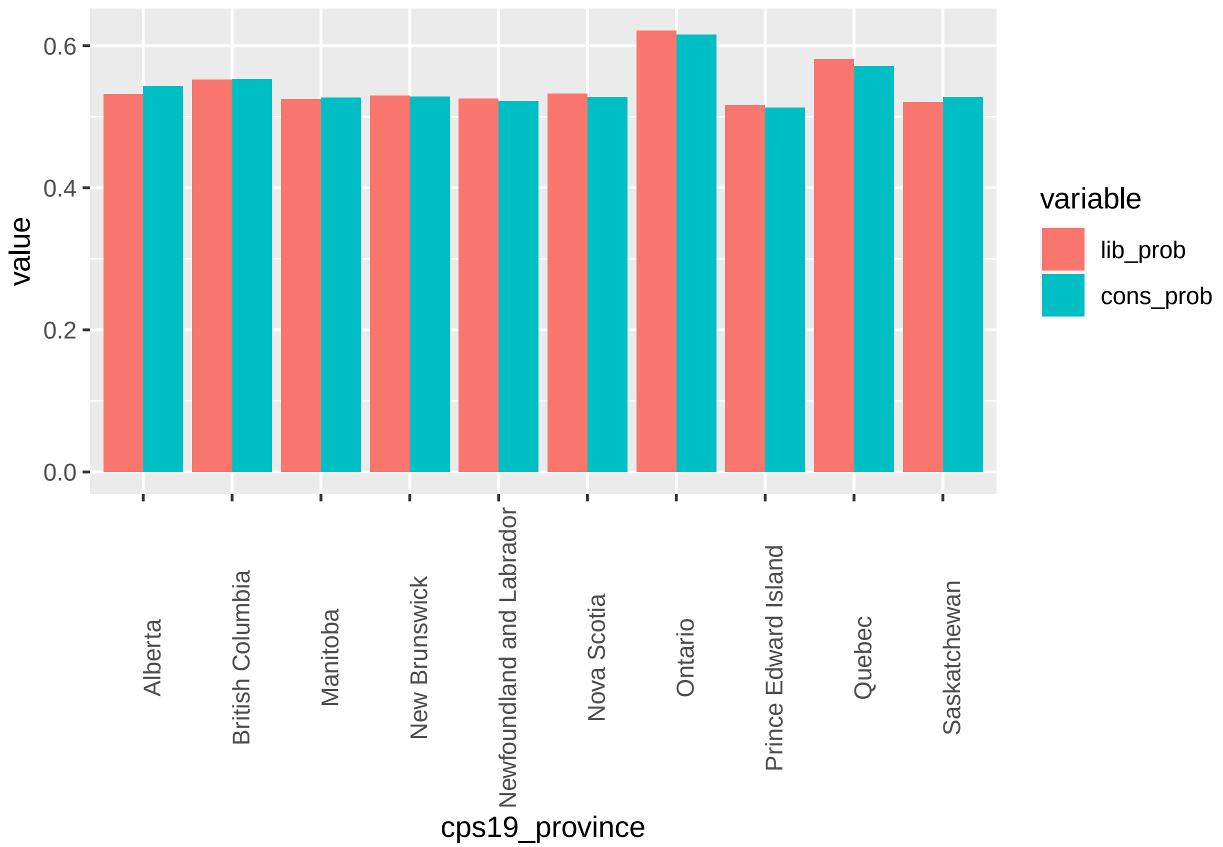


Figure 6. Bar chart comparing the probabilities of a province voting for the Liberal Party versus the Conservative Party.

After back calculating the probability of voting for the Liberal Party from the log odds, we obtain the above table. As seen in Table 3 and Figure 6, Ontario has the highest probability of voting for the Liberal party, and Prince Edward Island has the lowest probability of voting for the Liberal Party. Moreover, Ontario also has the highest probability of voting for the Conservative party, and Prince Edward Island also has the lowest probability of voting for the Conservative Party. This seems to suggest that people in Ontario are most likely among all provinces to vote for a major political party (liberal/conservative), and people in Prince Edward Island are least likely to vote for a major political party. This seems appropriate; Ontario has the largest population in the country and has the highest electoral quotient in the country (electoral quotient = 111,144), hence people may be more motivated to "compete" against a large number of people within their own province to secure a majority within the province. On the other hand, Prince Edward Island has the lowest electoral quotient among all provinces (electoral quotient = 35,726). Thus, people in Prince Edward Island may be less motivated to intentionally vote for a majority party in order to secure a majority within their province. ("House of Commons of Canada", 2021)

pred_lib	pred_cons
0.565724	0.5632055

Table 4. Post-stratified probabilities of observing a person voting for the Liberal party vs the Conservative party.

After performing post-stratification based on the provinces, we finally obtain the above probabilities. `pred_lib` is the probability of a person voting for the Liberal party, and `pred_cons` is the probability a person voting for the Conservative party. As we can see, according to our models, the Liberal party has a slightly higher probability of winning the 2023 Canadian federal election. This result seems to confirm our original prediction that the Liberal party will win the election. This result seems appropriate as a vast majority of opinion polls conducted by different polling firms for the 2023 Canadian federal election indicating that the Liberal Party is leading ("Opinion polling for the 44th Canadian federal election," 2021). In fact, all of the opinion polls conducted after October 2020 indicate that the Liberal Party is in the lead. Therefore, our final result agrees with current opinion polls for the 2023 Canadian federal election.

Conclusions

Now that we have created, run and collected the results of our two logistic regressions with post-stratification, we can determine the accuracy of our initial hypothesis and make a statement regarding the projected winner of the 2023 Canadian federal election's popular vote. Near the beginning of this report, the data collected from many different opinion polls regarding the 2023 federal election caused us to hypothesize that the Liberal Party would win the popular vote. We then created two Bayesian logistic regression models, one that would model the probability of a certain person voting for the Liberal Party and the other that would model the probability that a person would vote for the Conservative Party. Both of these probabilities would be based on a person's age, gender, religion, education level and the province of residence. We then ran both models on the population and post-stratified based on province, using the GSS Census Data. When looking at the results, we saw that our Liberal model estimated that a person would have a 56.6% chance of voting for the Liberal Party in the next election, therefore winning the popular vote. Our Conservative model surprisingly gave us a similar number. We saw that our Conservative model estimated that a person would have a 56.3% chance of voting Conservative in the next election, signalling that the Conservatives would win the popular vote. Although it may seem impossible that our results show that both parties have a greater than 50% chance of winning, there is a definite explanation. The reason for both of these values being greater than 50% is that the two models we ran were independent of each other, meaning that the value produced by the Liberal model has no effect on the value created by the Conservative model, and vice versa. Each model is only concerned with the probability of one party winning and is not attempting to determine the percentage of votes any other party will receive.

When looking at the results of our two models, it is difficult to determine whether our initial hypothesis was correct. In the Liberal model, the Liberals are projected to win, while in the Conservative model, the Conservatives are projected to win. Even though in the Liberal model, people are predicted to vote Liberal with a 56.6% chance, which is 0.3% more than the Conservatives are estimated to have in the Conservative model, we are unable to conclude that the Liberals are estimated to win the upcoming election. This is, once again, due to the two models being independent. One other potential weakness of this report is that we are using data that was collected in 2019 in hopes of estimating federal election outcomes in 2023. This has most likely had a damaging effect on the accuracy of our models as data can change greatly in the span of four years, especially during a global pandemic. The data we used in this report will most likely not be a true representation of the Canadian population in 2023. An interesting opportunity for the future would be to run an analysis similar to the one in this report with a model that would calculate the proportion of votes that each party will receive as opposed to calculating the proportion of votes for only one certain party. This, paired with data from a year closer to 2023, could give us a clearer picture of who we can expect to win the upcoming federal election.

Despite our inability to come to a clear conclusion on who can be expected to win the 2023 Canadian federal election, we were able to witness how logistic regression and post-stratification can be used to make predictions and gain insight into significant happenings in our daily lives. Moreover, we were shown how two nearly identical models using the same data are capable of producing such contrasting results. Overall, this report exposes the everlasting uncertainty and doubt that is present in Canadian politics.

Bibliography

1. Schenker, J. G., & Rabenou, V. (1993) *Contraception: traditional and religious attitudes.* European journal of obstetrics, gynecology, and reproductive biology, 49(1-2), 15–18. https://doi.org/10.1016/0028-2243(93)90102-i. (Last Accessed: May 22, 2021)

2. Elections Canada. (2020, August 11) *Voter Turnout by Sex and Age* https://www.elections.ca/content.aspx?section=res&dir=rec/eval/pes2019/vtsa&document=index&lang=e. (Last Accessed: May 24, 2021)

3. Breen, K. (2019, October 4) *The 'genderation' gap: political divisions exist between men, women, different age groups, polls show.* Global News. https://globalnews.ca/news/5988160/genderation-gap-political-divisions-men-women/. (Last Accessed: May 24, 2021)

4. Stephenson, L., et al. (2020, April 29) *Canadian Election Study 2019 Online Survey Codebook.* Harvard Dataverse. (Last Accessed: May 24, 2021)

5. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller. (2021). *dplyr: A Grammar of Data Manipulation.* https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr. (Last Accessed: May 24, 2021)

6. Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. International Journal of Forecasting, 31(3), 980-991.

7. Opinion polling for the 44th Canadian federal election. (2021, May 27). In Wikipedia. https://en.wikipedia.org/wiki/Opinion_polling_for_the_44th_Canadian_federal_election. (Last Accessed: May 27, 2021)

8. House of Commons of Canada. (2021, May 28). In Wikipedia. https://en.wikipedia.org/wiki/House_of_Commons_of_Canada#Members_and_electoral_districts. (Last Accessed: May 28, 2021)

9. Stephenson, L., et al. (2020). 2019 Canadian Election Study - Online Survey. Harvard Dataverse. V1. https://doi.org/10.7910/DVN/DUS88V.(Last Accessed: May 28, 2021)

The packages used: tidyverse, dplyr, reshape, brms, broom *The tidyverse, dplyr and reshape were used to clean the dataset and draw the graphs, such as `ggplot`.* The brms and broom packages were used for model building process.

All analysis for this report was programmed using `R version 4.1.0`, R markdown.