Forecasting 2025 Canadian Election Using 2019 Election Survey Data

STA304 - Assignment 3

GROUP NUMBER 13: Cong Liu, Yunfei Yu

November 5, 2021

Introduction

The federal government, along with its political orientation, is a crucial component of a country's development and future outlook. On the world stage, a shift in the federal government and political orientation can lead to changes in international relations and in the world position of a country; on the individual level, changes to the legal system and policies regarding environment, social welfare, immigrants and many other aspects can affect the well-being and living environment of every citizen in the country. Hence, the transition from one government to another, among other issues, becomes the issue of primary concern. For countries like Canada, transitions of the federal government are implemented via federal elections. According to the *Canada Elections Act*, federal elections are to be held on the third Monday of October every four calendar years (*Consolidated federal laws of Canada, Canada elections act* 2021). When a federal election is held, each political party decides on a candidate for each electoral district, also known as riding, who will try to convince the voters to vote for them during the campaign period; electors then vote for the candidates at the poll on the election day and candidate who win the most votes in each riding is the declared winner (*About Federal Elections* 2021).

In the 44th Canadian federal election held on 2021 September 20, Justin Trudeau, the leader of the Liberal Party of Canada, has won his third term as the Prime Minister since 2015. Despite winning, Trudeau leads a minority government which means that his Liberal Party does not hold the majority of seats in the House of Commons (Azzi & Kwavnick, 2012). By the end of the 44th election, the Liberal Party won in 160 electoral districts or ridings and the Conservative Party of Canada, which has been a powerful competitor of the Liberal Party, won in 119 ridings (Mason & Urback, 2021). Hence, as the Trudeau government has won three terms of office in a row, one would be curious about whether he is able to serve for a fourth term or would the Conservative Party win the next election.

Thus, our goal is to provide some insight into the possible outcome of the 2025 Canadian election. Given the pattern observed in the last several elections, we hypothesize that the winner of the next election is likely to be either the Liberal Party or the Conservative Party; however, it is unlikely that a majority government will be elected. To test the hypothesis, we will first analyze the outcome of the 2019 election based on data provided by the 2019 **Canadian Election Study (CES)** and then try to predict the 2025 election outcome based on 2019 results. Specifically, regression analysis will be conducted on the 2019 Canadian Election Study and the regression model created will then be applied to 2017 **General Social Survey (GSS)**data. Using post-stratification techniques, which will be explained in detail in the following sections, 2017 General Social Survey data will be stratified into cells or groups. Within each cell, the estimated voting proportion for each party will be calculated and then aggregated to yield a prediction on the population level. We expect that variables such as the province of residence, age group and religious beliefs would affect one's political attitude and thus the voting outcome. Hence, we would devote more attention to these variables, along with other demographic variables, while analyzing the data.

Finally, before ending this section, we would like to provide a brief overview of the report. The following **Data** section provides an introduction to, and some numerical summaries of, the 2019 Canadian Election Study dataset and the 2017 General Social Survey datasets, along with some cleaning that is done to the datasets to ensure the data is analyzable and can be interpreted by the reader. The **Method** and **Post-stratification** sections explain our choice of the regression model and rationale behind post-stratification techniques. The actual numerical results of the analysis and their interpretation will be included in the **Results** section, along with tables and graphs showcasing the results. Finally, comments on the results, limitations of the analysis and possible directions for future research are discussed in the **Conclusion** section.

Data

In order to perform post-stratification, two datasets are needed for this report where a set of survey data will be used to train the regression model and a set of census data to fit the model and to perform the actual analysis. To be more specific, the survey data used in model training comes from the 2019 Canadian Election Study, or CES2019 in short, which is a survey on political attitude and voting inclination (Stephenson et al.). The whole survey records responses of 4021 respondents (observations) to 278 questions (variables) where the answer to each question occupies a column. Questions of the survey include those that measure political attitudes and voting preferences, such as "which party did you vote for (in the last election)", "how do you feel about Justin Trudeau" and "where would you place the Bloc Québécois" (Stephenson et al); demographic questions asking for birth year, income level and household size of the respondents are also included in the survey in addition to politics-related questions (Stephenson et al). Data of CES2019 are collected via a non-probability computer-assisted telephone survey (Stephenson et al) consisting of key questions drawn from previous CES surveys to ensure continuity on issues such as vote inclination (Canadian Election Study). Since the survey is not probability-based, that is, not every individual in the population, namely, Canadians, had an equal and non-zero probability of being included in the survey, the results of analysis based on the data may be subjected to biases. For instance, some portion of the population may be under or overrepresented than others. However, since a rather large sample is included in the survey, the size and effect of the bias may be mitigated.

While CES2019 data is used to train the regression model, (near) census data from the General Social Survey or GSS is used for post-stratification. The general social survey is a telephone survey conducted annually over the ten provinces of Canada, excluding Yukon, Nunavut and Northwest Territories, to monitor and analyze trends in Canadians' living conditions and well-being and to represent any existing concerns with social policies (Statistics Canada, 2020). In particular, the 2017 GSS which is the census data used in this analysis, contains demographic information on respondents themselves as well as their family members (Statistics Canada, 2020). Information and data on 82 questions (variables) regarding respondents' age, sex, household and family characteristics are collected from 20602 respondents (observations). The survey is targeted at all non-institutional people living in the ten Canadian provinces aged 15 and older (Statistics Canada, 2020), from which the respondents are sampled. Specifically, the respondents of the survey are sampled from a list of combination of telephone numbers and Statistics Canada's Address Register (Statistics Canada, 2020). Again, data from 2017 GSS may be subjected to biases and drawbacks inherent from the collection process and the largely volunteer-based sample. One example of such drawbacks is the relatively low response rate, and indeed, the response rate of 2017 GSS is 52.4% (Statistics Canada, 2020); also, residents of Yukon, Nunavut and NWT are not included in the survey. Due to these drawbacks, the sample may be biased. However, as in the case with CES2019, the size and effect of these biases may not be as significant due to a rather large sample size of over 20000.

For analytic purposes, both CES2019 and 2017 GSS datasets are cleaned and reduced. Certain observations with invalid inputs are removed and some variables are recoded to ensure the results are interpretable. The first step of the data cleaning process regards selecting a) the response variable, namely, the party that a respondent voted for and b) predictor (independent) variables that are present in both CES and GSS datasets so that post-stratification methods can be carried out. In particular, variables p3, q2, q3, q4, p50, q62 from CES and variables age, sex, marital status, whether the respondent has a religious affiliation, citizenship status and province of residence from GSS are selected to be included in the cleaned dataset. The selection of variables for the datasets is done by using the `select()` function from `tidyverse` package, with input being the names of the variables to be selected. In addition, since the variable names from the original CES dataset can be confusing, we changed the names of p3, q2, q3, q4, p50, q62 into *voted*, *birthyear*, *gender*, *province*, *marital_status*, and *religious*, respectively, to better represent the aspects being measured by these variables. The renaming of variables is done using the `rename(newname = original name)` command in `r`. Since CES did not explicitly collect the ages of respondents, an *age* variable is added for CES with values being the year in which CES is conducted (2019) minus the reported birth year. The *age* variable is created by the `mutate()` function with input being the new variable name and corresponding values. Creating a variable for age is an intermediate step to the creation of the *age_group* variable, which contains 5 categories: aged between 18 and 24, 25 and 34, 35 and 44, 45 and 54, and finally aged 55 or older. Thus, for a respondent aged, for example, 27, the corresponding age group would be 25 to 34. The creation of the age group variable is done using `mutate()` function from `tidyverse` package, and the categories are created using the `case_when()` function with input being the condition for each category and the name for the corresponding category. The next step of data cleaning process is to omit invalid observations. In particular, observations with missing or NA values or having "don't know" recorded in any of the variables are considered invalid since such observations cannot provide information to the regression model. For CES, the respondents' answers to questions are coded in a consistent manner such that the value -7 represents a skipped question, -8 represents a refusal to answer a question and -9 represents a "don't know" (Stephenson et al.). Thus, observations with values -7, -8, -9 and/or NA for the variables for age group, gender, province, marital status, religious affiliation and the party voted are removed. In addition, observations with gender recorded as 3 or namely, "other" (Stephenson et al.), are also omitted to match the categories in GSS; thus the gender variable is now binary, being male and female. Observations with a recorded value of 8 for party voted, which represents that the respondent cast an invalid ballot (Stephenson et al.) in the election, are also removed. The removal of observations is done using the `filter(!(condition) & !is.na(variable name))` command where `filter` selects for observations satisfying the input condition. The "condition" after `!` is that a specific variable having values -7, -8, -9, as well as 3 for gender and 8 for party voted and `!(condition)` are observations not having these invalid values; `is.na()` function detects missing values in the variables. Then, the *religion* variable is modified into a binary one instead of having over 20 categories. Specifically, if a respondent report having no religion which is coded 21, then this observation is assigned to the category "no religious affiliation", and "has religious affiliation" otherwise. To do so, a new binary variable called *religious_affiliation* is created using the `mutate()` function; its categories are created via `ifelse()` function such that if `religious == 21` evaluates to TRUE then religious affiliation takes "no religious affiliation", and "has religious affiliation" otherwise.

The last step of cleaning CES data recodes the values of the variables. When CES data were collected, the responses were coded into a sequence of integers where each integer represents a specific response. For example, for the provinces of residence, each number is assigned to a province so that Ontario is coded as 6 and Quebec coded as 5. However, for the ordinary audience of the data, this coding can be confusing thus we recoded the values into English words to represent the exact response. After recoding, the response from someone living in Ontario would appear as "Ontario" instead of the number 5. For the columns recording information on gender, marital status, the party voted and the province of residence, the recoding is done through the `recode()` function with input being the corresponding variable name and the recoded values in order (so that values coded as 1 would appear as the first value after the variable name in the input). Gender is recoded as "Male" and "Female"; marital status recoded as having categories "Married", "Living with partner", "Divorced", "Separated", "Widowed" and "Never Married"; the party voted is recoded by the party names: Liberal, Conservative, NDP, Bloc Québécois, Green Party, People's Party and Other; and finally province of residence is recoded into actual names of the provinces in the order: Newfoundland and Labrador, Prince Edward Island, Nova Scotia, New Brunswick, Quebec, Ontario, Manitoba, Saskatchewan, Alberta, British Columbia, NWT, Yukon, and Nunavut. It is worth noting that although NWT, Yukon and Nunavut are coded, none of the respondents reported them as province or territory of residence. Also, the variable for religion is already recoded into the binary *religious_affiliation* in previous steps. Finally, variables are once again selected so that intermediate variables such as `birthyear` are removed from the dataset; this step is done through the `select()` function. The cleaned CES dataset contains 2377 observations on age group, gender, marital status, religious affiliation, province of residence and party voted.

GSS data also needs to be cleaned. Recall at the very beginning of the cleaning process, variables recording age, sex, marital status, whether the respondent has a religious affiliation, citizenship status and province of residence are selected from the full GSS dataset via the `select()` function. These variables, except for citizenship status, have a one-to-one correspondence with the six variables in the cleaned CES dataset and would go through similar cleaning processes as described above. Note that sex, although not necessarily the same as gender, is treated as if it was consistent with gender to match the *gender* variable in CES. Thus, sex is renamed into gender using the `rename()` function. Similarly, the GSS variable *religion_has_affiliation* is renamed into *religious_affiliation* to match the variable name in CES. An age group variable with the same partition as that in CES is also created for GSS. The GSS data is also filtered for observations with age over 18 and citizenship status is either by birth or by naturalization. The selection for respondents over 18 years of age and who are Canadian citizens is necessary since people who don't meet these two criteria are not eligible to vote, which is our outcome variable of primary concern. The creation of age group variable follows the exact same procedure as that for CES; and the selection of observations is done via `filter(age >= 18)` and `filter(citizenship_status %in% c("By birth", "By naturalization"))` command. In addition, since the names of the categories of marital status are not exactly the same as that in CES, we mutated the naming of categories of GSS marital status to ensure it is consistent with CES. Namely, the categories "Living common-law" and "Single, never married" are modified into "Living with partner" and "Never married". Modification of categories is done by using `mutate()` and `replace()` twice, where `replace(marital_status, marital_status == category name, new category name)` is the input to `mutate()` function. Similarly, the categories of religious affiliation in the two datasets do not match exactly since the CES religious affiliation variable is binary but the GSS religious affiliation has an additional "Don't know" category. For post-stratification purposes, this third category is omitted so that the variables in CES and GSS correspond with each other. This step is done by filtering out observations with reported religious affiliation as "don't know" using the `filter()` function. Finally, variables are once again selected so that the final dataset contains age group, gender, marital status, religious affiliation and the province of residence and all missing values are omitted using the `is.na()` function. The cleaned GSS dataset contains 18770 observations on the 5 variables.

After the cleaning processes, there are a total of six variables of interest, out of which five variables are in both datasets and act as predictors or independent variables, and the other is the response or dependent variable. Specifically, the predictors are age group, gender, marital status, religious affiliation and province of residence and are presented in both GSS and CES data; and the response variable represents the party voted is present only in CES data. **Age group** (*age_group*) is a categorical variable with 5 categories representing the age range (at the survey year) that respondents belong to; **gender** (*gender*) is a binary variable indicating the gender as male or female of the respondents; **marital status** (*marital_status*) is a categorical variable indicating whether a respondent is legally married, living with a partner, divorced, separated, widowed or has never married; **religious affiliation** (*religious_affiliation*) is binary indicating whether a respondent has religious affiliation; **province of residence**(*province*) is categorical, representing the province that a respondent resides in at the time of the survey; and finally, **party voted** (*voted*) is a categorical variable indicating which party did a respondent vote for. The party voted by respondents is selected as the response variable to fit the model on CES data because we are interested in the voting outcome in the upcoming election. It is expected that the voting outcome in the last election may convey some information about the outcome in the next election.

In the following paragraphs, we will present some summaries on both CES and GSS data to provide insight into the structure of the data and prepare the reader for the actual analysis.

Table 1: 2019 voting proportion across parties

Party Voted	Number	Proportion
Liberal	758	0.3188894
Conservative	750	0.3155238
NDP	435	0.1830038
Green Party	247	0.1039125
Bloc Québécois	138	0.0580564
People's Party	38	0.0159865
Other	11	0.0046277

The table above (Table 1) shows the proportion of CES respondents who voted for each political party. Not surprisingly, the Liberal party and the Conservative party of Canada are the two most popular parties, each occupying about 31% of the total vote with the Liberal party winning by a 0.3% margin. The New Democratic Party and the Green Party occupy 18.3% and 10.4% of the vote, respectively. Bloc Québécois and People's Party are the least popular parties, receiving only 5.8% and 1.6% of the total vote, respectively. The remaining 0.4% of the survey sample did not vote for any of the parties mentioned above, and their vote is recorded as "Other".

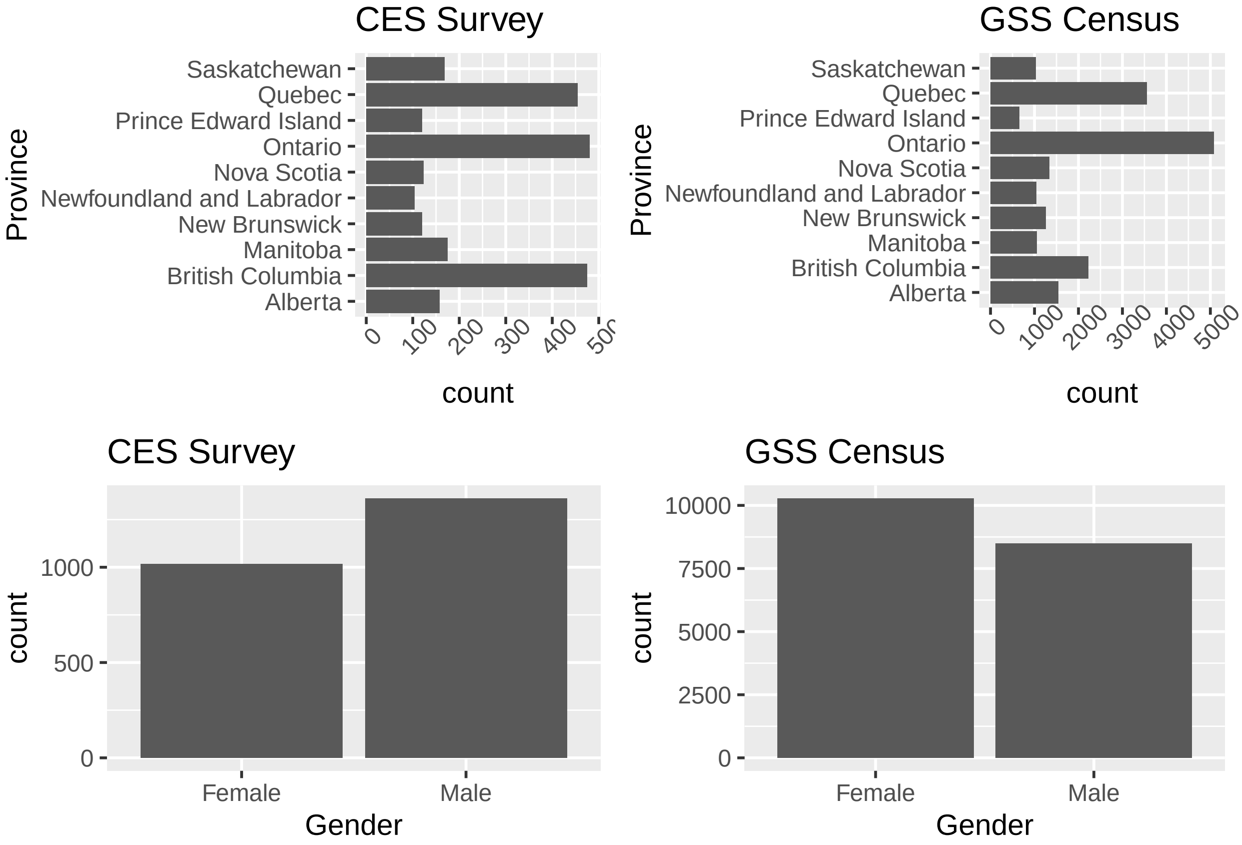

Figure 1: Comparison between gender and province distribution across CES2019 and GSS2017 data

Figure 1 demonstrates a graphical summary of both CES and GSS data on the variables **age** and **province** of residence. The bar plots in the first row show the distribution over the province of residence of the CES or GSS sample. By comparing and contrasting the two bar plots, we found that residents in Quebec, Ontario and BC are relatively better represented in both data sets. Yet CES includes a significantly larger proportion of BC and Quebec residents than included in GSS. Fewer residents from other provinces are included in both samples. The bar plots in the second row show distribution of sample observations over gender. As shown in the plot, male respondents occupy a slightly larger proportion of the sample than female respondents in the CES sample, but an opposite pattern is observed in the GSS sample where female respondents occupy a slightly larger proportion than male respondents. Despite the discrepancy, the proportions of male and female respondents in both samples do not differ significantly and both genders are not over or underrepresented.

Bar plots showing the distribution of CES and GSS samples over age group, marital status, and religious affiliation are also constructed (see Appendix). Based on the bar plots, we found that distributions over these three predictors are very similar for CES and GSS samples.

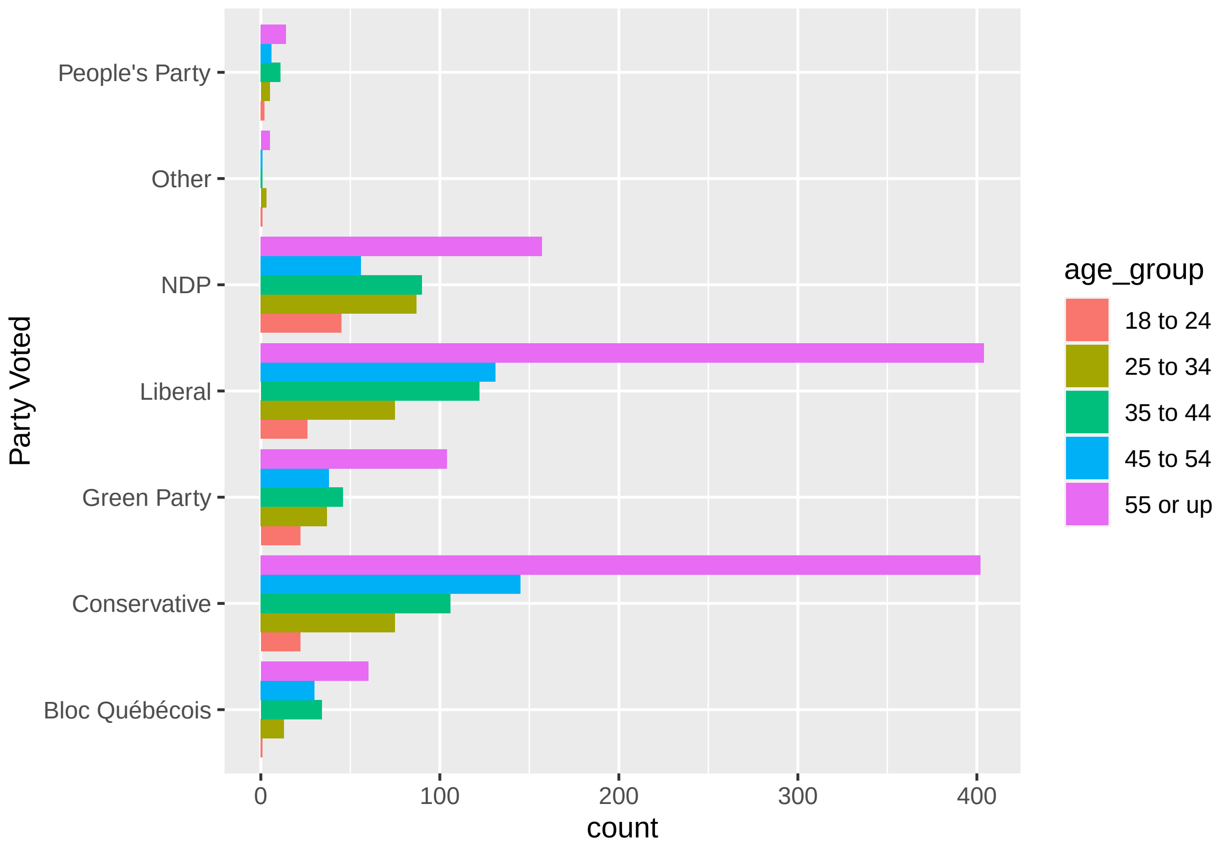

Figure 2: Age distribution of voters for each party

Figure 2 shows the age distribution of voters for each party in the CES sample, demonstrating the number of respondents in each age group that voted for each party. Across all six parties and an "Other" category, we noticed that votes from respondents aged 55 and older occupy a large proportion of total votes received for each party, which makes sense because the category "55 or up" covers a wider range of age than other age groups. The opposite pattern occurs with respondents aged 18 to 24 that their votes occupy a relatively small proportion of total votes received for each party. The votes received by the Liberal Party and the Conservative Party have similar constituent and the votes received by the NDP, Green Party and Bloc Québécois have similar constituent. The Liberal Party and the Conservative Party receive more votes from respondents aged 45 to 54 than from respondents aged 35 to 44; but for the NDP, Green Party and Bloc Québécois, the opposite pattern occurs.

The summaries above should have provided insight into the overall distribution of the CES and GSS sample over the predictors we are considering. Namely, the preditors are age, gender, marital status, religious affiliation and province of residence. The five predictors will be included in the multinomial logistic regression model, which will be introduced in the **Methods** section, aiming at unveiling their association with the election outcome as well as to predict the outcome for the next election.

Methods

This analysis aims to identify factors that significantly affect Canadian voting decisions and to build a predictive model based on these explanatory variables. Since the response variable, namely, Canadian voting decision is a nominal or unordered categorical variable with multiple levels (Liberal Party, Conservative Party, NDP etc.), we find that a multinomial logistic regression is an appropriate analytic approach to predict the proportion of the Canadian population voting for a certain party. Multinomial logistic regression is a generalized form of logistic regression - the model for categorical variables with binary responses, which we are more familiar with.

The dependent variable $y =$ `voted` has seven categories corresponding to six parties and an "other" category, as the Data section outlines. Let $\{\pi_1, \pi_2, ..., \pi_i, ...\pi_7\}$ denote the respective probabilities of voting for each party, satisfying $\sum_i \pi_i = 1$. Assuming all observations in the survey data are independent of each other, the probability distribution for the number of outcomes amongst the seven types is multinomial (Agresti, 2019).

In this analysis, we decide to use the last category (voting for Other) as the baseline or reference category. The multinomial logistic model first pairs each category with the chosen baseline category. The model then estimates the **baseline-category logits**, which are **log-odds** of voting for one party against voting for "Others". The odds is the ratio of the probability of choosing one party over the probability of choosing the baseline category. The model is actually estimating, for each possible pair,

$$log(\frac{\pi_i}{\pi_7}), i \in \{1, 2, 3, ..., 6\}$$

. Six pairs of logit correspond to six equations, with separate parameters for each. The effects vary according to the category paired with the baseline (Agresti, 2019). Using the `multinom()` function in the R package `nnet`, we are able to fit all six equations simultaneously (*Multinom: Fit multinomial log-linear models*).

The estimated probability of voting for each non-baseline party is

$$\hat{\pi_i}(X) = \frac{exp(\alpha_i + \beta_i X)}{1 + \sum_{j=1}^{6} exp(\alpha_j + \beta_j X)}, j \in \{1, 2, 3, ..., 6\}$$

, whereas the predicted probability for the baseline category (voting for Others) is

$$\hat{\pi_7}(X) = \frac{1}{1 + \sum_{j=1}^{6} exp(\alpha_j + \beta_j X)}$$

(Agresti, 2019). The denominator is the sum of the predicted probabilities for all categories, where the probability of choosing the baseline equals to 1 since all parameters are zero for the baseline category. The numerators of all estimated probabilities would add up to be equal to the denominator (Agresti, 2019).

To optimize the model performance as well as interpretability of the results, we performed model selection techniques to reduce the model to include only the most significant independent variables. To be more specific, we followed a variable selection procedure in a backward stepwise manner. Proceeding from the initial model containing all predictors of interest, variables are dropped one by one based on a variable selection criterion. The criteria for variable selection used is the Bayesian Information Criterion (BIC) since it is computationally convenient and is efficient in handling large-size datasets. BIC imposes a penalty term on the number of parameters: $qlog(n)$, where q is the number of model parameters and n is the size of the cleaned CES data. To summarize, a backward stepwise selection is applied to our initially proposed regression model. That is, deleting variables from the model as long as the dropping of the variable(s) results in a lower BIC value.

Model Specifics

To model the probability of voting for the Liberal Party, we propose an initial multinomial logistic regression model of the theoretical form:

$$log(\frac{\pi_i}{\pi_7}) = \alpha_i + \beta_i X + \epsilon_i, i \in \{1, 2, 3, ..., 6\}$$

where the intercept α_i is the average log odds of voting for the i-th party against voting for Other;

$$\beta_i X = \beta_{1i} age + \beta_{2i} gender + \beta_{3i} province + \beta_{4i} religion\_has\_affiliation + \beta_{5i} marital\_status$$

is the matrix form that encapsulates the slope estimates of model specified independent variables for the i-th party.

The multinomial logistic model specified above is built on the following assumptions (*Multinomial logistic regression using R* 2018):

- The response variable is nominal or unordered, which is clearly satisfied due to the nature of voting for different political parties. The voting response has multiple categories yet does not have an inherent order.

- All observations are independent, which is a reasonable assumption since the CES survey data is collected at the individual level and we can expect that people's voting response and demographic characteristics are fairly independent of one another.

- There is no perfect collinearity between predictors in the model such that the predictors are not significantly correlated with each other. This assumption is satisfied as the chosen categorical predictors clearly do not exhibit such a strong correlation.

- There are no influential points or outliers that would greatly affect the model. This should be true after the previous data cleaning process.

Post-Stratification

Although the CES survey sample is obtained by reaching out to randomly selected households, there are observable distributional differences in some demographic variables compared with the GSS data (as shown in Figure 1 of the Data section). For the validity of the analysis, we would like to eliminate the differences via adjusting the proportional weights for our estimates using the post-stratification technique. The success of this technique has been proven by previous studies where researchers had used non-representative poll data from the Xbox gaming platform to predict the 2012 US presidential election using the post-stratification technique (Wang, et al., 2014).

After finalizing the model fitted on the survey data, we first partition the census data into cells by demographic predictor variables as specified in the final model. Namely, age group (18 to 24; 25 to 34; …;, 55 and up), gender (Female, Male), province (Alberta, British Columbia, etc.) and whether a respondent is affiliated to a religion (Has/No). 200 bins are generated from such partitioning of the GSS data. Then we apply the fitted model to the partitioned GSS data and yield estimates of the voting probability within each cell. Lastly, we weigh the cell-level estimates by their respective proportion in the population and aggregate them to the population level. Thus, the final results will be a table outlining the estimated proportions of the population that will vote for each of the 6 parties and "other".

Using π_i to indicate the probability of voting for a certain Canadian party (which is the outcome of interest), the post-stratification estimate of π_i is

$$\hat{\pi_i}^{PS} = \frac{\sum_{j=1}^{J} N_j \hat{\pi_{ij}}}{\sum_{j=1}^{J} N_j}$$

where $\hat{\pi_{ij}}$ is the estimate of π_i in each cell j and N_j is the size of the j-th cell based off partitioning variables.

Through the application of the post-stratification technique, the difference in group proportions should be eliminated by adjusting the weight of the survey prediction in alignment with the census data. Doing so enables our prediction on the voting outcome to better generalize to the Canadian population.

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

Results

After applying model selection techniques based on BIC, the predictor marital status is dropped from the model. The reduced model uses age group, gender, province and religion affiliation to estimate the

category-baseline logits: $log(\frac{\pi_i}{\pi_7}), i \in \{1, 2, 3, ..., 6\}$.

Recall that the multinomial regression model is estimating **log odds** of voting for one party against voting for "Others". The odds is the ratio of the probability of voting one party over the probability of voting for the baseline "Other". An increase in the log odds implies an increase in the voting probability for that party. For simplicity, the effects of all coefficient estimates are to be interpreted in terms of log odds. Hopefully, the explanation of the log odds concept presented in the report is enough to prepare you to read the following results.

Table 3 (see Appendix) outlines the logistic coefficient estimates of each predictor for each non-baseline party. The logistic coefficient is the expected amount of change in the log odds for each one-unit change in the predictor (Starkweather & Moske). Since our model only includes categorical variables, the coefficients are estimated amount change in the log odds on average, if an individual respondent falls under a category.

There are a few coefficients which we would like to highlight from the lengthy table in Appendix. Amongst all non-baseline parties, Bloc Québécois is estimated to have the lowest intercept, which is in line with its low voted proportion in the Data section. The intercept is the log odds estimate for Bloc Québécois relative to "Other" when all predictor variables in the model are evaluated at zero (*Multinomial Logistic Regression | SPSS Annotated Output*). For a female respondent in the 18-24 years old age group who has religious affiliation and is living in Alberta at the time of the survey, the log odds for preferring Bloc Québécois to "Other" is -13.32. However, living in Quebec drastically increases the log odds of voting for Bloc Québécois by roughly 14.93 units relative to someone living in Alberta given the other variables are held at a fixed level. There is also evidence that people over 35 are more likely to vote for this party. Holding all other factors constant, someone in the 45 to 54 age group is estimated to increment his/her log odds of voting for Bloc Québécois by 3.47 units relative to someone in the 18 to 24 age group. On the other hand, someone living in British Columbia is predicted to be a lot less likely to vote for Bloc Québécois. Current residence in British Columbia is associated with a decrease in log odds of around 8.12 units relative to residing in Alberta, holding all other variables at a fixed level.

The Conservative Party has the highest intercept estimate of 3.43 amongst the six parties. The log odds for preferring the Conservative Party to "Other" is estimated to be 3.43 units for a female in the 18 to 24 years old age group, affiliated to a religion and is living in Alberta at the time of the survey. It seems that people living in Manitoba and Saskatchewan are highly in favour of the party. Living in these two provinces increases the log odds of voting for the Conservative party by around 9.03 and 9.00 units relative to living in Alberta, respectively when fixing all other variables. Elderly people also show a preference for the Conservative party. The increment in log odds of voting for the Conservative Party, relative to the 18 to 24 age group, is observably higher for CES respondents in the 45 to 54 age group, holding all other variables in the model constant.

The respondents' province of residence at the time of the CES survey affects the voting preference for other parties in a similar manner. People living in Prince Edward Island and Manitoba are a lot more likely to vote for the Green Party, relative to a resident in Alberta, shown by the respective increment in log odds of 10.84 and 10.21 units given the other variables in the model are held constant. These two provinces also show a strong favour in the Liberal Party and the NDP. On the contrary, residence in Manitoba and Prince Edward Island is associated with a considerably (13.01 units and 14.05 units) lower log odds of voting for the People's Party relative to someone living in Alberta, holding all other variables constant.

Gender and religious affiliation demonstrate a relatively moderate effect on the voting outcome. In particular, being a male is associated with a 0.4 to 0.6 decrease in the average log odds of voting for the Liberal Party, the Green Party, the NDP and Bloc Québécois relative to a female respondent; but is also associated with a 0.16 and 0.09 units increment in the log odds of voting for the Conservative Party and the People's Party, respectively on average. Relative to respondents who are affiliated with a religion, the log odds of voting for all non-baseline parties over "Other" are higher. Specifically, the log odds of voting for the Green Party against voting for "Other" is 1.40 units higher relative to a respondent who has a religious affiliation, holding all other factors constant.

Table 2: Estimated proportion

BQ_predict	Con_predict	Green_predict	Lib_predict	NDP_predict	PPL_predict	Other_predict
0.0577594	0.3179733	0.091992	0.3466389	0.1655702	0.0154721	0.0045834

By fitting the multinomial logistic regression model onto the partitioned GSS data, we were able to produce an estimated proportion for each cell; the estimates are then weighted by the cell size and aggregated to the population level where we arrived at the predicted proportion of the population that will vote for each political party. The results are demonstrated in Table 2. Based on Table 2, about 34.66% of the electorates will support the Liberal Party and about 31.80% support the Conservative Party. This result matches our expectation that the winner of the next election will be one of the two parties and is also consistent with the results in the 2019 federal election where approximately 31% voted for the Liberal Party and 31% voted for the Conservative Party. The NDP and the Green Party are estimated to receive support from an estimated 16.56% and 9.20% of the population, respectively. The Bloc Québécois, the People's Party and other political parties are supported by less than 6% of the population. The estimated results for parties other than the Liberal and the Conservative are also consistent with the outcome of the previous elections where NDP has been the third popular political party and the Green Party, the Bloc Québécois and other parties receive relatively less support.

Conclusions

Since the 2015 federal election, Justin Trudeau and his Liberal Party of Canada have served for two terms of office and are starting with his third term as a result of the 2021 election. Despite his winning, Trudeau leads a minority government and the competition between the Liberal Party and the Conservative Party is rather intense. While the Liberal Party has won three elections since 2015, the Conservative Party has always been almost as popular among Canadian citizens as the Liberal Party. Hence, it is natural that one would doubt whether Trudeau and his party will win the next election, or will there be a change in the political vibe that the Conservative Party, or even other political parties of Canada, will lead the federal government after the 2025 election. Based on previous election outcomes, we hypothesized that the winner of the next election will likely still be either the Liberal Party or the Conservative Party; other political parties such as the NDP, the Green Party and the Bloc Québécois are much less likely to win the election. Aiming at providing an estimated proportion of Canadian citizens that will vote for each political party in the next election, we analyzed the 2019 CES data using a multinomial logistic regression model. The fitted model is then applied to the 2017 GSS data, which included a larger and more comprehensive sample. Proportion estimates are computed using post-stratification techniques to arrive at the final results.

Amongst the predictors in the final model, the province of residence has the largest impact on the voting response. Holding all other variables in the model constant, a Quebec resident is estimated to have roughly 14.93 units higher log odds of voting for Bloc Québécois, relative to someone living in Alberta. By contrast, living in British Columbia is associated with a decrease in log odds of voting for Bloc Québécois of around 7.49 units relative to living in Alberta, holding all other variables at a fixed level. Manitoba and Saskatchewan show a strong favour of the Conservative Party. Living in these two provinces increases the log odds of voting for the Conservative Party by around 9.04 and 9.01 units, respectively, relative to living in Alberta when fixing all other variables. People currently living in Prince Edward Island and Manitoba are a lot more likely to vote for the Green Party, relative to a resident in Alberta, shown by the respective increment in log odds of 10.84 and 10.21 units given the other variables in the model are held constant. These two provinces also show a strong favour in the Liberal Party and the NDP. On the contrary, residing in Manitoba and Prince Edward Island is associated with a considerably (13.01 and 14.05 units) lower log odds of voting for the People's Party relative to someone living in Alberta, holding all other variables constant.

The elderly demonstrate higher participation in the election voting and also show a preference for certain parties. People over 35 show a preference for Bloc Québécois. Holding all other factors constant, someone in the 45 to 54 age group is estimated to increment his/her log odds of voting for Bloc Québécois by 3.47 units relative to someone in the 18 to 24 age group. The voting outcome for the Conservative Party shows a similar pattern. The increment in log odds of voting for the Conservative Party, relative to the 18 to 24 age group, is observably higher for CES respondents in the 45 to 54 age group, holding all other variables in the model constant.

The adjusted prediction of the voting proportion across parties is in line with our previous hypothesis that the winning party would be elected between Liberal or Conservative. The Liberal and Conservative Party take a large share of votes, 34.66% and 31.80% respectively. Yet the Liberal party wins by a small margin of approximately 2.87 percentage points and is likely to lead by a minority government.

Due to weaknesses inherent in the 2019CES and 2017GSS data, our analysis is subjected to some biases and limitations. First of all, there is a two-year discrepancy between the survey years of CES and GSS, and both survey years (2017 or 2019) are over 5 years from the next election (2025). It is possible that certain events happened or could happen between 2017 or 2019 and 2025 may impact the election outcome. Such events may include the COVID-19 pandemic and the economic turbulence followed from it, shift in political party leaders, and any possible changes in the political atmosphere and/or world relations. Specifically, the Conservative Party had its new Leader Erin O'Toole in 2020 (*2020 Conservative Party of Canada leadership election* 2021) and CES data fails to provide information on electorates' attitude toward him and the Conservative Party under his leadership. Thus, the results of this analysis are a rather simplistic and naive prediction of the 2025 election. Also, the GSS data does not include variables on respondents' political attitudes and voting preferences thus our prediction is based solely on demographic variables. To arrive at more meaningful and reliable results, future research may apply post-stratification to census data that collects responses on political attitude-related variables. In addition, electorates from the three territories (i.e. NWT, Yukon and Nunavut) are underrepresented in both GSS and CES data and Canadian citizens who have the right to vote but are in a foreign country at the time of the survey are also excluded from the data, thus future studies may look at data including observations from all 13 Canadian provinces/territories and considering Canadian citizens outside Canada so that the entire population is represented in the sample. Furthermore, since CES collected only respondents' gender but not sex and GSS collected only sex but not gender, gender and sex are treated as the same variable despite conceptual difference; also, to ensure consistency in the categories of the variable, only the categories "male" and "female" are included in the model since they appear in both the gender variable from CES and the sex variable from GSS. Treating gender and sex as the same binary variable not only neglects the fact that it is not uncommon that people's gender may not align with their sex at birth but also fails to account for the fluidity in gender or sex. Transgender and gender-diverse population are thus not well represented in the sample. Future analysis should consider gender and sex separately and as having multiple categories in addition to male and female to capture their influence, if any, on voting outcomes. Lastly, subject to the nature of the telephone survey, a large portion of the respondents reported "don't know" or invalid answers to at least one of the variables of interest; many also skipped or refused to answer certain questions. These missing or invalid values are removed from the dataset during the cleaning process since such values would impede analysis. However, treating missing values in this way is a huge waste of data where over 2000 observations are dropped from both datasets. Hence, we may look for other ways of handling missing values that would reduce the waste of data if this analysis is to be done again.

The model built in our analysis is far from perfect. The presented model only includes categorical predictors, which is not an ideal choice. The multinomial logistic regression model is not one that provides intuitive results and has drawbacks in its interpretability. Aside from the predictors included in our multinomial logistic regression model, there are other variables included in the GSS data that may be associated with one's voting preference and political attitude. An example of such a variable is the citizenship status column from the GSS data, which consists of categories including "by birth" and "by naturalization". We are aware that citizenship status may have an impact on one's attitude towards, for example, immigration policies, yet the variable is not included in our model because CES did not collect data on citizenship status. As mentioned previously, variables that explicitly indicate one's voting preferences are included in the CES dataset but not in GSS, thus we also did not include these variables in the model. Hence, future studies on the topic can look for survey and census data with more common variables to build a more comprehensive model and perform a more insightful post-stratification analysis. Considering that CES and GSS data are collected across multiple time periods, we may build a Bayesian model rather than a frequentist one in future analyses so that to learn patterns from past data.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)

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

3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)

4. Government of Canada. (2021, October 28). Consolidated federal laws of Canada, Canada elections act. Canada Elections Act. Retrieved November 5, 2021, from https://laws-lois.justice.gc.ca/eng/acts/e-2.01/page-10.html#h-204214.

5. About Federal Elections. Elections Canada. (2021, October 31). Retrieved November 5, 2021, from https://www.elections.ca/content2.aspx?section=faq&document=fedelect&lang=e.

6. Azzi, S., & Kwavnick. (2012, January 17). Minority governments in Canada. The Canadian Encyclopedia. Retrieved November 5, 2021, from https://www.thecanadianencyclopedia.ca/en/article/minority-government.

7. Mason, G., & Urback, R. (2021). Canada's 2021 federal election: Live results. The Globe and Mail. Retrieved November 5, 2021, from https://www.theglobeandmail.com/politics/federal-election/2021-results/.

8. Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, "2019 Canadian Election Study - Phone Survey", https://doi.org/10.7910/DVN/8RHLG1, Harvard Dataverse, V1, UNF:6:eyR28qaoYlHj9qwPWZmmVQ==fileUNF

9. Statistics Canada. (2020, April 20). Cycle 31 : Families. Public Use Microdata File Documentation and User's Guide. Microdata Analysis and Subsetting with SDA. Retrieved November 5, 2021.

10. Agresti, A. (2019). An introduction to categorical data analysis. John Wiley & Sons.

11. Multinomial logistic regression using R. Data Science Beginners. (2018, December 20). Retrieved November 5, 2021, from https://datasciencebeginners.com/2018/12/20/multinomial-logistic-regression-using-r/#:~:text=Multinomial%20Logistic%20Regression%20Using%20R%20Multinomial%20regression%20is,be%20predicted%20using%20one%20or%20more%20independent%20variable.

12. Wang, W., et al., Forecasting election with non-representative polls. International Journal of Forecasting (2014), http://dx.doi.org/10.1016/j.ijforecast.2014.06.001

13. Starkweather, J., & Moske, A. K. (n.d.). Multinomial Logistic Regression . Retrieved November 5, 2021, from https://it.unt.edu/sites/default/files/mlr_jds_aug2011.pdf#.

14. UCLA: Statistical Consulting Group. (n.d.). Multinomial Logistic Regression | SPSS Annotated Output. IDRE Stats. Retrieved November 5, 2021, from https://stats.idre.ucla.edu/spss/output/multinomial-logistic-regression/.

15. Wikimedia Foundation. (2021, September 22). 2020 Conservative Party of Canada leadership election. Wikipedia. Retrieved November 5, 2021, from https://en.wikipedia.org/wiki/2020_Conservative_Party_of_Canada_leadership_election.

Appendix

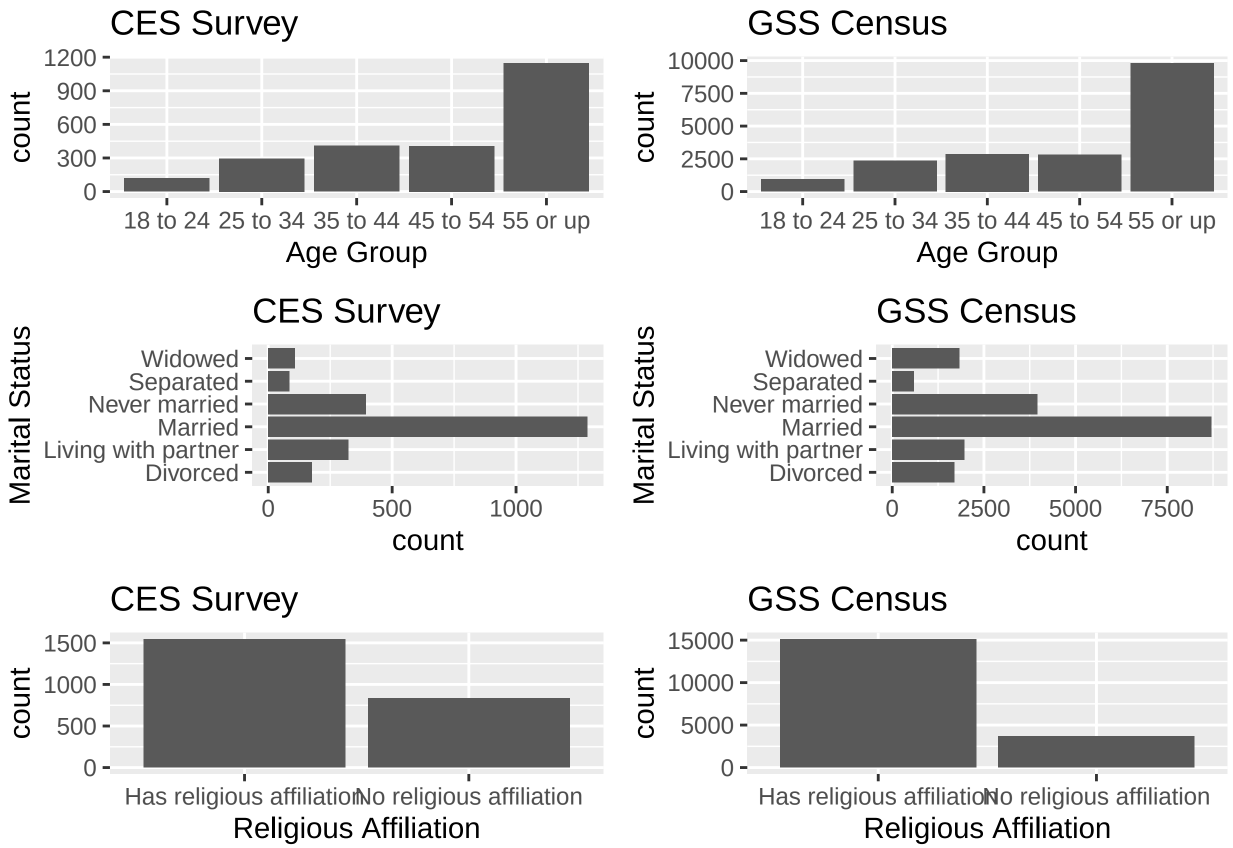

Figure 3: Distribution over age group, marital status, and religious affiliation

Table 3: Coefficient estimates for the final multinomial logistic regression

y.level	term	estimate	std.error	statistic	p.value
Bloc Québécois	(Intercept)	-13.3218708	1.3168054	-1.011681e+01	0.0000000
Bloc Québécois	as.factor(age_group)25 to 34	1.0631012	1.5831929	6.714919e-01	0.5019072
Bloc Québécois	as.factor(age_group)35 to 44	3.2192613	1.7712292	1.817529e+00	0.0691361
Bloc Québécois	as.factor(age_group)45 to 54	3.4705707	1.7691889	1.961673e+00	0.0498005
Bloc Québécois	as.factor(age_group)55 or up	3.0698307	1.5225975	2.016180e+00	0.0437811
Bloc Québécois	as.factor(gender)Male	-0.4081308	0.6641546	-6.145117e-01	0.5388772
Bloc Québécois	as.factor(religious_affiliation)No religious affiliation	0.8826715	0.7295991	1.209804e+00	0.2263543
Bloc Québécois	as.factor(province)British Columbia	-7.4942091	0.0031440	-2.383688e+03	0.0000000
Bloc Québécois	as.factor(province)Manitoba	3.9978035	0.0038123	1.048669e+03	0.0000000

y.level	term	estimate	std.error	statistic	p.value
Bloc Québécois	as.factor(province)New Brunswick	2.8533132	0.1143866	2.494446e+01	0.0000000
Bloc Québécois	as.factor(province)Newfoundland and Labrador	4.3538574	0.7172426	6.070272e+00	0.0000000
Bloc Québécois	as.factor(province)Nova Scotia	2.9509664	0.1481591	1.991756e+01	0.0000000
Bloc Québécois	as.factor(province)Ontario	-6.7134889	0.0033715	-1.991258e+03	0.0000000
Bloc Québécois	as.factor(province)Prince Edward Island	4.5154265	0.9472724	4.766766e+00	0.0000019
Bloc Québécois	as.factor(province)Quebec	14.9284941	1.3174145	1.133166e+01	0.0000000
Bloc Québécois	as.factor(province)Saskatchewan	2.4108619	0.1337675	1.802278e+01	0.0000000
Conservative	(Intercept)	3.4349408	1.4571378	2.357321e+00	0.0184074
Conservative	as.factor(age_group)25 to 34	0.1807872	1.1946322	1.513329e-01	0.8797131
Conservative	as.factor(age_group)35 to 44	1.6584986	1.4505393	1.143367e+00	0.2528863
Conservative	as.factor(age_group)45 to 54	2.0141518	1.4421855	1.396597e+00	0.1625348
Conservative	as.factor(age_group)55 or up	1.4283092	1.1323697	1.261345e+00	0.2071844
Conservative	as.factor(gender)Male	0.1602878	0.6367473	2.517290e-01	0.8012505
Conservative	as.factor(religious_affiliation)No religious affiliation	0.0366430	0.7002660	5.232720e-02	0.9582680
Conservative	as.factor(province)British Columbia	-1.4873898	1.1157417	-1.333095e+00	0.1825006
Conservative	as.factor(province)Manitoba	9.0393904	0.2173830	4.158279e+01	0.0000000
Conservative	as.factor(province)New Brunswick	7.0981432	57.0704955	1.243750e-01	0.9010184
Conservative	as.factor(province)Newfoundland and Labrador	7.5005210	96.2363546	7.793850e-02	0.9378769
Conservative	as.factor(province)Nova Scotia	6.9661092	65.7740704	1.059097e-01	0.9156540
Conservative	as.factor(province)Ontario	-0.9894261	1.1674016	-8.475456e-01	0.3966911
Conservative	as.factor(province)Prince Edward Island	7.4862539	91.9421435	8.142350e-02	0.9351051
Conservative	as.factor(province)Quebec	-1.2066964	1.2401344	-9.730367e-01	0.3305350
Conservative	as.factor(province)Saskatchewan	9.0090336	96.9071527	9.296560e-02	0.9259309
Green Party	(Intercept)	1.2824203	1.5246950	8.410995e-01	0.4002922
Green Party	as.factor(age_group)25 to 34	-0.6132232	1.2030870	-5.097081e-01	0.6102560
Green Party	as.factor(age_group)35 to 44	0.8459440	1.4562052	5.809236e-01	0.5612920
Green Party	as.factor(age_group)45 to 54	0.6921395	1.4506953	4.771088e-01	0.6332847
Green Party	as.factor(age_group)55 or up	0.1917843	1.1377334	1.685670e-01	0.8661372

y.level	term	estimate	std.error	statistic	p.value
Green Party	as.factor(gender)Male	-0.4987569	0.6452575	-7.729580e-01	0.4395473
Green Party	as.factor(religious_affiliation)No religious affiliation	1.4002168	0.7073420	1.979547e+00	0.0477544
Green Party	as.factor(province)British Columbia	0.7915982	1.2021372	6.584924e-01	0.5102218
Green Party	as.factor(province)Manitoba	10.2110056	0.4237782	2.409516e+01	0.0000000
Green Party	as.factor(province)New Brunswick	9.7057291	57.0723579	1.700601e-01	0.8649629
Green Party	as.factor(province)Newfoundland and Labrador	9.3340218	96.2381896	9.698880e-02	0.9227353
Green Party	as.factor(province)Nova Scotia	9.6475422	65.7757221	1.466733e-01	0.8833899
Green Party	as.factor(province)Ontario	1.1893391	1.2522550	9.497579e-01	0.3422353
Green Party	as.factor(province)Prince Edward Island	10.8368827	91.9431705	1.178650e-01	0.9061746
Green Party	as.factor(province)Quebec	1.0258988	1.3232134	7.753087e-01	0.4381573
Green Party	as.factor(province)Saskatchewan	8.7834826	96.9093885	9.063600e-02	0.9277818
Liberal	(Intercept)	1.9467121	1.4726466	1.321914e+00	0.1861968
Liberal	as.factor(age_group)25 to 34	-0.0692362	1.1917000	-5.809870e-02	0.9536700
Liberal	as.factor(age_group)35 to 44	1.5909561	1.4475113	1.099098e+00	0.2717255
Liberal	as.factor(age_group)45 to 54	1.6848507	1.4399803	1.170051e+00	0.2419804
Liberal	as.factor(age_group)55 or up	1.3040429	1.1295041	1.154527e+00	0.2482842
Liberal	as.factor(gender)Male	-0.4632122	0.6360858	-7.282228e-01	0.4664772
Liberal	as.factor(religious_affiliation)No religious affiliation	0.6586006	0.6992891	9.418144e-01	0.3462877
Liberal	as.factor(province)British Columbia	0.0739748	1.1387015	6.496410e-02	0.9482025
Liberal	as.factor(province)Manitoba	10.3174387	0.2729913	3.779402e+01	0.0000000
Liberal	as.factor(province)New Brunswick	8.9422790	57.0709533	1.566870e-01	0.8754915
Liberal	as.factor(province)Newfoundland and Labrador	10.0348884	96.2365223	1.042732e-01	0.9169526
Liberal	as.factor(province)Nova Scotia	9.4097460	65.7743375	1.430611e-01	0.8862420
Liberal	as.factor(province)Ontario	1.3428855	1.1874717	1.130878e+00	0.2581065
Liberal	as.factor(province)Prince Edward Island	9.9993120	91.9423177	1.087564e-01	0.9133957
Liberal	as.factor(province)Quebec	1.5154590	1.2568772	1.205734e+00	0.2279202
Liberal	as.factor(province)Saskatchewan	9.5297721	96.9075505	9.833880e-02	0.9216633
NDP	(Intercept)	3.1625260	1.4606849	2.165098e+00	0.0303802
NDP	as.factor(age_group)25 to 34	-0.4495909	1.1846886	-3.795013e-01	0.7043157
NDP	as.factor(age_group)35 to 44	0.7364337	1.4433088	5.102398e-01	0.6098834

y.level	term	estimate	std.error	statistic	p.value
NDP	as.factor(age_group)45 to 54	0.3644998	1.4383134	2.534217e-01	0.7999424
NDP	as.factor(age_group)55 or up	-0.1231961	1.1242777	-1.095780e-01	0.9127441
NDP	as.factor(gender)Male	-0.6377417	0.6390410	-9.979668e-01	0.3182955
NDP	as.factor(religious_affiliation)No religious affiliation	1.2798814	0.7010866	1.825568e+00	0.0679153
NDP	as.factor(province)British Columbia	-0.1407191	1.1302131	-1.245067e-01	0.9009141
NDP	as.factor(province)Manitoba	9.9702100	0.2592595	3.845648e+01	0.0000000
NDP	as.factor(province)New Brunswick	7.4075247	57.0713082	1.297942e-01	0.8967293
NDP	as.factor(province)Newfoundland and Labrador	9.5946175	96.2364704	9.969840e-02	0.9205838
NDP	as.factor(province)Nova Scotia	8.4294612	65.7743511	1.281573e-01	0.8980245
NDP	as.factor(province)Ontario	-0.0885425	1.1838904	-7.478950e-02	0.9403822
NDP	as.factor(province)Prince Edward Island	8.1250556	91.9426623	8.837090e-02	0.9295819
NDP	as.factor(province)Quebec	0.0496359	1.2540568	3.958030e-02	0.9684278
NDP	as.factor(province)Saskatchewan	9.3650735	96.9074095	9.663940e-02	0.9230128
People's Party	(Intercept)	0.5330952	1.6906086	3.153274e-01	0.7525131
People's Party	as.factor(age_group)25 to 34	-0.1233637	1.4375802	-8.581340e-02	0.9316147
People's Party	as.factor(age_group)35 to 44	1.8292529	1.6246723	1.125921e+00	0.2601989
People's Party	as.factor(age_group)45 to 54	1.2987401	1.6417055	7.910920e-01	0.4288903
People's Party	as.factor(age_group)55 or up	0.5696189	1.3452886	4.234176e-01	0.6719906
People's Party	as.factor(gender)Male	0.0945300	0.7192641	1.314260e-01	0.8954384
People's Party	as.factor(religious_affiliation)No religious affiliation	0.5118057	0.7768256	6.588424e-01	0.5099970
People's Party	as.factor(province)British Columbia	-0.7265126	1.2551459	-5.788272e-01	0.5627058
People's Party	as.factor(province)Manitoba	-13.0117421	0.0000021	-6.243505e+06	0.0000000
People's Party	as.factor(province)New Brunswick	6.7110174	57.0811788	1.175697e-01	0.9064086
People's Party	as.factor(province)Newfoundland and Labrador	-13.6882298	0.0016742	-8.175828e+03	0.0000000
People's Party	as.factor(province)Nova Scotia	8.1580972	65.7781988	1.240243e-01	0.9012960
People's Party	as.factor(province)Ontario	-0.3270036	1.3073438	-2.501282e-01	0.8024882

y.level	term	estimate	std.error	statistic	p.value
People's Party	as.factor(province)Prince Edward Island	-14.0493784	0.0011902	-1.180384e+04	0.0000000
People's Party	as.factor(province)Quebec	-0.3810350	1.3901281	-2.741006e-01	0.7840073
People's Party	as.factor(province)Saskatchewan	9.0622088	96.9096385	9.351190e-02	0.9254969