Predicting the Canadian Federal Election Popular Vote Using Logistic Regression Models and Post-Stratification

Alexandre Haulard - (1006173604)

Assignment 3: STA304 - 05/11/2021

Introduction

Motivation for the Analysis

The goal of this analysis is to use statistical models in order to predict who will win the upcoming Canadian federal election popular vote. More specifically, logistic regression models will be used along with post-stratification in order to predict whether the Liberal or Conservative party will win the popular vote. This will be achieved using both data from a survey and a census. These data provide a wide set of information about Canadian voters. A few examples are age, income, religion, candidate preference, whether they were born in Canada, and many others.

This analysis is motivated by the recent 2021 Canadian Federal election which arrived earlier than expected. The early election was a surprise to many Canadians who were not familiar with the "Dissolution of Parliament" process. The governor general, following the advice of the prime minister, had to issue proclamations to begin dissolution (*1. Senate of Canada.*).

This analysis should be of interest to many Canadians who were closely following the recent election and expressed confusion about the dissolution of parliament. Some argue that this election was an attempt from Prime Minister Justin Trudeau to gain back majority in the House of Commons, which he lost in the 2019 federal election. Others argue that the election is an opportunity for Canadians to have a say during this challenging pandemic (*2. Austen, Ian.*).

In any event, it would be interesting to create a model that predicts the popular vote for the tentatively 2025 election (or an early election that may occur). The analysis will attempt to answer the following research question: ***Can the Canadian Federal election popular vote be predicted using logistic regression models that depends on a set of variables such as age, education, sex, and income?***

Terminology

In order to understand the following analysis, let's define some niche political terms mentioned in this report.

- **Liberal Party of Canada**: The Liberal Party of Canada is the longest-serving and oldest active federal political party in Canada. The party has dominated federal politics of Canada for much of its history, holding power for almost 70 years of the 20th century.

- **Conservative Party of Canada**: The Conservative Party of Canada, colloquially known as the Tories, is a federal political party in Canada. It was formed in 2003 from the multiple right-leaning parties which had existed in Canada for over a century, historically grouped into two camps known as the "Red Tories" and the "Blue Tories". (*3. "List of Political Parties in Canada."*)
- **Popular Vote**: Popular vote, in an indirect election, is the total number of votes received in the first-phase election, as opposed to the votes cast by those elected to take part in the final election.

The notion of popular vote is important to keep in mind as Canada does not elect its Prime minister based on total votes. Canada elects its Prime Minister using a riding system. Each riding is an electoral district. The party who secures the most districts wins the election.

Hypothesis

Based on the census and the survey we are using; I hypothesize that our models will predict the conservative party will win the popular vote. This hypothesis is based on the most recent outcome of the Canadian federal election. The liberal party won the election as they obtained most seats, but the conservative had a higher share of total votes in Canada. There were 5,742,635 total votes for the conservative party against 5,556,835 for the liberal party (*4. "Federal Election 2021 Live Results."*).

Data

Collection Process

There are two sets of data used for this analysis. Both data provide intersecting information about Canadian voters. In other words, these data sets both provide information about Canadian citizens such as age, religion preferences, sex, highest education level, or income. The subsection 'Post-Stratification' will explain why we made use of two data sets that have intersecting information. Let's describe how each data set was collected:

- **Canadian Election Study, 2019, Phone Survey**: The 2019 Canadian Election Study was conducted to gather the attitudes and opinions of Canadians during and after the 2019 federal election (*5. "Welcome to the 2019 Canadian Election Study."*). The sample comprised of 66% wireless telephone numbers and 34% land-line telephone numbers. During these phone calls, participants were surveyed about attributes that describes who they are as well as their views regarding political issues (Economy, Education, Employment…). In the following subsection, we will refer to this survey as CES.
- **General Social Survey**: This census was made in the 2017 cycle. The data set was pulled from the CHASS Data Center (*6. "University of Toronto Libraries."*). The General Social Survey program, established in 1985, conducts telephone surveys across the ten provinces. The GSS is recognized for its regular collection of cross-sectional data that allows for trend analysis, and its capacity to test and develop new concepts that address current or emerging issues. In the following subsection, we will refer to this survey as GSS.

Cleaning Process

The cleaning process mostly consists of making both the CES and the GSS data match in variables type and value names. This is done in order for post-stratification (explained in 'Methods') to be performed correctly.

CES Data Cleaning

The variables we have used from the original CES data frame are 'q3', 'q4', 'q11, 'q61', 'q63', 'q69' 'p55'. These variables names provide little information. Therefore, all these variables were renamed to their equivalent English meaning using the data documentation. Most of them are demographic variables that will be used for our models. They were renamed as 'province', 'sex', 'age', 'education', 'family income', 'religion importance', 'born in Canada', and if 'one parent is born outside Canada'.

Two logistic regression models will be used. One determines the share of conservative popular vote, and the other determines the share of liberal popular vote. This is done using the variable 'p55' which provides information about which party a voter will likely vote for. From this CES survey, we create two data frames for each model. The first data frame contains all the demographic variables and 'p55' renamed as 'vote liberal' which is binary. A value of 1 means a citizen is likely to vote for the liberal party. A value of 0 means a citizen is not likely to vote for the liberal party which could be conservatives, NDP, Green, or any other Canadian party.

The second data frame contains the exact same demographic variables but uses 'p55' as a binary variable to know if a citizen is likely to vote for the conservative party. The variable is renamed as 'vote conservative'. A value of 1 means a citizen is likely to vote for the conservative party. A value of 0 means a citizen is not likely to vote for the conservative party which could be liberal, NDP, Green, or any other Canadian party.

The variable 'family income' was continuous. We have made it categorical in bins of $25,000 which matches the 'family income' variable from the GSS data.

Furthermore, all the demographic variables were converted from integers to strings. For example, a value of 7 means a person is from Manitoba in the original data. It is now 'Manitoba' in our cleaned data frames. The matching was done using the original study documentation. For most variables, missing values and values recorded as 'Don't Know' were removed. Only the variable 'religion importance has kept values recorded as "Don't know".

GSS Data Cleaning

The original census has 'age' recorded as a continuous variable with decimals. This variable was kept as continuous but rounded to integers such that it matches the CES data. Additionally, the variable 'education' was formatted such that it matches the CES. For example, "Bachelor's degree (e.g. B.A., B.Sc., LL.B.)" was renamed as 'bachelor'.

Finally, all observations with missing values were removed.

Definition of variables

This subsection serves as reference for the definition of the variables that are used throughout the analysis, and within our logistic regression models.

- **Province**: The variable contains the 10 Ontario provinces, but does not include the three territories. (*7. "Provinces and Territories of Canada."*)
- **Sex**: Contains the values 'Male' and 'Female'
- **Age**: Provides the age of a voter.
- **Education**: The variable provides the highest level of education of a voter. The self explanatory options are 'less than high school' degree, 'high school' degree, 'bachelor', and 'above bachelor'. The option 'trade or college' also includes participants who did not finish their college degree. The option 'some university' includes participants who did not finish their bachelor degree.
- **Income Family**: The sum of incomes earned by each member of a household.

- **Religion Importance**: Participants were asked the following question: In your life, would you say religion is VERY important, somewhat important, not very important, not important at all? "Don't know" was also an option that is included in our cleaned data frame.
- **Born in Canada**: Records whether a participant was born in Canada.
- **One parent born outside Canada**: Records whether at least one parent is born outside of Canada.
- **Vote Liberal**: A binary variable that records whether a person is likely to vote for the liberal party.
- **Vote Conservative** A binary variable that records whether a person is likely to vote for the conservative party.

Numerical Variable Summary

The only continuous variable for which we can calculate summary statistics is 'age'.

Table 1: Mean Summary of Age for Data Types

Data	Min	Mean	Median	Standard Deviation	Max	n
CES	18.00	54.43	56.00	15.46	95.00	1415
GSS	18.00	52.83	55.00	17.24	80.00	19357

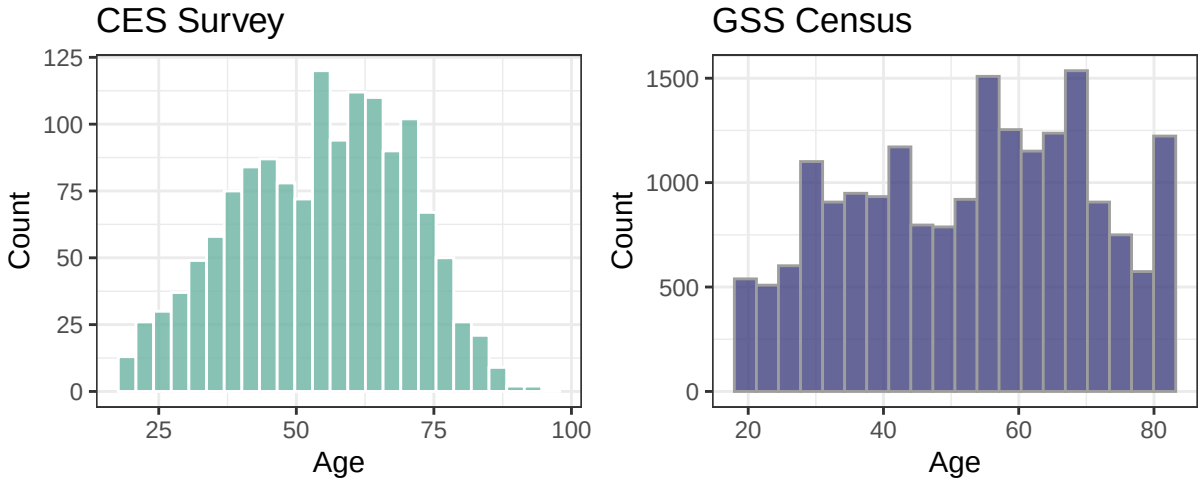
The CES data, which is our survey, contains 1415 observations. The census data contains 19357 observations. Theoretically, it is not a true census. Nevertheless, the GSS data can be reasonably considered a census for our purposes.

The min age is 18 which means all the participants were adults. This makes sense as Canadian citizens must be at least 18 years of age in order to vote. The max age is 95 for the survey data, but 80 for the census data. A census should cover the whole population, so we imagine that the max age of the census should be higher than 80. We can conclude that in the GSS data collection process, the population was defined such that participants cannot be older than 80.

The mean and standard deviation of the CES survey are quite close to the "true" population mean for age defined by the GSS census. Thus, the survey is representative of the population mean for age.

Finally, the standard deviation given by the survey is 15.46 while the population standard deviation is 17.24. Let's use histograms to visualize the distribution of age.

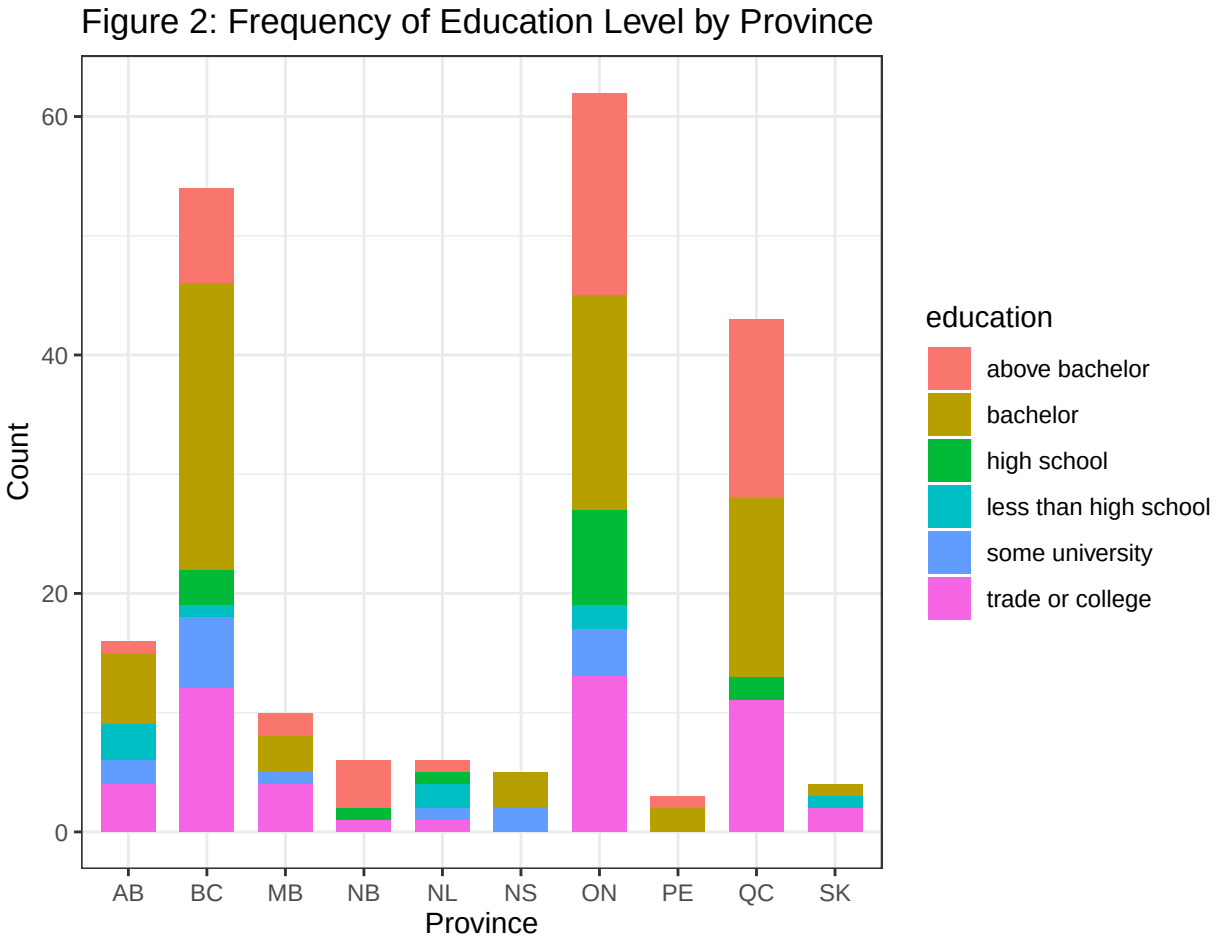
Figure 1: Distribution of Age for the Survey and Census



The data looks more normally distributed for the CES survey. The data in the CES survey is more concentrated around its mean. On the other side, the data from the census appears to be more spread out. This makes sense as the standard deviation from the census was larger.

Categorical Variables Summary

Let's use a frequency chart to visualize the education level conditional on the 10 Canadian provinces from the CES survey.



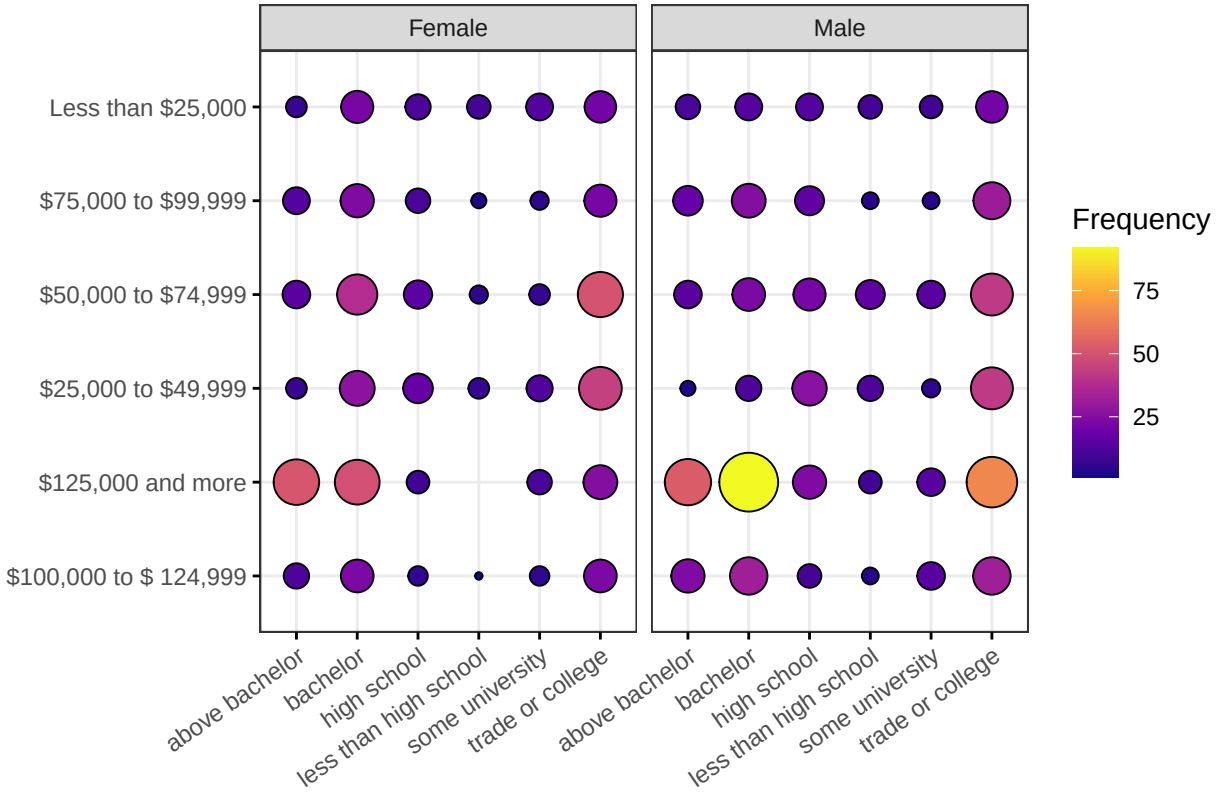
Notice that Ontario has the highest bar which makes sense as it is the province with the highest population. The predominant highest level of education is a bachelor degree in all provinces except in Manitoba, New Brunswick and Newfoundland and Labrador.

The proportion of people with an education level beyond a bachelor degree is quite high in New Brunswick.

Alberta has the highest amount of observation with an education level that is less than high school. Interestingly, there are no observation with a high school diploma in Alberta. This is unfeasible and can hint that the survey is not representative of the true population in terms of the education variable.

Let's use a balloon plot to visualize the frequency of 3 of the categorical variables at the same time. These variables are 'family income', 'education level', and 'sex'.

Figure 3: Balloon Plot for Salary Vs. Education Level given Sex



The plot has quite a lot of information. The vertical axis displays household income. The horizontal axis provides the education level. The left chart are observations conditional on being a female while the right chart are observations conditional on being a male. The size and color of each circle changes as the frequency for a certain observation increases. For example, there are more than 75 occurrences of males with a bachelor degree making $125,000 or more. Another example is females making $125,000 or more with an education level above a bachelor degree. The circle is red. This means there are between 50 to 75 occurrences for this example.

Finally, let's talk about the mean of the binary variables 'vote liberal' and 'vote conservative'. 'Vote liberal' has a mean of 0.252 while 'vote conservative' has a mean of 0.302. These numbers represent the proportion of participant who said they were likely to vote for a certain party on the survey.

In the next section, the logistic regression model will be introduced along with the post-stratification technique. Two models will be computed with the dependent variables 'vote liberal' and 'vote conservative'. Both models will have the exact same independent variables: 'sex', 'family income', 'education level', 'religion importance', 'born in Canada', and 'one parent born outside Canada'.

Methods

This section introduces the logistic regression model for a general science reader. We will later compute the model in the results section. Additionally, the post-stratification method will be explained. This technique is key as it will allow us to make a final conclusion about whether the conservative or liberal party will win the popular vote.

Logistic Regression

Before introducing logistic regression, let's note that we are not using a linear regression model as it is not appropriate to model a binary response variable.

The logistic regression model is a binary classification model in which the conditional probability of one of the two possible realizations of the output variable is assumed to be equal to a linear combination of the input variables, transformed by the logistic function (*8. Taboga, Marco. "Logistic Regression by Marco Taboga."*).

Some example of using logistic regression are: determining if an email is spam or not, identifying if a patient's tumor is malignant or benign (*9. "Introduction to Logistic Regression, Samantha-Jo Caetano"*). In our case, the model is perfect to determine binary outcomes such as whether a person will vote for a certain political party or not.

The general model is defined as follows: $log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$ where each x_i can be a continuous, categorical, and/or dummy variable. Each parameter which we aim to estimate represent change in log odds. The regression coefficients are usually estimated using maximum likelihood estimation. This is quite statistically rigorous and will not be defined here (*10. "Maximum Likelihood Estimation."*). Nevertheless, the R programming language will handle these computations for us without having to worry about hefty theory. Let's define the independent variables used for the two logistic regression models.

d_1 = Is Male
x_2 = Age

d_3 = Education: Bachelor
d_4 = Education: High School
d_5 = Education: Below High School
d_6 = Education: Some University
d_7 = Education: Trade or College

If all these variables are equal to 0, then we are estimating the outcome for an observation with an education level that is above a Bachelor degree.

d_8 = Family Income: \$125,000 and more
d_9 = Family Income: \$25,000 to \$49,999
d_{10} = Family Income: \$50,000 to \$74,999
d_{11} = Family Income: \$75,000 to \$99,999
d_{12} = Family Income: Less than \$25,000

If all these variables are equal to 0, then we are estimating the outcome for an observation with a family income that is between \$100,000 and \$124,999.

d_{13} = Religion Importance: Not at all important
d_{14} = Religion Importance: Not very important
d_{15} = Religion Importance: Somewhat important
d_{16} = Religion Importance: Very important

If all these variable are equal to 0, then we are estimating the outcome for an observation who "Doesn't Know" (as recorded in the survey) whether religion is important or not.

d_{17} = Is Born in Canada
d_{18} = Has one parent born outside Canada

Note that there is only one continuous variable which is 'age' defined as x_2. All other variables are dummy variables that takes on the value 1 or 0.

- **Logistic Regression for determining if a person will vote for the liberal party**

$$\begin{aligned} log(\frac{p_{lib}}{1-p_{lib}}) = \beta_0 + \beta_1 d_1 + \beta_2 x_2 \\ + \beta_3 d_3 + \beta_4 d_4 + \beta_5 d_5 + \beta_6 d_6 + \beta_7 d_7 \\ + \beta_8 d_8 + \beta_9 d_9 + \beta_{10} d_{10} + \beta_{11} d_{11} + \beta_{12} d_{12} \\ + \beta_{13} d_{13} + \beta_{14} d_{14} + \beta_{15} d_{15} + \beta_{16} d_{16} \\ + \beta_{17} d_{17} + \beta_{18} d_{18} \end{aligned}$$

- **Logistic Regression for determining if a person will vote for the conservative party**

$$\begin{aligned} log(\frac{p_{con}}{1-p_{con}}) = \beta_0 + \beta_1 d_1 + \beta_2 x_2 \\ + \beta_3 d_3 + \beta_4 d_4 + \beta_5 d_5 + \beta_6 d_6 + \beta_7 d_7 \\ + \beta_8 d_8 + \beta_9 d_9 + \beta_{10} d_{10} + \beta_{11} d_{11} + \beta_{12} d_{12} \\ + \beta_{13} d_{13} + \beta_{14} d_{14} + \beta_{15} d_{15} + \beta_{16} d_{16} \\ + \beta_{17} d_{17} + \beta_{18} d_{18} \end{aligned}$$

Again, these models are the same in regards to the independent variables. But, they differ only with the binary response. The first model estimates the log odds that a person will vote for the liberal party. The second model estimates the log odds that a person will vote for the conservative party.

Holding everything else constant, the interpretation of β_2 is as follows: for every 1 year increase in age we expect the log odds to increase by β_2. An example of the interpretation of parameters of binary variables is as follows: given an observation whose highest education level is a bachelor degree, we expect the log odds to increase by d_3. Same reasoning applies to all the other dummy variables.

In the results section we will find estimates $\hat{\beta_0}, \hat{\beta_1}, ..., \hat{\beta_{17}}$ for each of the parameters.

After computing these parameters, we will assess whether some variables are necessary in our model. If we find variables that have no statistical significance in explaining the response, these variables will be dropped. The assessment will be made using the following selection techniques:

- **t-test**: This is is an hypothesis test. The null hypothesis is formulated as follows: $H_0 : \beta_i = 0$. If we can reject H_0, it means that the i predictor is significantly related to the response. If the null cannot be rejected, then we can make a second model without this parameter and check if the new model is more efficient using an AIC or BIC test. A 5% significance level will be used. If most categories in a variable set are statistically insignificant, we will remove the whole variable set.
- **AIC**: The equation of this test is as follows; $AIC = 2k - 2\ln(\hat{L})$. The number of predictors is k, so more predictors increases the value of AIC. $\hat{L}$ is the maximum value of the likelihood function. Therefore, if our model fits well, the second term of AIC would be higher. This means AIC would be lower. Thus, when comparing models, we are looking for the one with the lowest AIC (*11. "Akaike Information Criterion."*).
- **BIC**: The equation of this test is as follows; $BIC = k\ln(n) - 2\ln(\hat{L})$. The intuition behind this formula is the same as described above for AIC. A lower BIC is preferred (*12. "Bayesian Information Criterion."*).

Post-Stratification

Post-stratification is a statistical technique used for correcting model estimates for known differences between a sample and a population. The technique involves using data from, for example, censuses relating to various types of people corresponding to different characteristics such as age, race, or education level (*13. "STA304", Multilevel Regression & Poststratification*).

The model is mathematically defined as follows: $\hat{y}^{PS} = \frac{\sum N_j \hat{y_j}}{\sum N_j}$. We can then compare $\hat{y}^{PS}$ for both the liberal and conservative models to determine who will win the popular vote.

In order to estimate the proportion of voters, we can finally use the census data to apply post-stratification. It is appropriate to use post-stratification as the data from the CES survey is not properly balanced. We have determined in the data section that certain variables in the survey are not representative of the true population.

The census data contains all variables that were used in our model that is computed using the survey data. This will allow us to first compute estimates of probabilities from our models. The second step is to create bins by grouping estimates. These bins will be implemented using provinces because it is likely to influence voter outcome. Therefore, there will be 10 bins as there are 10 provinces.

The results of $\hat{y}^{PS}$ for both models will be computed and compared in the results section. The highest $\hat{y}^{PS}$ will determine which party wins the popular vote assuming other parties cannot do better.

Results

This section is divided into two subsection. One for displaying the logistic regression models results. The second one for deciding who will win the popular vote using the post-stratification results.

Logistic Regression Models

Let's compute the estimate of parameters for both models. We will then check the statistical significance of these parameters. If necessary, a second version of each model without statistically insignificant parameters will be created. Post-stratification will be performed with the models that have the lower AIC and BIC. The table of estimates is found on the next page.

Before interpreting the parameters, let's analyze their statistical significance. As said in the methods section, a set of variables which does not have at least one category with a statistical significance of 5% will be removed. This is the case for 'Family Income' and 'Religion Importance'. T-tests shows that these variables give poor justification for the response as opposed to other predictors. Age is statistically insignificant in the conservative model but not the liberal one. Age will not be dropped in the conservative model so that both model can stay similar in predictors. We now have a second version of both models. Both the AIC and BIC are lower in the second version of the liberal model. Therefore, the Liberal V2 model is better than the previous model and will be used for interpretation and post-stratification. The V2 Model for conservatives has lower BIC than the previous one but a higher AIC. In that case, we could argue that both models are almost as efficient. Therefore, we will use the Conservative V2 model for post stratification in order to be consistent with the Liberal V2 model.

The 'is male' estimate is negative for liberal and positive for conservative. The interpretation is that given a male, log odds of voting for liberal decreases by 0.212. Given a male, log odds of voting for conservative increases by 0.697. There is a positive relationship regarding age in the liberal model, but none in the conservative model. Given a person with an education level 'trade or college', logs odds of voting for liberal decreases by -0.909, but increases by 0.609 for conservative. Same logic applies for the interpretation of other education levels. Given a person is born in Canada, the log odds of voting for liberal decreases by 0.734, while it increases by 0.134 for conservative. Note that the parameter estimate of this variable has a high statistical significance in the liberal model, but none in the conservative model.

Table 2: Estimation of Logistic Regression Models' parameters for Liberals and Conservatives

	Liberal	Conservative	Liberal V2	Conservative V2
(Intercept)	−14.652	−2.496*	−0.649*	−1.836***
	(426.194)	(1.246)	(0.304)	(0.313)
Is Male	−0.216+	0.701***	−0.212+	0.697***
	(0.128)	(0.127)	(0.126)	(0.122)
Age	0.014**	0.000	0.014**	0.000
	(0.004)	(0.004)	(0.004)	(0.004)
Education: Bachelor	−0.294	0.393+	−0.292	0.357+
	(0.183)	(0.205)	(0.181)	(0.200)
Education: High School	−0.392+	0.769**	−0.392+	0.617**
	(0.230)	(0.238)	(0.221)	(0.227)
Education: Below High School	−0.547+	0.863**	−0.574+	0.660*
	(0.310)	(0.298)	(0.300)	(0.284)
Education: Some University	−0.678*	0.743**	−0.676*	0.596*
	(0.267)	(0.266)	(0.263)	(0.257)
Education: Trade or College	−0.904***	0.733***	−0.909***	0.609**
	(0.198)	(0.204)	(0.192)	(0.195)
Family Income: $125,000 and more	0.009	0.420*		
	(0.212)	(0.201)		
Family Income: $25,000 to $49,999	−0.070	−0.085		
	(0.247)	(0.235)		
Family Income: $50,000 to $74,999	0.056	0.071		
	(0.230)	(0.220)		
Family Income: $75,000 to $99,999	0.121	−0.090		
	(0.248)	(0.243)		
Family Income: Less than $25,000	−0.018	−0.357		
	(0.260)	(0.257)		
Religion Importance: Not at all important	13.780	−0.133		
	(426.194)	(1.198)		
Religion Importance: Not very important	13.813	−0.047		
	(426.194)	(1.187)		
Religion Importance: Somewhat important	13.903	0.393		
	(426.194)	(1.181)		
Religion Importance: Very important	13.715	0.907		
	(426.194)	(1.182)		
Is Born in Canada	−0.617**	0.233	−0.734***	0.134
	(0.198)	(0.210)	(0.161)	(0.172)
Has one parent born outside Canada	0.207	−0.024		
	(0.163)	(0.158)		
N	1415	1415	1415	1415
AIC	1565.7	1670.4	1552.4	1700.7
BIC	1665.6	1770.2	1599.7	1748.0

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Post-Stratification Results

First, let's compute a table with the bins for which we will perform post-stratification.

Table 3: Cell/Bin splits use for Post-Stratification

Province	n	Probability of Voting Liberal	Probablility of Voting Conservative
Alberta	1619	0.25	0.30
British Columbia	2353	0.27	0.29
Manitoba	1104	0.25	0.31
New Brunswick	1257	0.23	0.30
Newfoundland and Labrador	1031	0.22	0.31
Nova Scotia	1358	0.24	0.30
Ontario	5268	0.28	0.29
Prince Edward Island	665	0.23	0.30
Quebec	3624	0.24	0.30
Saskatchewan	1078	0.23	0.31

The variable 'n' represents the number of observations in each province. Analyzing results about voting for a certain party, we see that the probability of voting for liberals is highest in Ontario at 28%. It stands at 29% for conservatives. This means we expect 57% of Ontarians to vote for the two major parties. The rest is expected to vote for other parties (E.g. NDP, Bloc Québécois, …). Newfoundland and Labrador has the lowest probability of voting for the liberal party at 22%, but the highest for conservative at 31%. In every province, the overall probability of voting for conservative is higher than the probability of voting for liberals. From here, we can already infer that conservatives will win the popular vote without performing post-stratification, but let's still compute it for the sake of completion. Note that the 'n' from the table is the equivalent of N_j in our post-stratification equation.

Probability that a person votes for the liberal party of Canada:

$$\hat{y}^{PS_{lib}} = \frac{\sum N_j \hat{y}_j^{lib}}{\sum N_j} = 25.2079\%$$

Probability that a person votes for the conservative party of Canada:

$$\hat{y}^{PS_{con}} = \frac{\sum N_j \hat{y}_j^{con}}{\sum N_j} = 29.8811\%$$

Since $\hat{y}^{PS_{lib}} < \hat{y}^{PS_{con}}$, we can conclude that conservatives will win the popular vote in the next federal election. According to our model and post-stratification results, 29.8811% of voters will choose the conservative party and 25.2079% will choose the liberal party. To make that conclusion, we assumed that liberals and conservative cannot be beat by other parties (E.g. NDP, Bloc Québécois, …) in terms of popular vote.

Comparing to past elections, our results seem reasonable but not ideal. During the 2021 election, conservative won the popular vote at 33.7% with liberals just behind at 32.6% (*14. Tahirali, Jesse, and Phil Hahn.*). The $\hat{y}^{PS_{lib}} = 25.2079\%$ is quite far from 32.6% which is why I would not conclude our model is ideal. Nevertheless, it is reasonable since it rightly predicts conservatives would win the popular vote.

Conclusions

In the introduction section, we asked the following research question: ***Can the Canadian Federal election popular vote be predicted using logistic regression models that depends on a set of variables such as age, education, sex, and income?*** To answer this question, I would re-iterate the interpretation of the results section. Our model is not ideal to estimate the exact proportion of popular vote for each party, but does a fairly good job at predicting the outcome for who wins the popular vote. To arrive at this conclusion, logistic regression models were used in addition to the post-stratification technique.

The hypothesis we made earlier has guessed the right outcome of the study. Nevertheless, the results don't provide support that is convincing enough since the proportion estimate of liberal voters greatly differs from the one in the closest election. The key results were that the proportion of conservative voters was higher than the proportion of liberal voters.

Talking about the big picture, predicting the outcome of an election is complicated even when using appropriate statistical methods. There are too many unknowns that change voter opinion. The obvious and most recent unknown is the COVID-19 pandemic. This event which started in March 2020 can greatly change the opinions of voters.

Weaknesses

The first weakness of our study is that the CES survey was made in 2019, and the GSS census in 2017. The pandemic most likely changed the political views of many voters. The second weakness is sampling biases present in the CES survey. There is most likely a selection bias due to the fact it was conducted by phone. Additionally, it is very likely that the survey contains response biases. Many participants could have inaccurately answered questions asking about their salary, religion, or any other information that they find sensitive to report. The third weakness concerns the choice of predictors in our model. It is quite hard to guess which independent variable provides the best explanation to our response. Intuitive ones we have used are income and education, but family size, spouse's age, spouse's education, parents' occupation, and/or many other unforeseeable predictors might explain the response.

Next Steps and Discussion

Future work would involve correcting the above mentioned weaknesses. Conducting a more recent survey with less biases is a first step. Testing different models and/or different predictors is another option.

Finally, to summarize this report, despite some weaknesses we have shown using logistic regression and post-stratification that conservatives would win the popular vote in an upcoming election.

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

Bibliography

In Line Referencing

1. Senate of Canada. "What Is Dissolution of Parliament?" Senate of Canada, https://sencanada.ca/en/sencaplus/how-why/what-is-dissolution-of-parliament/. (Last Accessed: November 4, 2021)

2. Austen, Ian. "Why Did Justin Trudeau Call for an Early Election?" The New York Times, The New York Times, 20 Sept. 2021, https://www.nytimes.com/2021/09/20/world/canada/justin-trudeau-why-early-election.html. (Last Accessed: November 4, 2021)

3. "List of Political Parties in Canada." Wikipedia, Wikimedia Foundation, 27 Aug. 2021, https://en.wikipedia.org/wiki/List_of_political_parties_in_Canada. (Last Accessed: November 4, 2021)

4. "Federal Election 2021 Live Results." CBCnews, CBC/Radio Canada, https://newsinteractives.cbc.ca/elections/federal/2021/results/. (Last Accessed: November 4, 2021)

5. "Welcome to the 2019 Canadian Election Study." Canadian Election Study, http://www.ces-eec.ca/. (Last Accessed: November 4, 2021)

6. "University of Toronto Libraries." My.access - University of Toronto Libraries, https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/index.htm. (Last Accessed: November 4, 2021)

7. "Provinces and Territories of Canada." Wikipedia, Wikimedia Foundation, 26 Oct. 2021, https://en.wikipedia.org/wiki/Provinces_and_territories_of_Canada. (Last Accessed: November 4, 2021)

8. Taboga, Marco. "Logistic Regression by Marco Taboga." Logistic Classification Model (Logit or Logistic Regression), https://www.statlect.com/fundamentals-of-statistics/logistic-classification-model. (Last Accessed: November 4, 2021)

9. "Introduction to Logistic Regression, Samantha-Jo Caetano", STA304 (Last Accessed: November 4, 2021)

10. "Maximum Likelihood Estimation." Wikipedia, Wikimedia Foundation, 21 Oct. 2021, https://en.wikipedia.org/wiki/Maximum_likelihood_estimation. (Last Accessed: November 4, 2021)

11. "Akaike Information Criterion." Wikipedia, Wikimedia Foundation, 21 Oct. 2021, https://en.wikipedia.org/wiki/Akaike_information_criterion. (Last Accessed: November 4, 2021)

12. "Bayesian Information Criterion." Wikipedia, Wikimedia Foundation, 5 Nov. 2021, https://en.wikipedia.org/wiki/Bayesian_information_criterion. (Last Accessed: November 4, 2021)

13. "STA304", Multilevel Regression & Poststratification (Last Accessed: November 4, 2021)

14. Tahirali, Jesse, and Phil Hahn. "Six Charts to Help You Understand the 2021 Federal Election." CTVNews, CTV News, 27 Sept. 2021, https://www.ctvnews.ca/politics/federal-election-2021/six-charts-to-help-you-understand-the-2021-federal-election-1.5598419. (Last Accessed: November 4, 2021)

Software Packages Used to Complete this Report

1. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

2. Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.1. https://CRAN.R-project.org/package=stargazer

3. Vincent Arel-Bundock (2021). modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready. R package version 0.9.2. https://CRAN.R-project.org/package=modelsummary

4. David B. Dahl, David Scott, Charles Roosen, Arni Magnusson and Jonathan Swinton (2019). xtable: Export Tables to LaTeX or HTML. R package version 1.8-4. https://CRAN.R-project.org/package=xtable