Predicting the Next Canadian Federal Election

STA304 - Assignment 2

Group 86: Mary Jones, Michael Tsimidis, Patrice Yee and Aryan Niraj Kishan

December 1, 2022

Introduction

The main objective of this report is to use a regression model with post-stratification to predict the overall popular vote for the upcoming Canadian federal election (tentatively 2025).

For the analysis involved in this report, we will use two types of data — the census data and the survey data. The 2017 General Social Survey or GSS will serve as our census data, this includes every non-institutionalized individual over the age of 15 that is residing in one of the ten provinces of Canada. This 2017 GSS data was collected from February 2nd to November 30th, 2017 via telephone. The 2019 Canadian Election Study (CES) phone survey will serve as our survey data, this includes Canadian citizens and permanent residents aged 18 and above. This data was collected during and after the 2019 Canadian federal election through computer-assisted telephone interviews.

This analysis is important because we will combine a multiple logistic regression model with post-stratification. Here, post-stratification will allow us to adjust the weights of each of the predictor variables to represent the population. This would make our final predictive model built using sample data more accurate [2]. More specifically, the variables "Province" , "University Completion", and "Employment" will be used to conduct post-stratification.

The research question for our multiple logistic regression model with post stratification can be framed as follows: "Are province, university completion, and employment factors that predict if a person will vote for the Liberal Party of Canada?".

Furthermore, we can represent this question in the form of a null and alternative hypothesis. Here the null hypothesis states that none of the predictive covariates have a statistically significant relationship with our dependent variable. On the other hand, the alternative hypothesis states that at least one predictive covariate has a statistically significant relationship with our dependent variable [4].

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$
$$H_1 : \beta_1 = \beta_2 = \beta_3 \neq 0$$

Overall, this hypothesis test is important because it will help us determine whether our model is statistically significant and whether it can actually predict if an individual will vote for the Liberal Party of Canada.

Here, "Province" refers to the province of residence of the individual, "University Completion" is a binary variable that indicates whether or not the person has attained a university degree, and "Employment" is a binary variable that indicates whether or not the individual has worked in the last week.

Data

The 2017 General Social Survey data is a sample survey with a cross-sectional design, facilitated from February 2nd through to November 30th, 2017. The target population of the survey is all non-institutionalized persons above the age of 15, living in Canada. The survey uses telephone numbers combined with Statistics Canada's Address Register to collect data via telephone. The data is subject to sampling and non-damping errors. The data collected issued to meet two primary objectives; to monitor changes in living conditions and well-being of Canadian individuals and to gather information on specific social policy issues of current or arising interest. In order to meet these objectives data collected by the GSS contains core content and classification variables. The core content data contains information about individuals living conditions and well-being to help inform specific policies. Classification variables such as age, income, education, marital status, and sex help to classify population groups to better analyze the core data within them.

Data collection

In order to collect the data each of the ten provinces in Canada was divided into strata, and within the ten provinces, many of the Census Metropolitan Areas were considered separate strata, for a total of 27 strata. A simple random sample without replacement of records was performed within each stratum. The data was collected via computer-assisted telephone interviews. Responses from the interviews were entered directly into computers as the interview went on. The data was then electronically transmitted to Ottawa. Several questions in the data allowed for write-in responses. These responses were either categorized into existing groups or left in "other specify" if a match was not possible within the existing data. As the survey was conducted the computer would create edits and identify "out of range" values that the interviewer could clarify and resolve with the individual being surveyed. If issues were unable to be resolved the errors were forwarded to the Head Office for resolution. Non-response was not permitted for questions that the survey later required to use for the weighting of individuals. In the case where important values were missing values were imputed. In 2017, personal income information was not obtained through the survey but rather collected from tax data for the respondent. Finally, the timing that respondents experienced given life events was very important to the survey and was collected from individuals' recollections. Originally the month and year of a particular event were asked, however, if individuals were unable or refused to provide such information then the age at which he or she experienced the event is asked. For some situations, the answers required imputation to derive the age, month, and year of the occurrence of events for individuals.

In the 2019 Canadian Election Study, a Phone survey was conducted to gather information on the underlying reasons why people make the voting decisions they do, to evaluate what does and does not change during the campaign from one election to another, and to highlight similarities and differences between elections and voting in Canada to other democratic countries. The study was conducted on Canadian citizens and permanent residents age 18 or older during and after the 2019 federal election. The data was collected through computer-assisted telephone interviews, consisting of 4021 cases and 278 variables. The sample consisted of 66% wireless telephone numbers and 34% landline telephone numbers.>

Data Cleaning

The census data was cleaned by changing the values of variables to be easier for computation and reduced to contain only important variables with individuals with all selected variable values recorded. The variable province was mutated to two different categories of east and west to be easier for computation. Religious importance is converted to a dummy variable with a value of 1 if religion is important at all and 0 if not. Age is mutated into categories representing the same age categories as the survey data. Education is also converted to a dummy variable with a value of 1 for a University certificate, diploma or degree above the bachelor's and 0 for College, CEGEP or other non-university certificate or diploma. The variable married and employed are also converted to dummy variables with a value of 1 if the individual is married or employed and 0 otherwise. The data was then reduced by selecting to only contain variables province, religion_important, age, income_family, completed_university, married, and employed. The data then was lastly cleaned by omitting any rows of data that have NA values in the selected variables.

Variables of Interest

The relevant variables of interest we will be using when determining the model and for post-stratification are *vote_liberal*, *age*, *province*, *completed_university*, *religion_important*, *income_family*, *married*, and *employed*. There are many redundant variables in the survey data set and so in the cleaning process, we removed variables that seemed to be overlapping with others by using the select function. The variables we selected to use are especially informative as to which way an individual tends to lean politically. Age is a categorical variable for the age of the individual over 18. Province points to whether an individual lives in the east or west of Canada. Completed University is a dummy variable that has a value of one if an individual obtained university certification. Religion_important is another dummy variable which has value one if religion is important to the individual. en_first_language is another dummy variable which has value one if the individual has English as their first language and 0 otherwise. Income_family is a categorical variable that depicted the income of the individual. The last two variables denoted married and employed are also dummy variables with values 1 if the individual is married/employed and 0 otherwise. All of these variables will be important in predicting the vote_liberal dummy variable which has a value of 1 when an individual votes liberal and 0 otherwise.

Table 1 Summaries of Survey Data

Variable	Proportion	Number of Observations	Standard Deviation	Standard Error
Voted Liberal	0.31	1769	0.46	0.011
University Graduates	0.43	1769	0.5	0.0118
Religion is Important	0.72	1769	0.45	0.0107
Currently Employed	0.54	1769	0.5	0.0119
English First Language	0.69	1769	0.46	0.011

Table 2 Summaries of Census Data

Variable	Proportion	Number of Observations	Standard Deviation	Standard Error
University Graduates	0.32	18766	0.46	0.0034
Religion is Important	0.64	18766	0.48	0.0035
Currently Employed	0.52	18766	0.5	0.0036

The numerical summaries in *Table 1* and *Table 2* look at the proportion of the population or survey sample that exhibits super characteristics. The summaries also state the devotion from the proportion, the number of observations or individuals analyzed and the standard error. In looking at the vote_liberal variable in the survey data we can see that from 1769 individuals approximately 31% of the sample voted liberal. We can also see from the survey data that approximately 43% of the survey sample completed university, whereas in the population data of 18766 individuals only approximately 32% completed university. In investigating another variable, religion_important, we can see that in the survey sample, 72% reported religion being of importance to them, whereas in the population data a smaller 64% reported religion being of importance. In the survey data for employment, we can see that only approximately 54% of individuals were employed. The population census data from employment is very similar to the sample with approximately 52% of individuals being employed. From the survey data, we also see that roughly 69% of individuals have English as their first language. In comparing the survey and census data the census data consistently has a lower standard error which makes sense for the increase in observations. The standard deviation of variables between the population and sample is relatively similar.

Figure 1: Surveyed Family Income

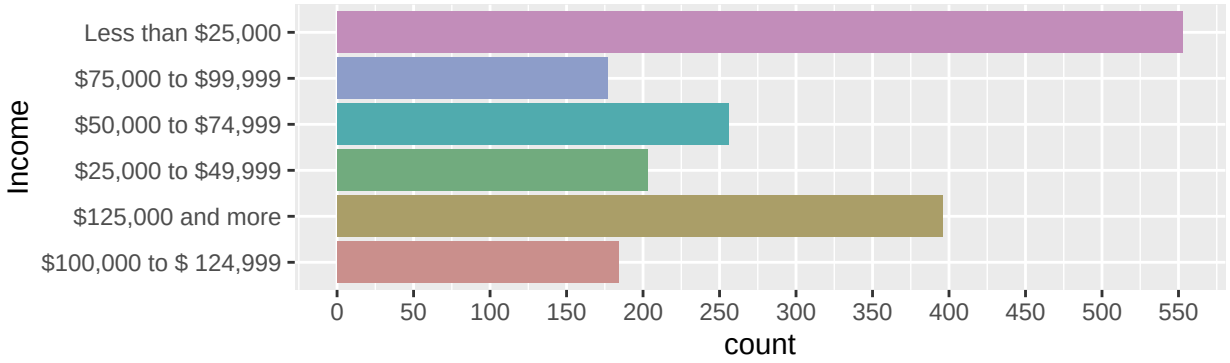
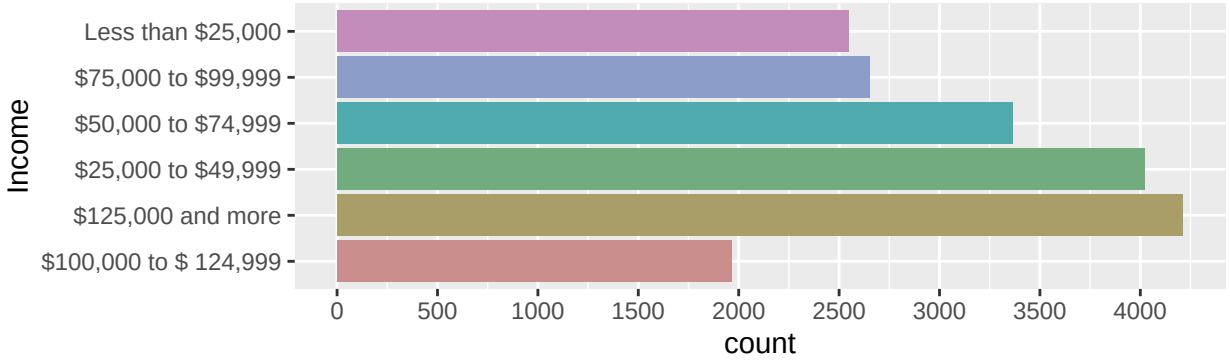
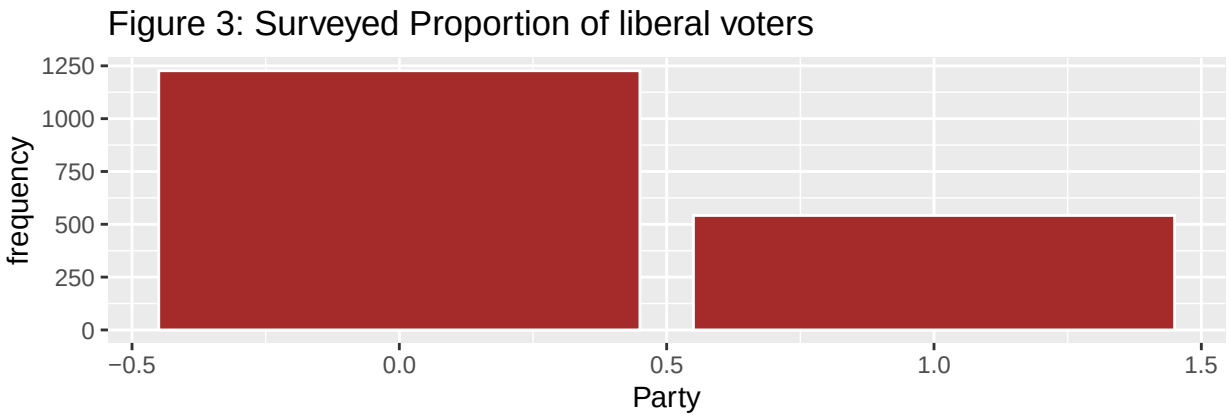


Figure 2: GSS Family Income





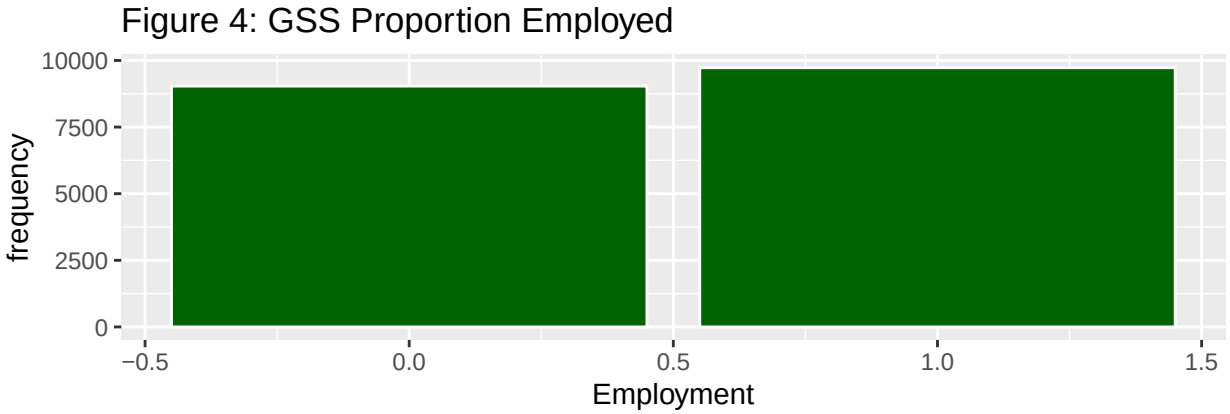


Figure 1 and 2: "Survey and Census Family Income: Depict the household incomes of those surveyed in CES2019 and GSS. CES2019 has a much larger prevalence of poorer households than in GSS.

Figure 3. "Surveyed Proportion of Liberal Voters": Frequencies of liberal vote values. The variable takes values 0 and 1 which gives us the two columns at 0 and 1 on the x-axis. From the plot, we can see the survey data contains substantially more non-liberal voters than liberal voters.

Figure 4: "GSS Proportion Employed": Frequencies of individuals in the survey being. We can see from the plot that there is almost an equal amount of employed and unemployed individuals from the census level data.

Methods

In the endeavor of producing an estimate of the outcome of the next Canadian Federal Election, the results from the Canadian Election Survey from 2019 are utilized to create a model of election results. The outcome of a Canadian election is for a government to be formed and this is accomplished by that party winning the largest proportion of electoral zones, referred to as seats or ridings where citizens in that area cast their votes for their preferred candidates. In Canada, citizens are entitled to vote for one party which makes the results for each individual voter a binary outcome. From this, the outcome variable in the model will be binary which reflects the conditions of the election. The binary variable was chosen to represent a vote cast in favor of the Liberal Party of Canada, since that party is the current government and would be incumbents in the next federal election, barring unforeseen circumstances.

A multiple logistic model was selected as they are designed to operate with categorical and binary outcome variables. A logistic model produces an estimate of the odds of a predicted response rather than the explicit predicted response that is produced from a linear model. The odds of a logistic model are denoted by $log(\frac{p}{1-p})$, where p represents the probability of the outcome. Comparatively, a linear model which attempts to fit a line that corresponds to every data point is inappropriate as the binary outcome variable that

describes whether or not a voter will vote liberal will mean the line attaches at the average value of the two extremes of the result. As such, the linear model suffers dramatically in predictive ability as it believes the data is simply a plethora of outliers.

The predictions from this model will be extrapolated to a more representative population in a process called *post-stratification*, in which data points from the General Social Survey that are characteristically similar to data points in the CES2019 data set will be categorized and classified in order to determine a proportional frequency of those classes. From the proportions, the estimates produced by the logistic model can be updated to correspond with the additional information provided by the GSS data.

Model Specifics

The chosen model was constructed with the express goal of minimizing a metric called the *Bayesian Information Criterion* (BIC) which measures the quality of a model by the degree it produces estimates that match the observed values and by incentivising models with fewer predictor variables. BIC is minimised through a process called stepwise selection which iterates between adding variables and removing them from the model until a minimum is determined. The presented model originated from a larger model that contained all relevant variables which were determined to be relevant and were appropriate for post-stratification. Namely, this initial model includes age, province, religion importance, family income, university completion, marital status, and employment. The model produced from the BIC stepwise selection process comprised of province, university completion, and employment. The results from the analysis of BIC are depicted in *Table 3* wherein the differences of BIC between three models, the original model, the model from BIC stepwise selection, as well as a model from Akaike's Information Criterion (AIC) stepwise selection. The AIV and BIC metrics are similar but differ in that BIC places more weight on producing a model with comparatively fewer predictor variables which is desirable when attempting to generate a model for the purposes of predicting rather than explaining already observed data. *Table 3* depicts that the BIC model, shown below, produces the largest difference in BIC compared to the other two models. As well, the BIC model also has a relatively low difference in AIC compared to the AIC model which indicates that the AIC model is only slightly better in terms of minmising the AIC criterion. As such, the BIC model depicted below was selected to perform the election predictions.

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{province} + \beta_2 x_{university} + \beta_3 x_{employed}$$

The variables in model can be interpreted: p represents the probability of a liberal vote, β_0 represents the average change in log odds when all predictor variables are at 0. β_1 represents the average change of the log odds depending on the province where the voter presides. β_2 represents the average change in logs odds depending on if the voter has completed university. β_3 represents the average change in log odds depending on if the voter is currently employed .

Table 3 Model Comparisons

Model	Number of Predictors	ΔAIC	ΔBIC
Original Model	8	0	0
BIC Model	3	6.6	61.4
AIC Model	5	8.8	52.7

Post-Stratification

In the context of this model, post stratification will be applied by categorizing the data in the CES data such that it matches data in GSS which is a much larger data set. From this, proportions of each category can be calculated and applied to estimates generated by the model previously developed. Post-stratification would thus seek to adjust estimates such that over- or under- represented data points from the survey are adjusted

to reflect their prevalence in a wider context. The assumption of post-stratification is that the census level data used is truly more representative and is less biased than the surveyed data.

Mathematically, post-stratification follows where the predicted value for the response is weighted by the number of population level responses:

$$\hat{y}_{ps} = \frac{\Sigma N_j \hat{y}_j}{\Sigma N_j}$$

Assumptions of the Logisitic Model

There are four assumptions present in a Logistic model. First, due to the nature of the construction of the logistic model, the response variable must be a binary categorical variable. This assumption is was addressed by utilising a dummy variable to represent the particular political party which represented whether or not that observation included a vote for that party. The next assumption requires linearity between the log odds $log(\frac{p}{1-p})$ and any continuous variables. However, since the model's presented above do not include continuous variables, this assumption is (not important). Third, there is an assumption about the lack of multicollinearity which can be tested by analysing a metric called the Variance Inflation Factor for each variable. As a rule of thumb, any predictor variable with a VIF greater to 5 to 10 can be classified as being correlated with other predictor variables. As is depicted *Table 4*, the VIF of the predictor variables are low and thus the model does not likely suffer from multicollinearity

Table Multicollinearity

Table. 4 Variance Inflation Factor

Variables	VIF
Province	1.002
Completed University	1.019
Employed	1.021

Results

The final model created is:

$$\log \frac{\hat{p}}{1-\hat{p}} = \beta_0 + \beta_1 x_{province} + \beta_2 x_{university} + \beta_3 x_{employed}$$

By generating a summary of the model, the following estimates for the coefficients were found:

Coefficient	Estimated value	Error	p-value
β_0	-0.53	0.09	2.21×10^{-8}
β_1	-0.95	0.12	8.13×10^{-16}
β_2	0.47	0.11	1.20×10^{-5}
β_3	-0.36	0.11	6.92×10^{-4}

From this table, we can see that all of the p-values for the predictor variables are very small (all less than 1%), therefore all of the predictor variables are statistically significant. Additionally, the errors on the estimations of the coefficients are relatively small, so we can assume that these variables provide a fairly reasonable fit to the survey data.

However, the accuracy of the model depends on the sample data. The survey data had many missing observations which greatly reduced the amount of usable data. Thus, the reduction in the number of sampled elements affects how accurately the sample represents the population. Additionally, in a report from Statistics Canada, it was found that in addition to province, education, and employment, other factors that may affect voting are age, family status, immigration status, and economic well-being (Uppal and LaRochelle-Côté,2012). Since these factors in the survey data were not found to be significant enough to be in the final model, this may also affect the accuracy of the model.

Using poststratification, it was found that the predicted probability that the Liberal Party would receive the majority vote is 0.28 or 28%. From the official results of the 43rd General Election, it was found that the Liberal Party received 33.1% of votes while the Conservative Party received 34.3% of votes (Elections Canada, 2019). Since the Liberal Party did not have the majority vote in 2019, it aligns with the predicted probability that was found from the 2019 census data.

Conclusions

In conclusion, the overall goal of our report was to predict the overall popular vote for the upcoming Canadian federal election (tentatively 2025). We had planned to achieve this objective with the help of a multiple logistic regression model with post-stratification.

We started this report by stating our research question and resulting hypotheses. Here, our research question was as follows: "Are province, university completion and employment factors that predict if a person will vote for the Liberal Party of Canada?". Furthermore, our null and alternative hypothesis were as follows:

$$H_0 : \beta_Province = \beta_UniversityCompletion = \beta_Employment = 0$$
$$H_0 : \beta_Province = \beta_UniversityCompletion = \beta_Employment \neq 0$$

Knowing this, we went on to describe in detail the data used in the report in the "Data" section, the methods utilised to build the model in the "Methods" section, and the final interpretations in the "Results" section. More specifically, we started the "Methods" section by conducting variable selection. Our final model was the result of BIC (Bayesian Information Criterion) stepwise selection. This process reduced our model from 7 variables to just 3 (province, university completion, and employment). Finally, we built our multiple logistic regression model which was as follows:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{province} + \beta_2 x_{university} + \beta_3 x_{employed}$$

Through the "Results" section, we provided a summary of our model which included estimates for the coefficients as well as their respective P-values. Due to the fact that the P-values for each variable is lower than the 5% significance level, we rejected the null hypothesis that "none of the predictive covariates have a statistically significant relationship with our dependent variable". Moreover, the errors for each of the coefficients was also small. Thus we can say that our model as a whole is statistically significant.

Finally, through the implementation of post-stratification, we found out that there was a 28% chance that the Liberal Party of Canada would get a majority of the vote. When comparing this result with that of the 2019 and 2021 federal election, we can see that our prediction is correct.

However, it is important to note that there are drawbacks to our model. The sample data upon which we built our model had a significant amount of missing data. We dealt with this issue by completely committing the rows which had data missing. Due to the fact that sample data is extremely important, this could have impacted the accuracy of our model.

Moving forward, future analyses/reports can improve on this model by utilizing sample data that has less missing values. In addition, due to the reproducible nature of this report, similar analysis can be done using AIC (Akaike information criterion) stepwise variable selection or by using P-values.

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. Zach. (2021, September 29). Understanding the null hypothesis for logistic regression. Statology. Retrieved December 1, 2022, from https://www.statology.org/null-hypothesis-of-logistic-regression/

5. Elections Canada. (2019) *Forty-Third General Election 2019: Official Voting Results.*https://www.elections.ca/res/rep/off/ovr2019app/51/table9E.html

6. Uppal, S., & LaRochelle-Côté, S. (2012, February 24) *Factors associated with voting.* Statistics Canada.https://www150.statcan.gc.ca/n1/pub/75-001-x/2012001/article/11629-eng.htm