# Forecasting the Election Results of Liberals vs. Conservatives in 2025

Group 67 - STA304 - Assignment 2

Fadi Maamarbachi (1004841380) Md Abrar Nasir (1005684127)
Mehreen Anhar Uzma (1005958215) Sumaita Imam Anika (1006666644)

December 01, 2022

## Introduction

Forecasting different phenomena have become increasingly popular with the advancement of technology and knowledge. Having an insight into not only the current happenings but projections of what is likely to happen, enables individuals to gauge the magnitude of the subject. Forecasting is also common in political and statistical analysis. Political events can take many different forms, including diplomatic choices, activities of political leaders, and other situations involving political institutions. Elections are a significant part of democracy and politics as it reflects the choice of the population. Election forecasting is a very popular form of statistical analysis used by political scientists. Mathematics, statistics, and data science are frequently used in political forecasting methodologies. Forecasting elections has a variety of uses including academic and public reasons (Lewis-Beck, 2005).

Election projections appeal to a primal human desire to peek into the future (Norpoth & Stegmaier, 2017). Humans have always been drawn to the idea of predicting election results with some degree of precision, since elections were created to choose leaders. It informs leaders and followers about a likely outcome, enabling them to make any required adjustments. It might also indicate a change in the policy goals of political activists. Additionally, any interested party can evaluate a campaign by keeping track of the election forecasts. Effective election forecasting models satisfy intellectual curiosity of the population. As citizens, the majority of us are interested in knowing who will win and by how much. In a functioning democracy, this knowledge is inherently fascinating. Since, election is the process through which leaders are chosen, it is an extremely significant matter for the population. This is because numerous financial, social and economical factors such as unemployment benefits, tax rates, interest rates, human rights, subsidies and more are dependent on the party elected and its leaders.

In the 2019 federal election, the Liberal party, led by Justin Trudeau, won the election to form a coalition government. With the second-highest number of seats in the parliament, the Conservatives became the official opposition. The 2019 election was ranked the second for the lowest vote share for a party to create a single-party minority administration. This shows the popularity of the party and its leadership qualities declined over the years.

### Terminology

This statistical analysis report includes numerous political and statistical terminologies. Some examples of statistical terminologies include population, sample, response variable, explanatory variable, hypothesis, mean, distribution, weight, regression, parameter, variance, survey, target population, non-response biased sample and histogram.

Population: is a group of elements about which we wish to make an inference (Jo-Caetano, 2022).

Sample: is a condensed, controllable representation of a larger group. It is a subgroup of people with traits from a wider population. When population sizes are too big for the test to include all potential participants or observations, samples are utilized in statistical testing.

Response variable: is a variable that researchers are trying to predict or explain.

Explanatory variable: is one that explains or predicts the change in the response variables and thus, it can be anything that affects the response variable.

Target population: population that the intervention is intended to study and take conclusions from.

Hypothesis: proposition tested using observations and statistical experiments.

Maximum likelihood estimate: It is a likelihood function that calculates the conditional probability of observing the data sample, given a probability distribution and distribution parameters.

Some political terms include candidate, federal and elector:

Candidate: is an individual running for public office and vying for a seat in the House of Representatives (Elections Canada).

Elector: is a Canadian citizen who is at least 18 years old and thus qualified to vote (Elections Canada).

Federal: a system of government or nation where authority is split between a central administration and many regional governments (provincial governments).

These terminologies are used numerous times throughout the report to analyze the data set and answer the research question accordingly.

### Data and Hypothesis

In this analysis, we use two distinct sets of data. We consider the General Social Survey (GSS) dataset as the 'census' data and the Canadian Election Study 2019 dataset as the 'survey' data. For the 2019 CES, there are 2 data files: an online survey and a phone survey. In this analysis, we are using the phone survey to build a model to predict which party the population is leaning towards.

The analysis in this paper aims to predict which factors influence the overall popular vote amongst the top two electoral parties in the next Canadian federal election to be held in 2025 using the statistical methodologies of multivariable logistics regression model and post-stratification. More specifically, this paper investigates the probability of Canadians voting for the Liberal Party in contrast to the Conservative party based on 3 specific combinations of demographic categories.

In the next section we introduce the dataset we used to perform the analysis to predict the election outcome and provide a numerical summary of the important variables, before we delineate in the third section, the statistical methods we used to build and assess the relevant forecasting model and explain the parameters of interest. The fourth section then outlines the results derived using the methodologies with interpretations of the figures and tables. The sixth section concludes the analysis with a brief recap of the hypotheses and comments on drawbacks regarding the model.

## Data

The General Social Survey was established in 1985 and every survey includes a key problem, exploratory questions, and a standard set of socio-demographic questions used for classification (Statistics Canada). The main aim of the GSS is to collect information on social patterns for the purpose of addressing specific social policy concerns. Datasets from GSS and scholarly articles issued by Statistics Canada are available to all concerned parties which they can use to study numerous phenomena (Statistics Canada). A complete statistical portrayal of Canada and its citizens based on their demographic, social, and economic aspects is provided by census data. The output of gathering responses from study participants is survey data. It serves as the foundation for making educated decisions in various circumstances and is a fair depiction of the viewpoints and perceptions of the target audience.

The GSS was established in 1985 and every GSS survey includes a core topic, focus or exploratory questions, and a standard set of socio-demographic questions used for classification (Statistics Canada). The primary objective of the General Social Survey is to collect data on social trends for the purpose of addressing certain

social policy challenges. All interested parties have access to GSS datasets and analytical articles issued by Statistics Canada and can use them to study numerous phenomena (Statistics Canada).

The main goals of the Canadian Election Study are to give a comprehensive account of the election, to highlight the key factors that influence people's voting decisions, to show what changes and remains constant throughout the campaign and from one election to the next, and to draw attention to the similarities and differences between voting and elections in Canada and other democracies. The second goal is to advance scientific understanding of voter motives and the significance of elections and election campaigns in democratic nations. The third mandate of this election study is to gather information about Canadians' attitudes and beliefs regarding a wide range of social, economic, and political issues. This information was then made available to researchers in the fields of political science, sociology, economics, communications, and journalism (Stephenson et al, 2020).

In order to learn more about Canadians' attitudes and beliefs before, during, and after the 2019 federal election, the 2019 Canadian Election Study was carried out. It carries on the tradition of Canadian Election Studies, which was established in 1965.

The outcome information gathered from a sample of survey respondents is known as survey data. This information is thorough data that was acquired from a target audience regarding a particular subject to perform research. The methods for gathering survey data and doing statistical analysis are numerous. Since 2013, GSS data have been gathered by combining telephone interviews and self-completed online questionnaires (Statistics Canada). In order to lessen the burden on respondents and increase data accuracy, some data, most frequently income statistics, are derived from tax or other administrative files as opposed to being directly collected through surveys. On the other hand, the online sample for the 2019 Canadian Election Study was composed of a two-wave panel with a modified rolling-cross section during the campaign period and a post-election recontact wave (Stephenson et al, 2020). The data collection process of the CES data we are analyzing in this paper is also telephonic. Traditional element sampling methods, like stratified random sampling and systematic selection, are used for telephone sampling for list frames. The data from these two surveys were then analyzed in Rstudio, a programming language for statistical analysis and graphics.

Before the cleaning process, the GSS dataset contained 20602 observations and the CES 2019 dataset contained 4021 observations of 278 variables. However, some rows had observations set to N/A because they lacked sufficient information. To keep the flow of the observations consistent, the rows of observations that were set to N/A were removed from the datasets. Additionally, we compared the GSS and the CES datasets and cleaned them to only include the questions that overlap to help make this analysis feasible using statistical tools. We found that most of the overlapping questions were demographic based questions, and these important variables are described in the Data section of the paper. After cleaning the datasets and selecting the common variables, the GSS (census) dataset contained 17034 observations of 8 variables and the CES 2019 (survey dataset) contained 1027 observations of 9 variables. Additionally, the variables in both the datasets were adjusted to be in sync and of the same datatype where each numeric was referred to the same identifier for both datasets.

**Description of the important variables:**

The response variable in our analysis is 'party to vote for' which represents the party the individuals in the sample would most likely vote for. The question for this variable was structured as "Which party do you think you will vote for?" This variable will help directly answer the research question as to which party is likely to get the most votes.

The common predictor variables in both the datasets are sex, age, province they currently live in, marital status, education attainment, whether or not they have any religious affiliations, the language they speak, and their household income which are explained in detail in Table 1.

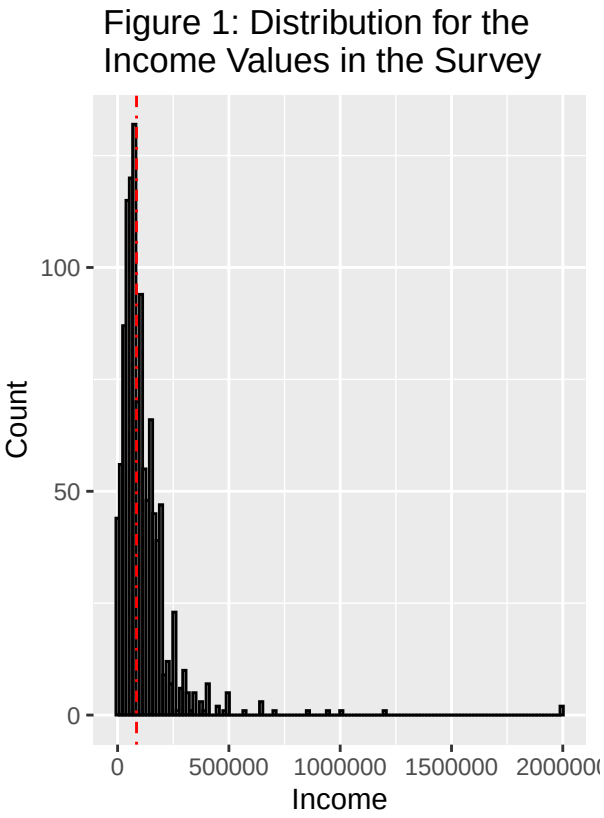
**Table 1** explains the important variables of the analysis:

| Variable | Description |
|---|---|
| Age | It refers to the period of human life measured by years since birth characterized by certain mental and physical developments with legal responsibilities. |
| Province | The country of Canada is divided into regions termed as provinces for administrative purposes. |
| Marital status | It refers to the distinct options that show a person's relationship with his/her significant other. |
| Education | It refers to the highest level of education or achievement of the learning objectives of a certain level that an individual has achieved successfully through validated assessment of acquired knowledge, skills and competencies. |
| Religious affiliations | It refers to the religious and spiritual beliefs and practices a person adheres by. |
| Language | It is the main medium of human communication that consists of words and is expressed through speech, writing, and gesture. |
| Household income | It is the combined gross cash income of all members of a household who are above 15 years of age. It is used to evaluate the economic conditions and compare living conditions between geographic regions that is often dependent on government policies. |
| Sex | It refers to the two main categories of human which are male and female, based on their reproductive functions. |

**Table 2** provides the numerical summary of the age and household income of the individuals who took part in the survey. It shows the center, spread and the maximum-minimum statistics of the sample data:

| | Min | 1st Quartile | Median | Mean | Std. Deviation | 3rd Quartile | Max |
|---|---|---|---|---|---|---|---|
| Age | 18.00 | 38.00 | 51.00 | 50.89 | 16.84 | 64.00 | 100.00 |
| Income | 0 | 50000 | 85000 | 114261 | 130298.5 | 150000 | 2000000 |

The table above shows the summary of the important numerical variable: Age. Specific features of the variables in the dataset are described in numerical summaries, such as the mean, spread, and maximum-minimum statistics, as well as the center or median of the distribution.



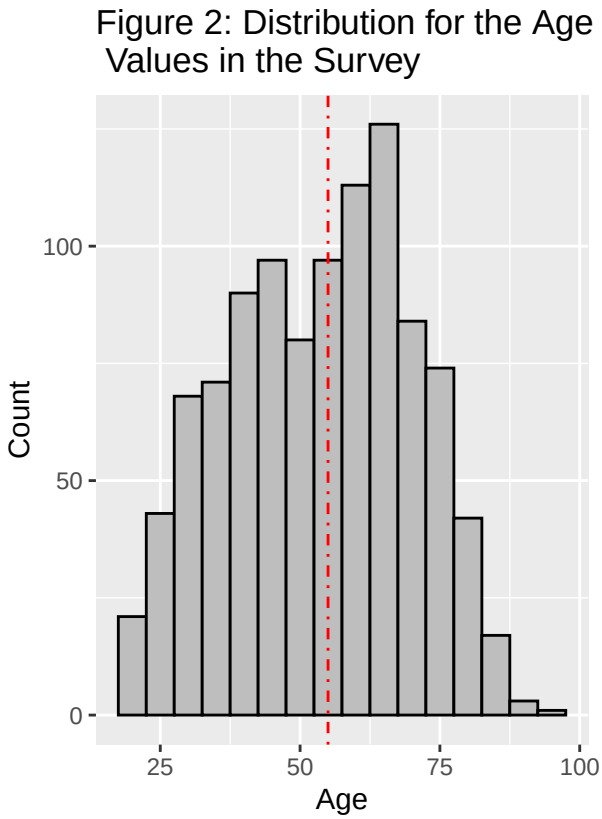




**Figure 1: Histogram of income**

The illustration above shows the histogram of the variable "Household Income". The diagram appears to be highly positively (right) skewed, showing an inequitable distribution of income across Canadian households.

**Figure 2: Histogram of age**

The occurrence of the normal distribution can be loosely classified into three categories: exactly normal distributions, approximately normal distributions, and distributions modeled as normal (lumen Boundless Statistics). We can comment whether the distribution of our dataset is normal just by looking at the histogram above. The illustration above shows the histogram of the variable "Age". The histogram appears to be almost symmetric and does not look skewed. The graph shows that the distribution of the data is approximately bell-shaped but with a slight skew on the left.

## Methods

Regression is a predictive modeling technique that is used to find the relation between a response variable and one or more predictor variables. When two or more predictor variables are used to predict the response variable, it is termed as multivariable regression. Regression analysis can be broadly classified into two categories: (1) linear regression and (2) logistic regression. Linear regression is used to establish a linear relation between the response and dependent variables. On the other hand, logistics regression is used to calculate or predict the probability of a binary outcome from a set of predictor variables. Binary outcome refers to a scenario where only two outcomes are possible: the event happens or the event does not happen. For the purpose of analysis using statistical softwares, the possible outcome of the event occurring is assigned a value of 1 and the event not occurring is assigned a value of 0 (Thanda, 2022). For our analysis, our response variable is trying to predict which specific political party will gain more votes in the federal elections of 2025. We will be dealing with binary response data and therefore, logistic regression will fit our purpose the best.

In reality, it might not always be possible to collect a sample that is a true representative of the target

population that we are interested in measuring. In such situations, a large sample is collected with many different types of demographic data on the individuals in the sample. This will enable us to partition the data into different demographic cells and estimate the response variable for each cell. We will then extrapolate the cell-level estimates of the response variables to population-level by weighing each cell by its relative population proportion. In doing so, post-stratification gives more accurate or precise inference on the population-level data based on a poorly representative or non-response biased sample. In our analysis, we will be estimating the response variable of each cell using multivariable logistic regression and then use the GSS data since it is a good representative of demographic data of the entire population to extrapolate the cell-level estimates as per relative population proportion of that specific cell. (Jo-Caetano, 2022).

**Model Specifics**

There are numerous statistical tools that can be used to select the best predictors to use in our logistic regression model. In our analysis, we first built a model using all the 7 overlapping variables between the GSS and the CES dataset. At this step, we factor all the categorical response types to get columns of binary values to support the mechanism of our model, given the nature of our dataset. We then reduced the model by comparing the significant p-values based on t-test to select predictors that best explain variations in the response variable and are most influential in explaining the logistic regression relationship. We chose 4 predictors with the smallest p-values which are the identifiers if the respondent was from Saskatchewan or Alberta, if they were Female-identifying and if their household income was between $75,000 and $99,999 per annum. In other words, these 4 predictors are the most significant variables to explain the logistic relationship. Using the 4 selected predictors, we built a reduced logistic model and to validate this regression model, we used a few statistical tools.

When building a logistic regression model, there are a few assumptions that must be taken into account.

1. Absence of multicollinearity: It is necessary to check if our model satisfies the assumptions of a multivariable logistic model. Non-collinearity between the independent variables is an assumption that is measured using variance inflation factors (VIF). When there are many highly associated independent variables in the data, this is referred to as multicollinearity. This is a concern since it lessens the accuracy of the calculated coefficients, which diminishes the logistic regression model's statistical potency. A set of independent variables' multicollinearity is gauged by the variance inflation factor (VIF). VIF can have a minimum value of 1 only which would indicate absence of collinearity whereas a VIF number greater than 5 denotes multicollinearity that exists and can affect the accuracy of our data.

2. Appropriate Outcome Type: As logistic regression often serves as a classifier, the response variable in the dataset must match the kind of logistic regression being utilized (binary, multinomial, or ordinal). Logistic regression by default posits a binary outcome variable with a two-outcome maximum.

3. Linearity in logit for continuous variables: The relation between each continuous independent variable and the logit (also known as log-odds) of the outcome must be linear in order for logistic regression to work.

4. Lack of strong influential and outlier observations: Since outliers and influential observations might affect model's accuracy, logistic regression makes the assumption that there are no extreme outlier and highly influential data points. An outlier is a point that is different from other observations while influential data points affect the skewness of your model.

To further compare and establish that our reduced model is indeed a better model to predict the response to our research question, we used Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and McFadden's Pseudo $R^2$ value. BIC is a statistical tool for assessing how well a model fits the dataset. It helps us in choosing the predictor variables that best describe the changes in the response variable. The smaller the value of BIC, the better is our model in representing the changes in the response variable (Shelton, 2019). Another goodness of fit for logistic regression models is McFadden Adjusted $R^2$. McFadden's Pseudo $R^2$ ranging from 0.2 to 0.4 indicates a good model fit (Wikipedia Foundation, 2022). Furthermore, AIC is another different measure that explains how well a model fits that data. The calculation of AIC uses the number of parameters and maximum likelihood estimate. There is no particular threshold for AIC values. So,

the model with the lowest AIC score is the better one as it has the least number of parameters explaining the response variables the most.(Bevans, 2020)

$$log(\frac{\hat{p}}{1-\hat{p}}) = -0.09978 - 1.67467x_{Saskatchewan} - 2.14980x_{Alberta} + 0.66626x_{female} + 0.41586x_{income75000-99999}$$

$\hat{p}$ is the probability of Canadians voting for Liberal Party and the coefficients in the equation represent the average change in log odds for every one one unit change in the independent variables.

This paragraph explains the logit euqation above. The y-intercept, $\beta_0$ of our model is -0.09978. Y-intercept represents the log probability of our designated event when all predictor variables are 0. If the Canadian citizen is currently residing in Saskatchewan, this will decrease the log odds by 1.67467 units when everything else is constant. If the individual is currently residing in Alberta, this will decrease the log odds by 2.14980 units when everything else is constant. Furthermore, if the individual is a female, it will increase the log odds by 0.66626 units when everything else is constant. Finally, if the individual's household income is between CAD75000 to CAD99999, it will increase the log odds by 0.41586 units when everything else is constant.

## Post-Stratification

For the post-stratification process, we used the GSS (census) data and only include the variables from our reduced model, sex, province and household income. However, to accurately map survey and census information, we create a dummy census dataset to proceed with post-stratification. This dummy dataset, again, includes with the identifiers if the census respondent was from Saskatchewan or Alberta, if they were Female-identifying and if their household income was between CAD75,000 and CAD99,999 per annum. For every combination of demographics based on the mentioned variables, we calculate the proportion of each category. We found 3 unique bins for these variables which we will use to decode to what extent specific demographic classes of the population are most likely to vote for the liberals in contrast to the conservatives.

Now, to predict the $\hat{y}^{PS}$ values which is our parameter of interest, we use the probability of each bin to vote for the Liberal Party in contrast to the Conservatives. Using this estimate and the proportion of each bin we found earlier, we calculate the respective Yps (probability) values by finding the product of the estimate and the proportion.

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

where $N_j$ is the count of the occurrences in a bin based on the combination of demographics [the individual is a female living in Saskatchwan or Alberta with household income between CAD75000 to CAD99999]. $\hat{y}$ is the probability of the Liberals gaining the vote from each bin mentioned above.

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

## Results

**Table 3:**

| | Saskatchewan | Alberta | Female | Income (CAD75000-CAD99999) |
|---|---|---|---|---|
| VIF (Reduced model) | 1.008583 | 1.004308 | 1.006372 | 1.001157 |

Table 3 shows the VIF values of the predictors of the reduced model. All the VIF values are observed to be less than the threshold value of 5. This means that the predictors in the model are not significantly collinear which otherwise would have reduced the accuracy of the model.

**Table 4:**

| Model | AIC | BIC | McFadden pseudo R-squared |
|---|---|---|---|
| Full model | 1233.8 | 1406.533 | 0.1815105 |
| Reduced model | 1302.4 | 1327.046 | 0.09110833 |

Table 4 above shows the goodness-of-fit measures of the logistic regression model. As we can see that both the BIC and McFadden pseudo $R^2$ values have decreased in our reduced model relative to the full model. Therefore, the reduced model is a better representative of the logistic regression relationship.

**Table 5** showing the count and the proportion relative to the entire census population and also provides information on their associated probabilities of voting for the Liberals:

| Saskatchewan | Alberta | Female | Income (CAD75,000-CAD99,9999) | Proportion | Estimate | $\hat{y}^{PS}$ |
|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 40.44% | 56.65% | 22.91% |
| 0 | 0 | 0 | 1 | 5.65% | 31.61% | 1.79% |
| 0 | 0 | 1 | 1 | 6.75% | 98.23% | 6.63% |

Referring to the *Table 5* above, we can conclude the the following statements: 1. For every female individual not living in Alberta or Saskatchewan with a household income off the range CAD75,000 to CAD99,999, the probability of them voting for the Liberals in contrast to the Conservatives is 22.91%. 2. For every male individual not living in Alberta or Saskatchewan with a household income in the range CAD75,000 to CAD99,999, the probability of them voting for the Liberals in contrast to the Conservatives is 1.79%. 3. For every female individual not living in Alberta or Saskatchewan with a household income in the range CAD75,000 to CAD99,999, the probability of them voting for the Liberals in contrast to the Conservatives is 6.63%.

We have verified one of the assumptions of non-multicollinearity between predictors using variance inflation factors (VIF). Additionally, since the outcome variable in our research model has only two outcomes, 0 and 1, where 1 refers to the Liberals receiving the voting preference and 0 refers to the Conservatives getting the vote, assumption 2 described above is also verified.

## Conclusions

In this paper, we built a logistic regression model to predict which party, Liberal or Conservative, Canadians have a greater probability of voting for. We used a logistic regression since the outcome variable is categorical. From the survey sample, we can see that the Conservative Party is the most popular as people have identified the party as their preferred one. However, from this paper, we can conclude that the likelihood of the Liberals winning votes in the 2025 election depends on the combination of demographic categories which are that if they are females, if they are living in Alberta or Saskatchewan or if their household income is between CAD75,000 to CAD99,999.

There are 3 main results of this paper. For every female individual not living in Alberta or Saskatchewan with a household income off the range CAD75,000 to CAD99,999, the probability of them voting for the Liberals in contrast to the Conservatives is 22.91%.For every male individual not living in Alberta or Saskatchewan with a household income in the range CAD75,000 to CAD99,999, the probability of them voting for the Liberals in contrast to the Conservatives is 1.79%. For every female individual not living in Alberta or Saskatchewan with a household income in the range CAD75,000 to CAD99,999, the probability of them voting for the Liberals in contrast to the Conservatives is 6.63%.

Forecasting election results can involve different methods of statistical analysis. It can be analyzed from numerous angles and perspectives. This paper only focuses on the two main electoral parties of Canada since

these two parties garner the most attention of the voters based on significant demographic categories. The CES survey was structured in a way that was not analysis-friendly since numerous questions were repeated and heavily worded. A decrease in goodness-of-fit test measures such as BIC and McFadden's adjusted $R^2$ reflects that the new model is a better fit for our dataset. However, another drawback of the paper is that the model built in this analysis is that the AIC value of the reduced model has increased.

This is an issue that should be further analyzed for the sake of accuracy of the analysis. We should also create a well-rounded model that includes all the involved electoral parties in the election.

## 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. Norpoth, Helmut & Stegmaier, Mary. (2017). Election Forecasting. Oxford Bibliographies
5. Grolemund, G. (2014, July 16) Introduction to R Markdown. RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: May 5, 2021)
6. Jo-Caetano, Samantha (2022), STA304 Lecture Slides. Elections Canada
7. Lewis-Beck, Michael S. (2005). Election Forecasting: Principles and Practice. The British Journal of Politics and International Relations
8. Elections Canada. Glossary. https://www.elections.ca/content.aspx?section=res&dir=glo&document=index&lang=e
9. Thanda, Anamika. (2022). What is Logistic Regression? A Beginner's Guide. Careerfoundry.
10. Statistics Canada. https://www.statcan.gc.ca/en/start
11. Shelton, Nick. (2019). What is BIC, how do you use it and what is a good BIC value? https://community.jmp.com/t5/Learning-Center/What-is-BIC-how-do-you-use-it-and-what-is-a-good-BIC-value/ta-p/194416
12. Bevans, Rebecca. (2020). Akaike Information Criterion | When & How to Use It (Example).
13. Bartlett, Jonny. (2018). Probability concepts explained: Maximum likelihood estimation. https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1
14. Stephenson, Laura B. et al. (2020). Canadian Election Study, 2019, Phone Survey.