Predicting Voting Behavior in Canadian Elections

Mohammad Ettouhami Sanjam Sigdel Ziyue Yang

May 28th, 2021

Introduction

Voting is not only one of the pillars of democracy, but also the pillar of a well functioning society. Exercising our right to vote means electing official representatives whose interests and concerns align with our own. As a result, our political participation can have a huge impact on the policies, laws and regulations that are put into place. Not using our vote runs the risk of giving power to elected members who will implement policies that do not represent us. Keeping this in mind, it becomes extremely important that political parties understand 'psephology'. Coined by William Francis Hardie to describe the statistical study of elections and trends in voting, psephology provides insight into why certain groups of people tend to vote for specific political parties over others (Merriam-Webster, n.d.; Wikipedia contributors, 2021).

A few socio-demographic factors have been identified to have an influence on voting behaviour (Anderson & Stephenson, 2011). As such, the main objective of this report is to determine if factors such as age, gender and household income can be used to predict whether Canadians will vote for the Liberal or Conservative political parties in the upcoming 2023 federal elections. We hypothesize that age, gender and household income will be predictive of which party someone will vote for. Our reasons for only focusing on the Liberal and Conservative parties are that they are the two largest political parties in Canada with the Liberals forming the current government and the Conservatives being their official opposition (Political Parties, 2020). They are also the major vote holders and their values and mission are the most distinct from one another (Political Parties, 2020). While the relationship between socio-demographic factors and voting behaviour is not that simple, we hope that the distinction between parties would also translate to providing a clearer image of the demographics both parties' attract. For our report we will be using a logistic regression model with post stratification to analyze the General Social Survey (GSS) and Canadian Election Study (CES) survey data (Statistics Canada, 2020; Stephenson et al., 2020).

Established in 1985, the GSS is a collection of cross-sectional data that serves to measure the trends in society relevant to living conditions and subjective well being (Statistics Canada, 2013). It also provides data that is then used to inform current or future social policies. The GSS is very useful in answering questions that center around the demographics and socio-economic conditions of a family. The CES is a large-scale online survey that is conducted each election year. The main objective of it is to provide insight on voting behaviour (Stephenson et al., 2020). More specifically, why people do or don't decide to vote at all and if they choose to vote what pushes them to give their vote to one political party over the other. Together with the GSS these datasets provide us with the necessary information we need to determine if age, gender and household income can predict what political party Canadians will vote for.

In this report, we will start by introducing the data and highlighting its key characteristics. Next, we will walk through our statistical models and give an explanation of their underlying concepts. After presenting our methodology, we will discuss the findings of our analysis and end with a discussion drawn from our results. Our analysis will not only contribute to our understanding of election and voting trends in Canada but will also provide political parties with information that will allow them to implement policies and select representatives that would appeal to the demographics that their party attracts.

Data

Introducing the Data

The 2019 CES survey was conducted to observe the opinions of the Canadian population during and after the 2019 Canadian Election. The data that we are interested in is the CES phone survey conducted in a one month period during which data was gathered from 66% wireless telephone numbers and 34% landline telephone numbers (Stephenson, 2020). The target population of this survey is the entire adult population of Canada, since the objective of the survey is to understand the population demographics, motivations and voting behaviors. The frame population in this survey are the respondents who chose to complete the survey by phone, and the sample population selection was done by selecting random Canadian landlines, and wireless phone numbers from phone books. It is important to note that the respondent selected for the landline survey was the eligible voter with the closest birthday, in the case of there being multiple eligible respindents in a single household (Stephenson, 2020).

According to the GSS overview provided by Statistics Canada (2013), the GSS is a collection of sample surveys that were conducted via telephone using a cross-sectional design. These surveys were administered across the ten Canadian provinces to all non-institutionalized individuals ages 15 years and up. The sampling frame used for this survey was a combination of cellular and landline phone numbers, along with Statistics Canada's Address Register. The primary objective of the GSS was to measure the trends in society relevant to living conditions and wellbeing along with providing data that would be used to inform current or future social policies.

In this section, we will walk through the cleaning process, go over the important variables and take a look at any interesting summaries that we can derive from our data. In later sections, we will discuss our model and the results from our analysis.

Cleaning the Data

CES Data

Before we build a model on the CES data, we must first clean it. The following three variables required similar cleaning:

- `q69`: Could you please tell me your total household income before taxes?
- `q11`: Which party will you likely to vote for?
- `q3`: With which gender do you identify?

We first renamed these three variables to *Household Income*, *Party Vote* and *Gender*, respectively. For all three variables, respondents could select one of the following three options: "Don't know", "Refused" or "Skipped". If respondents selected one of these options, they were given a negative value as their response. Our code removes these negative values by filtering the columns and removing rows where the value is less than zero in the column. Additionally, since our focus is primarily on two parties, Liberal and Conservative, we needed to remove any values that represented the other political parties that we were not interested in. The values representing the other parties ranged from 3 to 7 in the *Party Vote* variable. We filtered out and removed any rows containing these values. The *Party Vote* variable also had observations whose values ranged between 8 to 10, representing the following three options: "Will not vote", "None of these" or "Will spoil ballot". We removed any rows containing these values as we were not sure of the respondent's voting intentions.

In order to analyze our *Household Income* variable, we decided to create a new categorical variable called *Income Category*, where we categorized respondent's household income into six income categories. We selected the categories based on a similar variable found in the GSS data set, *Income Family*. There were some missing data points across most of our variables which meant that we had to filter out any rows that contained them. We also created a new column based on *Party Vote* which indicated 1 if the respondent selected the Liberal party as their answer, or 0 if they voted for Conservative. This new variable was called *Liberal Vote*.

GSS Data

Before we apply our model on the GSS data, it must be cleaned to fit our model parameters. We grabbed the following three variables:

- age
- sex
- income_family

In order for our model to work, we needed to have the same variable names in both sets of data, so we renamed *sex* to *Gender*. The minimum value for the *age* variable in the CES data set was 18, since that is the minimum legal age to vote. The GSS data has respondents younger than 18, so we filtered and kept only rows where respondents indicated that they were 18 years of age or older.

During both cleaning processes, we converted any categorical variables into a factor so that R treats them as categories instead of actual numerical or character values. *Table 1* provides an overview of how some of our CES data looks like after cleaning, while *Table 2* shows our cleaned GSS data:

Table 1: First three CES data observations

Sample ID	Age	Gender	Household Income	Household Income Category	Voted Liberal
39	25	Male	20000	Less than $25,000	Yes
165	56	Male	35000	$25,000 to $49,999	No
329	41	Male	100000	$75,000 to $99,999	Yes

Table 2: First three GSS data observations

Sample ID	Age	Gender	Household Income Category
1	53	Female	$25,000 to $49,999
2	51	Male	$75,000 to $99,999
3	64	Female	$75,000 to $99,999

Important Variables

CES - Survey Data

In order to explore our research question and build a model, we must first try to understand all of our important variables: the independent variables (*Age*, *Gender*, *Household Income* and *Household Income Category*) and the dependent variable (*Party Vote*). *Party Vote* is a categorical variable identifying two Canadian political parties: Liberal and Conservative. Out of 1502 survey respondents, 731 indicated that they will vote Liberal while 771 respondents indicated that they will vote Conservative.

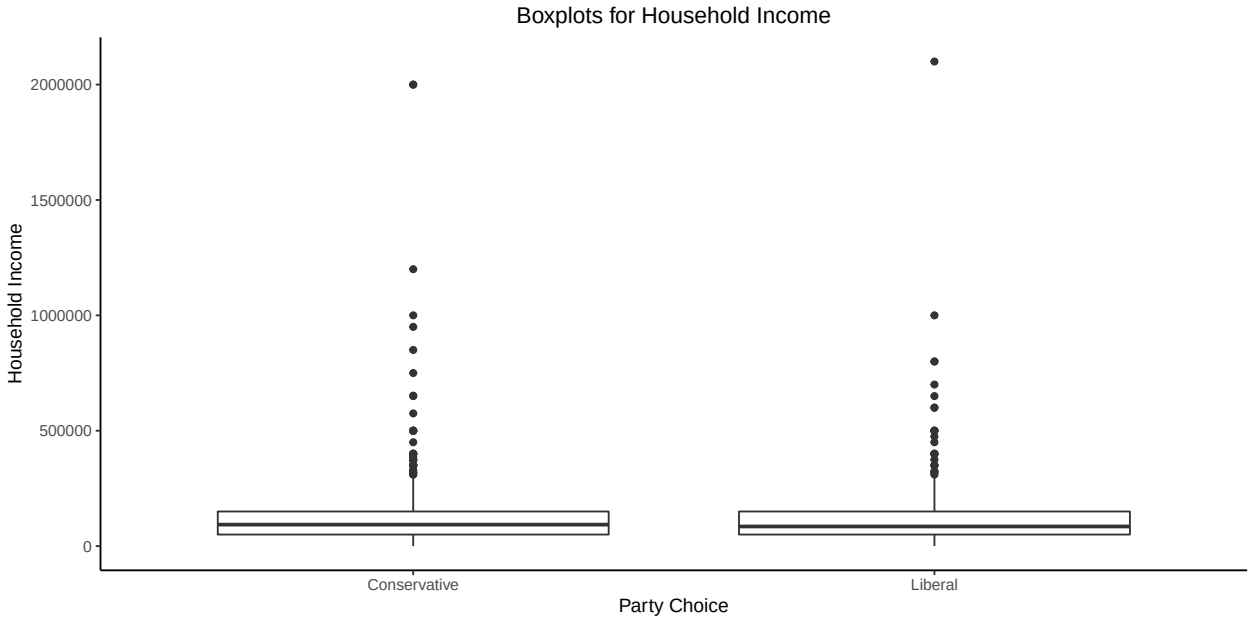
Household Income is a quantitative variable representing household income (in Canadian dollars) as numerical discrete values. A brief summary statistic for our *Household Income* variable can be seen in *Table 3*. It shows that the distribution has a mean of 117276.3 and a standard deviation of 137639.95.

Table 3: Summary statistics for the *Household Income* variable

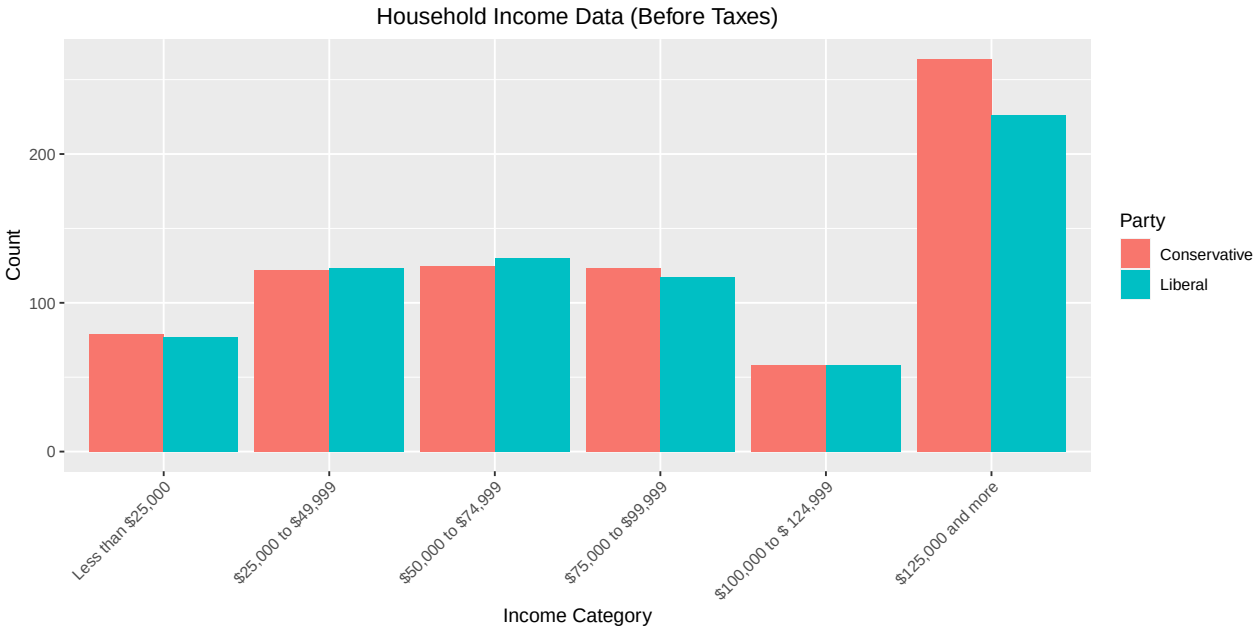
Summary Statistics	Value
Mean	117276.3
Median	90000
Standard Deviation	137639.95
Minimum	0
Maximum	2100000

Summary Statistics	Value
First Quartile	50000
Third Quartile	150000

Boxplots can be used to get a better idea of the difference between different groups. The following two boxplots display the distribution of the *Household Income* variable for both political parties.

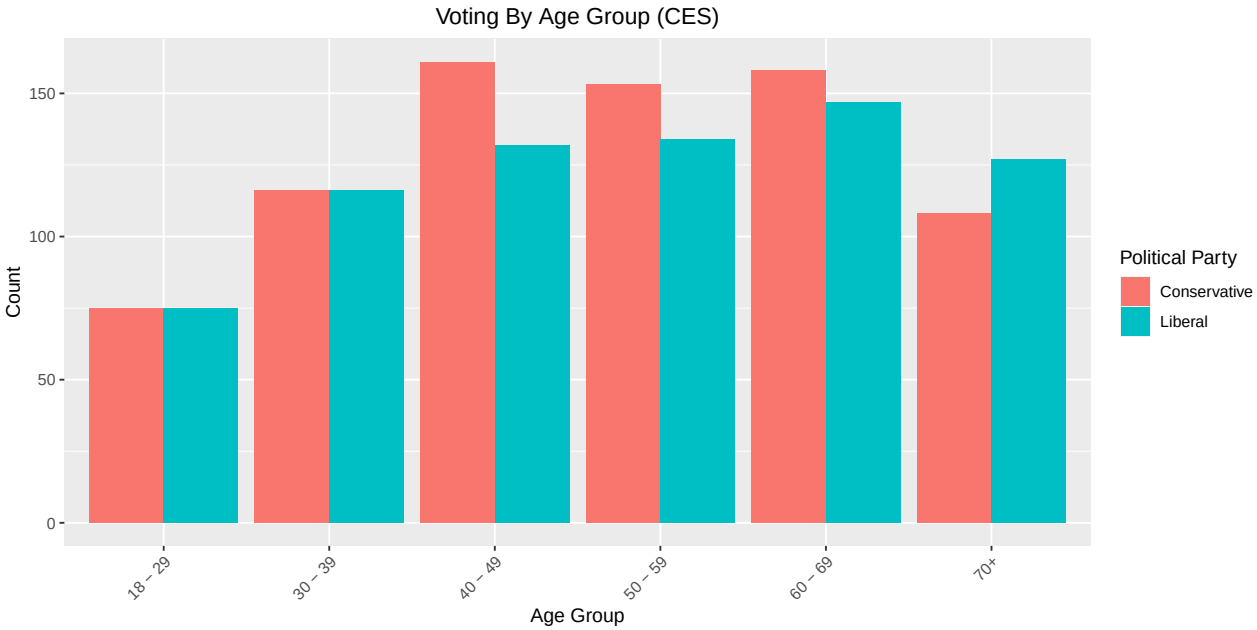


For both groups, we can immediately notice that there are quite a few outliers. We also realized that the GSS data does not have a variable which contains household income as numerical discrete values, which is necessary if we later want to apply our model on GSS data. However, the GSS data does have a variable that contains household income as a category. We made the decision of creating a new variable, *Household Income Category*, in the CES data set. It is a categorical variable identifying which of the six possible categories the respondent's household income falls under, using data from the *Household Income* variable. We can visualize our *Household Income Category* variable by showing the six categories separately for both *Party Vote* groups, Liberal and Conservative. The following double bar graph shows the frequency distribution:



This barplot gives us more insight into the voting behavior of each income group. It seems that people with a household income considered to be "middle class" are split evenly between the Conservatives and Liberals. We can see the Liberals get a little advantage in the $50,000 to $74,999 income group, but they lose this bump to the Conservatives in the $75,000 to $99,999 income group. This data shows a very close race between Liberals and Conservatives, until we get to the highest income group in our data, $125,000 and more. We can see that in this household income category, there are significantly more respondents (38) who voted for Conservative than Liberal.

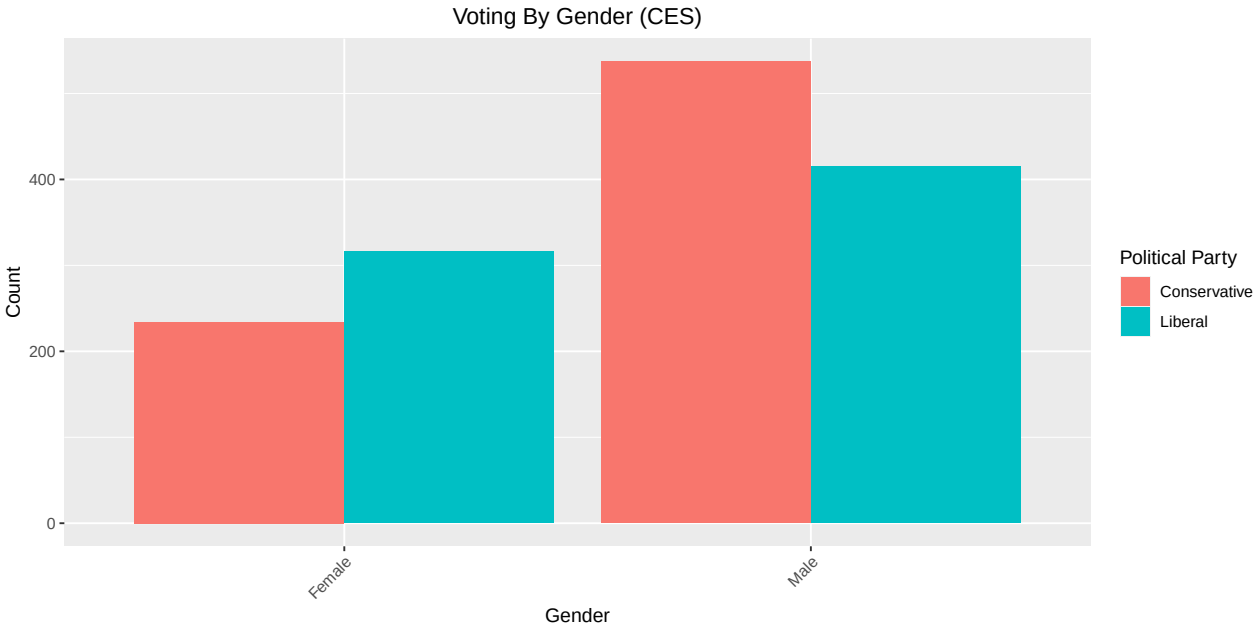
Our last important variables from the CES data are *Age* and *Gender*. *Age* is a quantitative variable representing the respondents age as numerical discrete values. The mean age is 51.99. The following double bar graph shows voting by age:



For the sake of visualization purposes, we converted *Age* into a categorical variable to see if we could gain any valuable insights before building our model. The two youngest age groups have an even split between

voting for Liberal and Conservative. The next three older age groups have more respondents that vote for Conservatives than Liberal. The last age group, 70+, had more respondents voting for Liberal than the other party.

Gender is a categorical variable, representing the respondents gender as either "Male" or "Female". There are 952 male respondents and 550 female respondents in our CES cleaned data set. The following double bar graph shows voting by gender. We can see that females vote more for the Liberal party while males vote more for the Conservative party. This may indicate that *Gender* may be a relevant variable in determining the probability of *Party Vote*. This has to be confirmed by a more detailed statistical analysis.



GSS - Census Data

In the GSS data, we consider three variables to be important. Similar to the CES data set, *Age* is a quantitative variable representing the respondents age as numerical discrete values. The mean age is 52.8. *Gender* is a categorical variable, representing the respondents gender as either "Male" or "Female". There are 9220 male respondents and 11033 female respondents in our GSS cleaned data set. Lastly, *Household Income Category* is a categorical variable representing which of the six categories the respondents household income falls in. *Table 4* shows the frequency of each category:

Table 4: GSS: Household Income Category variable frequency

Household Income Category	Count
Less than $25,000	2725
$25,000 to $49,999	4301
$50,000 to $74,999	3655
$75,000 to $99,999	2888
$100,000 to $ 124,999	2120
$125,000 and more	4564

Methodology

In this section we will introduce our methodology that we will be using to analyze the data. We will provide an explanation of our model and the poststratification process. We will also discuss why our model is appropriate

for this analysis and any model assumptions, and follow with a discussion of the parameters of interest. The goal of this analysis is to predict the overall popular vote of the next Canadian federal election using a model with post-stratification. We will set up our null hypotheses based on our first variables of interest, *Gender*:

$$H_0 : \beta_{Gender} = 0$$

The second hypothesis concern the *Age* variable:

$$H_0 : \beta_{Age} = 0$$

The third hypothesis concern the *Household Income Category* variable:

$$H_0 : \beta_{HouseholdIncome} = 0$$

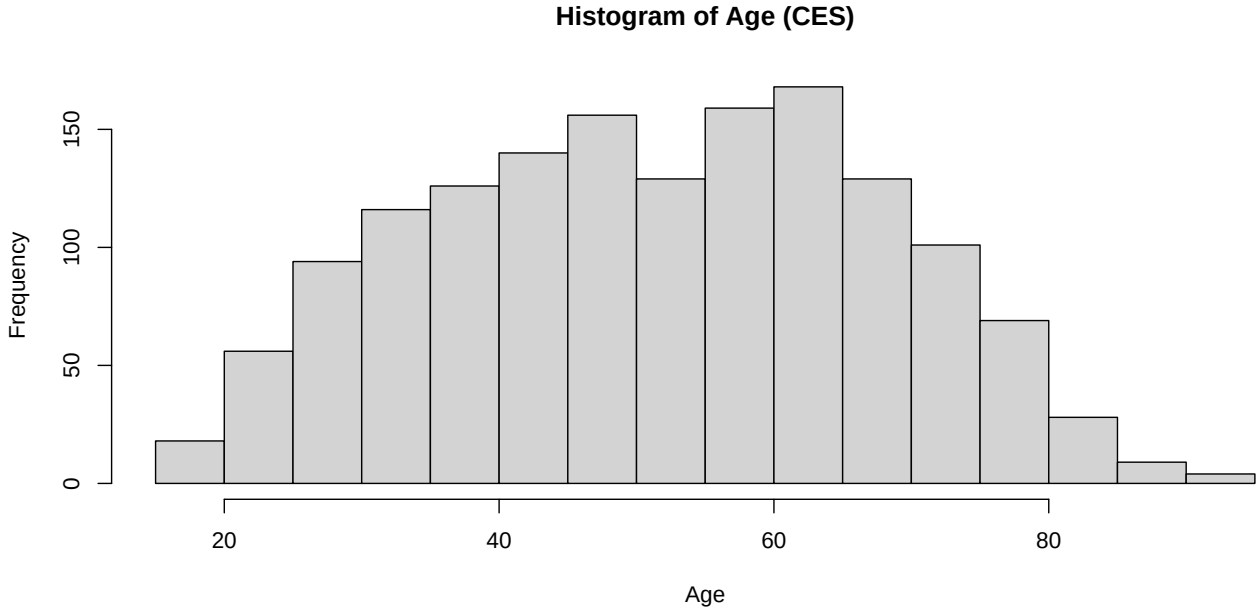
Model Assumptions

We will construct a regression model on the sample CES data to help us determine whether there is a relationship between our variables of interest and to predict GSS respondents voting outcomes. We must pick the regression model that is most appropriate for our data. Since our dependent variable, *Party Vote*, is a binary outcome (Liberal or Conservative), we will be using logistic regression.

In order to build a valid logistic regression model, the data should meet the following conditions (Kassambra, 2018):

- The dependent variable is binary
- The data values are independent
- There are no extreme values (outliers) in the continuous independent variables
- There is a linear relationship between the logit of the dependent and each independent variables
- There is no high multicollinearity among the independent variables

We know that the dependent variable is a binary outcome. The observations are independent since the respondents age, gender and household income in the CES data are not dependent on another respondent. The following histogram shows the data for our continuous independent variable, *Age*:



We can see that there are no extreme outliers. The maximum age is 94 and the minimum age is 18. We will assume that the final two assumptions are met, since their proof is outside the scope of this report. Based on the above observations, we can conclude that logistic regression is an appropriate regression model.

Model Specifics

In this section we will provide a simplified explanation of logistic regression. This is the statistical model that we will use to find the relationship between the variables of interest and the probability p of voting Liberal.

Consider a model with n independent variables, $x_1, x_2, \ldots, x_n$. We denote our dependent variable, *Party Choice*, as Y. Logistic regression assumes a linear relationship between the dependent variable and the log-odds, which can be written in the following form:

$$\ell = \log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Solving for p, we obtain:

$$p = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n}}{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n}+1} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}}$$

This can be written using the sigmoid function:

$$f(x) = \frac{1}{1+e^{-x}}$$

In the form:

$$p = f(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)$$

A plot of $f(x)$ shows that the sigmoid function varies between 0 and 1 which implies that p can be interpreted as a probability.

This method allows us to train a model based on the survey data. Luckily, R has a built in function to help us build our logistic regression model, *glm()*. Once we create our model, we will apply it on the census data so that we can predict whether respondents will vote Liberal or Conservative.

Post-Stratification

Poststratification is a technique used to reduce differences between sample and target populations, by combining population distributions of variables into survey estimates (Little, 1993). In this section, we classify the survey samples by household groups into various cells to match the income groups in the CES data, then weigh individual samples in each cell to improve the accuracy of survey estimates.

Aiming to transform the CES survey data (Stephenson el al., 2020) into accurate estimates of voter intent in the federal election, we use the household income information provided by respondents, and *poststratify* the survey responses to mimic a representation of similar voters. Following the poststratification process provided by W. Wang et al. (2014), we partition the population into cells based on various household income categories, and use the sample to estimate the response variable within each cell. We then aggregate the household cells' estimates up to the Census data estimate by weighting each household income cell by the relative proportion in the population. Denoting y as our outcome of interest, the poststratification estimate is defined by

$$\hat{y}^{\mathrm{PS}} = \frac{\sum_{j=1}^{J} N_j \hat{y}_j}{\sum_{j=1}^{J} N_j} \tag{1}$$

where $\hat{y}_j$ is the estimate of y in cell j, N_j is the size of jth household income cell in the population.

We generated cell-level estimates via a logistic regression model, and combine a poststratification strategy, namely logistic regression and poststratification.

All analysis for this report was programmed using `R version 4.0.5`.

Results

In this section, we will present the results of our model and poststratification and discuss our interpretation.

Model

After building a logistic regression model using the *glm()* function, we summarize the results in the following table:

Table 5: Logistic regression model results

term	estimate	std.error	statistic	p.value
(Intercept)	0.3431326	0.2599649	1.3199191	0.1868620
GenderMale	-0.5503703	0.1092939	-5.0356927	0.0000005
Age	0.0011630	0.0032635	0.3563756	0.7215593
`Household Income Category`$125,000 and more	-0.1692796	0.2082329	-0.8129340	0.4162559
`Household Income Category`$25,000 to $49,999	-0.0917060	0.2291566	-0.4001891	0.6890172
`Household Income Category`$50,000 to $74,999	-0.0503782	0.2270369	-0.2218945	0.8243960
`Household Income Category`$75,000 to $99,999	-0.1050227	0.2284622	-0.4596939	0.6457360
`Household Income Category`Less than $25,000	-0.1217144	0.2486017	-0.4895959	0.6244198

Let us try to interpret the coefficients of the logistic regression. We will start with the variable *Age*, which is a numerical variable. The estimate for the coefficient β_{Age} is 0.00116, which means that a 1 unit increase in *Age* is associated with an increase in the log odds of *Party Vote* by 0.00116 units. The standard error of this coefficient $SE(\beta_{Age})$ is 0.00326, and the z-statistic given by $\beta_{Age}/SE(\beta_{Age})$ is 0.356. The fact that the z-statistic is small indicates that the null hypothesis $H_0 : \beta_{Age} = 0$ holds. This null hypothesis implies that the probability of voting Liberal does not depend on *Age*. This conclusion is confirmed by the p-value of 0.722 which is large, indicating that the *Age* variable is not significant.

Looking at the data for the *Household Income Category* variable, we reach similar conclusions as for the *Age* variable. For each one of these categories, the z-statistic is small and the p-value is larger than the significance level of 0.05, indicating that these coefficients are not statistically significant.

The final variable is *Gender*, which is a categorical variable taking two values, male or female. It is encoded as a dummy variable called *GenderMale* in R which is equal to 1 when the respondent is male, and 0 when the respondent is female. The coefficient estimate $\beta_{GenderMale} = -0.550$, and the standard error is 0.109. This implies that the z-statistic $\beta_{GenderMale}/SE(\beta_{GenderMale})$ is -5.04 which is relatively large, indicating evidence against the null hypothesis $H_0 : \beta_{GenderMale} = 0$. Also, the p-value associated with *GenderMale* is very small (0.000000476), allowing us to reject H_0. In other words, we conclude that there is indeed an association between *Gender* and the probability of *Party Vote*. This result agrees with the data visualization that we did earlier in the report.

Poststratification

We now create the cells needed for poststratification and record the size of each cell. Cells are divided by *Age*, *Gender* and *Household Income Category*. There are a total of 756 cell splits. Table 6 shows a sample of how our cells look like:

Table 6: First four cell splits

Household Income Category	Age	Gender	n
$100,000 to $ 124,999	18	Female	6
$100,000 to $ 124,999	18	Male	15
$100,000 to $ 124,999	19	Female	5

Household Income Category	Age	Gender	n
$100,000 to $ 124,999	19	Male	6

Next, we create the *Log-odds* using the *predict()* function. Notice that these log odds can be negative. We then transform the log odds into actual probabilities using the relation $p = 1/(1 + exp(-\text{Log-odds}))$. *Probability Estimate* is the probability that a given cell will vote Liberal. Note that we can calculate these probability estimates directly from the *predict()* function by setting the type parameter to "response".

Table 7: Calculating the probability estimates

Household Income Category	Age	Gender	n	Log-odds	Probability Estimate
$100,000 to $ 124,999	18	Female	6	0.3640673	0.5900246
$100,000 to $ 124,999	18	Male	15	-0.1863031	0.4535585
$100,000 to $ 124,999	19	Female	5	0.3652303	0.5903060
$100,000 to $ 124,999	19	Male	6	-0.1851400	0.4538468

Finally, we multiply the probability of each cell by the cell size and store the results in a new column. These are the numerator terms in Eq. (1). Summing all the values in this new column and dividing by the sum of the values in the n column, we arrive at the final result of the poststratification calculation of the probability of the Liberal vote to be: $Y^{PS} = 0.51$.

Given that this probability is greater than 0.5, we conclude that the Liberal party will get a majority of the vote according to our logistic regression analysis with post-stratification.

Conclusions

Summary

In this report, we performed a logistic regression based on phone survey data of Canadians using three variables of interest: *Age*, *Gender* and *Household Income*. Our regression analysis indicated that the *Age* and *Household Income* variables were not significant, but that *Gender* plays a role in determining voting preferences of Canadians. We then performed post-stratification to transpose the results of our logistic regression from the phone survey data to a much larger census data. The result of our analysis seems to suggest that the Liberal party will win the 2023 elections. However, this conclusion should be taken with a grain of salt given the role that other important parties like the New Democratic Party (NDP) or the Bloc Quebecois can play in the election.

Limitations

The greatest limitation to our analysis was the difference in how one of our variables was categorized by our two datasets. In the CES survey, when asked about gender, respondents were given the following options: a. man, b.woman, and c. other (e.g. Trans, non-binary, two-spirit, gender-queer). In comparison, the GSS survey did not collect gender identity and only asked for the sex of the respondants. While the two terms 'sex' and 'gender' may have been used interchangeably before, considerable amount of research has shown that they are not synonymous (Westbrook & Saperstein, 2015). While sex can easily be reported by a respondant simply based off on their biology, gender identification is more complex and fluid. Interestingly enough, when surveys ask respondents about their sex, they are taking that data and making assumptions that have less to do with the respondents biological make up and more about their gender identity and its role in context to their research question. Therefore, it is more beneficial to have surveys that include gender, especially when such surveys are used to inform public policies. In respect to our analysis, this is also an issue when we attempt to create some sort of mapping from genders to sex. The best approach we found was to impute

gender in the population by applying non-binary relative to the proportion of those who choose to respond as either Male of Female. This approach is found to be the one which produces the lowest mean squared error (MSE) (Kennedy, 2020).

An additional limitation to our analysis was that the GSS data collected and listed personal income which included only income from yourself and/or your spouse. The CES dataset however, asked the total household income, not specifying if this was restricted to the respondent, their spouse and/or any additional earning individual in the household. Therefore, there is a chance that this question was answered based on the respondents own evaluation of the question. This could mean that the two surveys were measuring two different things. Lastly, it would've been ideal that the CES and GSS datasets we were using, contained data from the same year however that was not the case. The CES data was conducted in 2019, while the GSS data was conducted in 2017.

Next Steps

Future analysis examining whether age, gender, and household income can predict voting behaviour should aim to use datasets which can be more easily mapped on to each other. The way in which variables of interest are categorized and collected by each survey should be compared to avoid issues similar to the ones we faced with our gender and household income variables.

In our results we found that less than 50% of individuals between the ages of 18-23 years old were more likely to vote for the Liberal political party. Based on our analysis this would mean that the remaining majority of this age group would be more likely to vote for the Conservative party. We found these results to be quite odd since this age group is more socially progresive with attitudes and beliefs that are in line with Liberal party's policies and mission. Therefore, we would expect that the number of 18-23 year olds voting for the Liberal Party to be substantially higher than those voting for Conservative. We suspect that in future analyses which include more political parties results would show that it is untrue that the majority of this age group is voting for the Conservatives. Rather, an additional percentage of the group would be found to be voting for other socially progressive political parties such as the NDP. Perhaps our own data would conform with the notion that younger age groups (like the one in question), are more socially progressive if we had considered more political parties.

Bibliography

Anderson, C. D., & Stephenson, L. B. (Eds.). (2011). Voting Behaviour in Canada. UBC Press.

Frost, Jim, et al. "Multicollinearity in Regression Analysis: Problems, Detection, and Solutions." Statistics By Jim, 14 May 2021, statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/.

Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.2.1. https://CRAN.R-project.org/package=kableExtra

Kassambara, & U, M. (2018, March 11). Logistic Regression Assumptions and Diagnostics in R, 24, 122 - 123.

Kennedy, L., Khanna, K., Simpson, D., & Gelman, A. (2020). Using sex and gender in survey adjustment. arXiv preprint arXiv:2009.14401.

Little, R. J. (1993). Post-stratification: a modeler's perspective. Journal of the American Statistical Association, 88, 1001–1012.

Merriam-Webster. (n.d.). Psephology. In Merriam-Webster.com dictionary. Retrieved May 27, 2021, from https://www.merriam-webster.com/dictionary/psephology

Paul A. Hodgetts and Rohan Alexander (2020). cesR: Access the CES Datasets a Little Easier.. R package version 0.1.0.

Political Parties. (2020, November 17). The Canada Guide. https://thecanadaguide.com/government/political-parties/

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Statistics Canada. (2013). The General Social Survey - An Overview. Statistics Canada. https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2013001-eng.htm

Statistics Canada. (2020). General Social Survey- Cycle 31: Families (Version 45250001) [Public Use Microdata File Documentation and User's Guide]. Social and Aboriginal Statistics Division.

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

Tables. (2016). R Markdown from RStudio. https://rmarkdown.rstudio.com/lesson-7.html

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

Westbrook, L., & Saperstein, A. (2015). New categories are not enough: Rethinking the measurement of sex and gender in social surveys. Gender & Society, 29(4), 534-560.

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

Wikipedia contributors. (2021, February 8). Psephology. Wikipedia. https://en.wikipedia.org/wiki/Psephology