Predicting the Canadian Election Results through Post-Stratification

Haoying Shen

May 28, 2021

Introduction

Elections are a time for citizens of their respective country to vote for an authority figure that they believe will lead their country in the best way. Different people will have their own opinions based on their own circumstances on which country to vote for. Unfortunately, what specific circumstances may not be clear. This study will investigate relationships between the backgrounds of Canadian citizens and which party they will vote for. The data that will be used includes a phone survey done from 2019 and Canadian census data from 2017, both of which includes information on sex, age, education, province, and income for the participants. Additionally, the phone survey includes what party is preferred by the surveyed individual. With this analysis we will also attempt to predict which party is more favorable based on sex, age, income, education, and location. Understanding what factors contribute to favourability to a party can be important for government parties to realize what features they must improve upon to gain votes [1]. This leads us to our overall research question of if economic, status, age, education level, and sex can be used to predict the election results for 2023. Specifically, we are looking to predict which of the 2 large parties of Canada will win, the Conservative party, or the Liberal party [2]. Since economic status, age, education level, location, and sex all affect the circumstance and outlook of life for the individual, we believe that there may be some effect on what party any individual favors based on these factors, and these effects will be investigated.

Data

The data was collected from the General Social Survey of Canada (GSS), which is a completely voluntary survey that occurs annually. The survey is done through online questionnaires and interviews over the phone [3]. The GSS used in this study in the 2017 version which focuses on family aspects, such as number of children, marriage status, and family origins. The data used to take for this paper includes the sex, financial status, education level, and age of survey participants. The second dataset used is the Canadian Election Study (CES), which is a survey that is done in two steps, first being during the election campaign, and second being after the election [4]. The survey is collected through phone interviews and online. For this study, the phone interviews during the election campaign were used, and information on sex, financial status, education level, age, as well as political party preference was taken.

Data Cleaning

All data cleaning for this report was programmed using `R version 4.0.2` using functions from `tidyverse` [5,6].

The CES data was cleaned by using the r's `select()` function to reduce the dataset to the 6 questions important to our study this was done by running `survey_data %>% select(q2, q11, q4, q3, q69)` . Since question 11 is party preference and it also includes parties aside from the 2 largest parties which is represented by 1 and 2, the responses that were not 1 and 2 are filtered out using `survey_data %>% filter(q11 == c(1,2))`. Question 3 represents the sex of the participant, which originally includes male, female, and transgender, represented by 1, 2, and 3 respectively. The transgender category was removed as there was only one datapoint, which is not enough of a sample size to create any analysis. This was done

using `survey_data %>% filter(q3 == c(1,2))`. Question 69 asks for the income of the individual, which is a raw number; to get rid of the negative values which represents those who did not answer `survey_data %>% filter(q69 >= 0)` was run to get rid of these values. Next the following code was run `survey_data %>% mutate(age = 2021-q2, income = cut(q69, c(-Inf, 24999, 49999, 74999, 99999, 124999, Inf)), party = ifelse(q11 == 1, 1, 0), sex = ifelse(q3 == 1, "Male", "Female"))`. This creates a variable age, which subtracts 2021 from the participant's birth year to give an age. This also sets party to a proper binomial variable where 1 represents the liberal party, and 0 represents the conservative party. Sex is set so that 1 is now Male, and 2 is now Female. Next the `cut()` function is used on income to group the raw income into categories using `survey_data %>% mutate(income = cut(q69, c(-Inf, 24999, 49999, 74999, 99999, 124999, Inf)))`. The categories are then renamed using the `levels()` function `levels(survey_data$income) = c("Less than $25,000", "$25,000 to $49,999", "$50,000 to $74,999", "$75,000 to $99,999", "$100,000 to $ 124,999", "$125,000 and more")`. The same process was done for education, `survey_data %>% mutate(education = cut(q61, c(0, 4, 6, 7, 8, 9, 11)))` where the numerical values were grouped, and then renamed using `levels(survey_data$education) = c("Less than high school diploma or its equivalent", "High school diploma or a high school equivalency certificate", "College/Trades", "High school diploma or a high school equivalency certificate","Bachelor's degree/University certificate","Above University Degree")`. Similarly, the numerical answers of provinces was also converted using `cut()`. `survey_data %>% mutate(province = cut(q4, c(0, 1, 2, 3, 4, 5,6,7,8,9,10)))` and the numbers were renamed using `levels(survey_data$province) = c("Newfoundland and Labrador", "Prince Edward Island", "Nova Scotia", "New Brunswick", "Quebec","Ontario", "Manitoba","Saskatchewan","Alberta",\n "British Columbia")`. The territories were removed from the dataset as there were no datapoints for them. Continuing, `survey_data %>% select(party, age, income, sex, province, education)` was called to take only the variables party, age, income, sex , province, and education. Finally `na.omit(survey_data)` is used to remove any missing values.

Important Variables

Sex – Categorical variable which determines the sex of the surveyed individual. This variable is used to investigate the differences in preferred party between the different sexes.

Age – Continuous variable which determines the age of the surveyed individual. This variable is important in investigating how the relationship between party preference.

Province – Categorical variable which determines the provincial location of the surveyed individual. This variable is important in investigating how the province of an individual influences the which federal party is favored.

Income – Categorical variable that determines the financial status of the surveyed individual. This variable is crucial in investigating the relationship between the financial status of an individual and the party preference.

Education – Categorical variable that is indicates the highest level of education achieved by the surveyed individual. This variable will be used to determine if there is any relationship between education level and the party preference of the survey participant.

Party – Binomial variable that represents which of the 2 major political parties is preferred, conservative or liberal.

All tables are created using the kableExtra package's `kable()` function [7].

Summary and Spread of Data

From Table 1 we can see that the census and survey data show very similar numbers with means and medians around the 50s and roughly the same standard deviation of 16.15 and 17.74, the minimum and maximum values tend to be higher in the survey compared to the census. Table 2 shows the frequency and percentage of income, and we can observe that the survey has a relatively even distribution throughout all income categories, apart from $125,000 and more. Unlike the census, which has most of the observations in the $125000 and more as well as the $75000 to $99999 categories. Table 3 presents the frequency and percentage

Table 1: Summary of spread for participant ages in Survey and Census Data

	Survey Age Summary	Census Age Summary
Min.	20.00	15.00
1st Qu.	42.25	37.00
Median	55.37	53.09
Mean	55.66	50.86
3rd Qu.	66.00	67.00
Max.	95.00	80.00
SD	16.15	17.74

Table 2: Frequency and Percentage of Income Categories of Survey and Census Data

	Less than $25,000	$25,000 to $49,999	$50,000 to $74,999	$75,000 to $99,999	$100,000 to $ 124,999	$125,000 and more
Survey Frequency	49	54	60	47	48	140
Survey Percentage	12	14	15	12	12	35
Census Frequency	839	864	6072	3838	2010	6638
Census Percentage	4	4	30	19	10	33

Table 3: Frequency and Percentage of Education Categories of Survey and Census Data

	Less than high school diploma or its equivalent	High school diploma or a high school equivalency certificate	College/Trades	Bachelor's degree/University certificate	Above University Degree
Survey Frequency	22	108	85	120	63
Survey Percentage	6	27	21	30	16
Census Frequency	1843	4485	6049	4848	3036
Census Percentage	9	22	30	24	15

Table 4: Frequency and Percentage of Survey and Census Participant's Provinces

	Newfoundland and Labrador	Prince Edward Island	Nova Scotia	New Brunswick	Quebec	Ontario	Manitoba	Saskatchewan	Alberta	British Columbia
Survey Frequency	20	17	18	19	67	95	22	29	42	69
Survey Percentage	5	4	5	5	17	24	6	7	11	17
Census Frequency	1698	2480	1166	1318	1076	1405	5531	694	3764	1129
Census Percentage	8	12	6	7	5	7	27	3	19	6

Table 5: Frequency and Percentage of Survey and Census Participant's Sex

	Female	Male
Survey Frequency	142	256
Survey Percentage	36	64
Census Frequency	11018	9243
Census Percentage	54	46

of education levels where both the census and survey sample show similar distributions with the high school, college, and university levels being the most frequent. Table 4 Showcases the distribution of both sample between the provinces, the survey shows a majority in British Columbia, Ontario, and Quebec, which match the population data. Surprisingly, we see that Manitoba and Alberta are the most frequent observations in the census data. Finally, Table 5 shows the distributions of respondent sex, in both datasets, where in the survey males are more frequent and inversely, the opposite is true for the census. Overall, we see that the summary and spread of the data is generally similar between the two datasets.

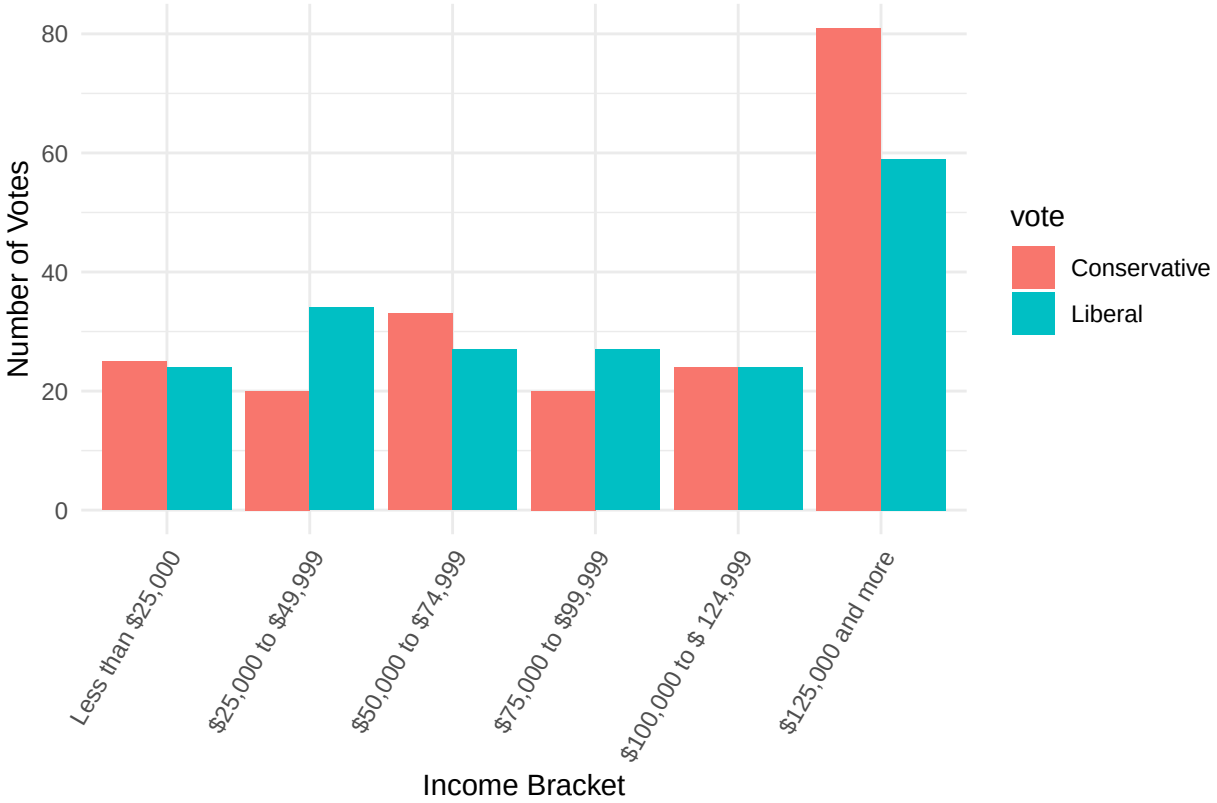

Figure 1. Distribution of votes between income brackets

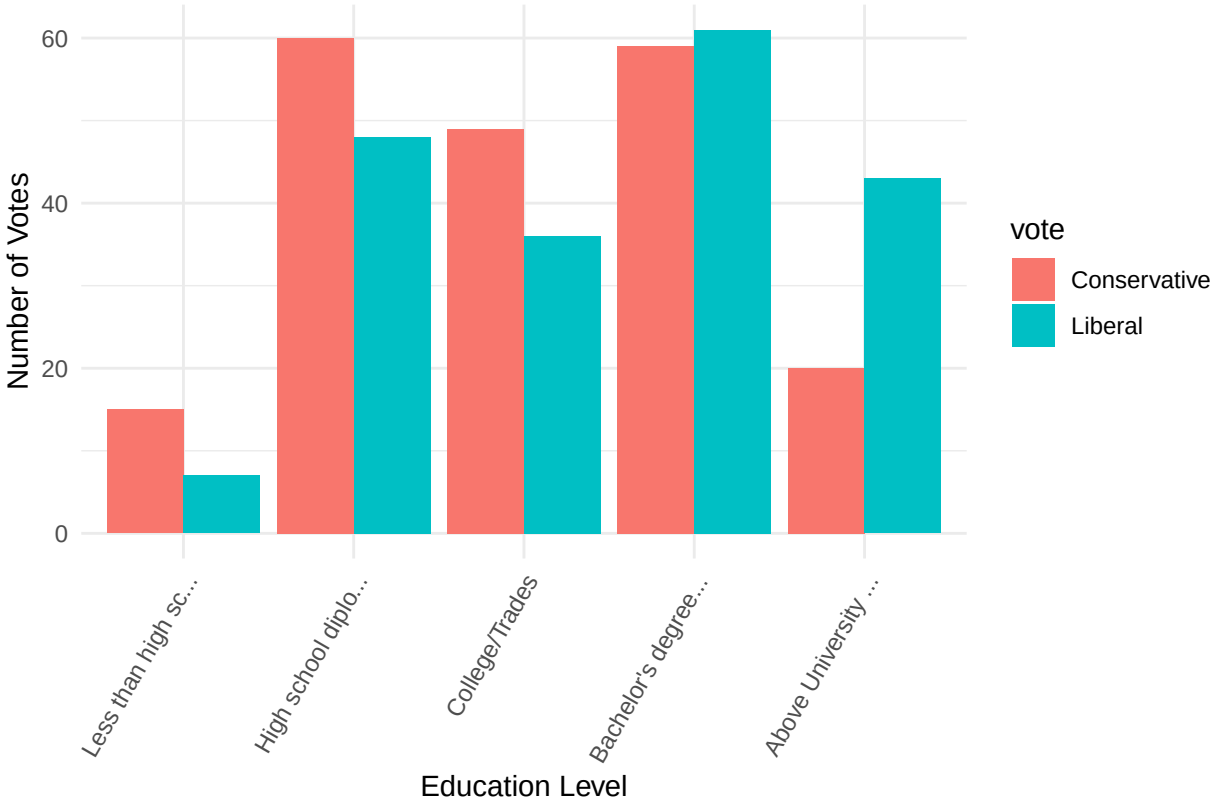

Figure 2. Distribution of votes between education levels

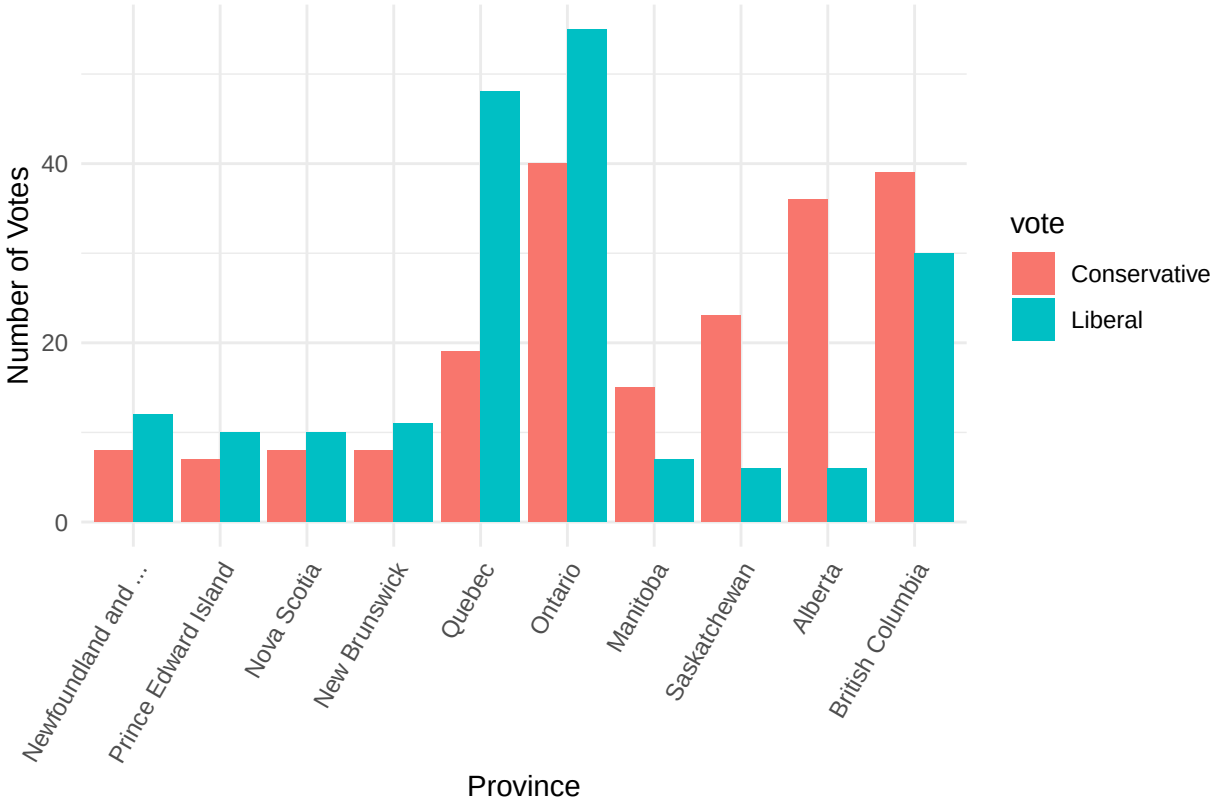

Figure 3. Distribution of votes between Provinces

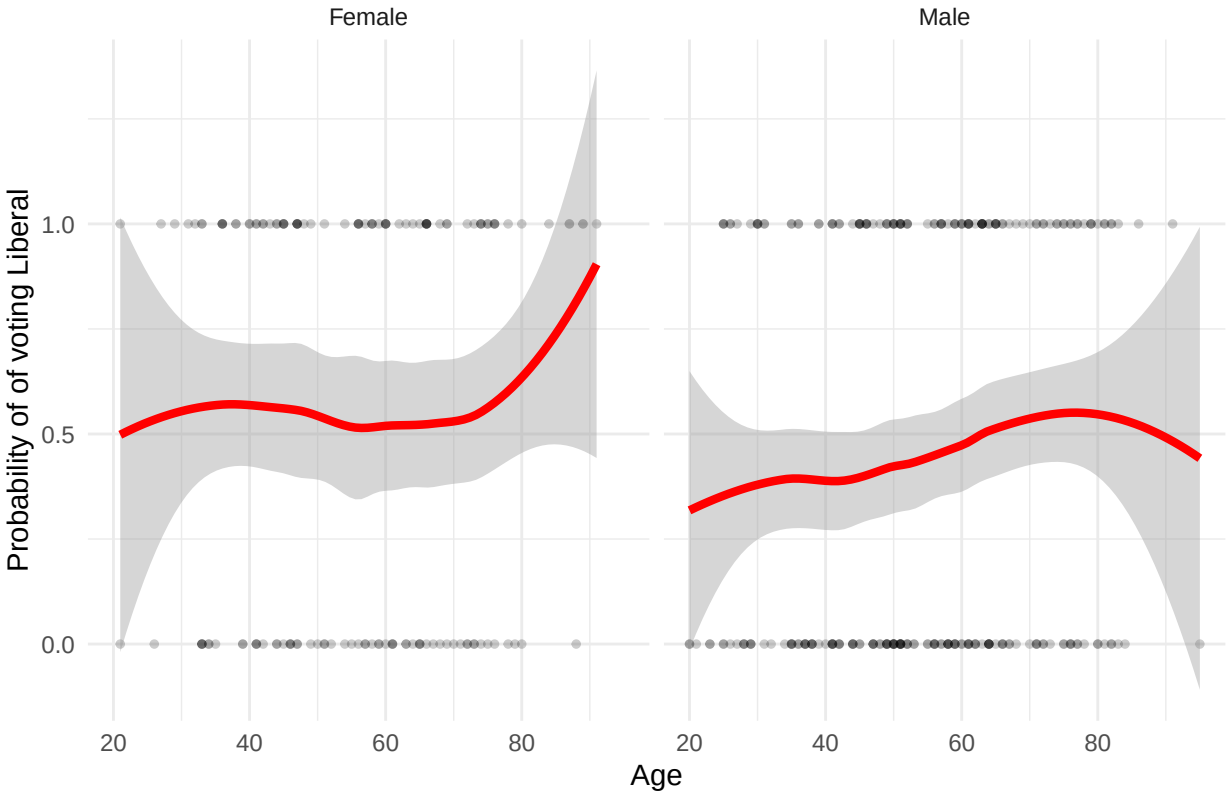

Figure 4. Probability of voting liberal over age, split between sex

All plots in this analysis is created using `ggplot` from 'tidyverse [6].

Figure 1 shows the number of votes for conservatives or liberals between the different income categories. From the bar graph, we can observe that those who make over $125,000 tend to favor the conservatives more. This is similar in the category of $50,000 to $74,999 where conservatives are also favored. Continuing those who make between $25,000 to $49,999 and $75,000 to $99,999 are observed to have voted for liberals more frequently. The other two categories seem to be rather even with their votes.

Figure 2 shows the number of votes for conservatives or liberals between the different education levels. From the figure we can see that those who have not completed high school, those who have graduated highschool, and those in college or trades tend to favor the conservative party. Inversely, those who have education higher than a bachelor's heavily favor the liberal party. Whereas those in the bachelor category show votes between the parties.

Figure 3 shows the number of votes for conservatives or liberals between the provinces. The figure shows that 6 out of 10 countries favor the liberals including Newfoundland and Labrador, Prince Edward Island, Nova Scotia, New Brunswick, Quebec and Ontario. While Manitoba, Saskatchewan, British columbia, and especially Alberta favor the conservatives

Figure 4 shows the probability of of voting liberal given age and sex, where 1 represents a vote for liberals, and 0 represents a vote for the conservatives. we can see that males tend to be more likely to vote conservative while females are more likely to vote liberal. Additionally both sexes tend become more likely to vote liberal as age increases.

Methods

For this study, we will assume that none of the predictors share any relations with each other even should certain variables such as education and income be related in some sense. We will use post stratification to

predict the results of the election, we will do this by using a model to determine the effects of each of the variables: sex, age, province, income, and education on the choice of party on the phone survey data, or in other words their weights. Then we will use those weights on the GSS data, which has far more participants to determine what is the popular vote.

Model Specifics

To be able to predict the outcome of election, we must first use the phone survey data to determine what effects sex, age, province, income, and education on the preferred major party. As our response variable consists of only two outcomes, it is binary. Although, province is a level 2 variable, meaning it centers around a group rather than the individual, indicating a multi-level logistic regression. We would also like to observe the effect of an individual's province on their preferred party. Therefore, ordinary logistic regression is the model to be used. The logistic regression model is as following:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{sex} + \beta_3 x_{income} + \beta_4 x_{education} + \beta_5 x_{province} + \epsilon$$

Where p represents the probability of voting for liberals, and $1-p$ represents the probability of voting for conservatives. $log(\frac{p}{1-p})$ then represents the log-odds of voting liberals β_0 represents the intercept, which occurs should all of the other predictors be 0, or in other words the baseline. β_1 represents how much of an effect age has on determining the log-odds. β_2 represents the coefficient determining how much of an effect sex has on determining the log-odds. β_3 represents the coefficient determining how much of an effect income has on determining the log-odds. β_4 represents the coefficient determining how much of an effect education level has on determining the log-odds. Finally, β_5 represents the coefficient determining how much of an effect the province of an individual has on determining the log-odds

Post-Stratification

We will predict how the GSS dataset will respond when given the parameters of the model created above. To perform this prediction, we must account for the fact that the two datasets may be extremely different as shown in the tables above (table 1, 2, 3, 4, 5). Post stratification allows us to circumvent this by using the model above to determine how each of the individual effects will affect the response variable (party preference) and will match the most likely response of those who fall into certain cells [8]. Cells are one specific and unique combination of all the predictors, including: sex, age, province, income, and education.

To perform the post stratification, we first grouped all the observations in the GSS dataset with groups created by the variables: sex, age, province, income, and education. Where those with the same sex, age, province, income, and education were grouped together and the total number in each category was counted. Since the GSS dataset contains only those of ages 15-80, there are 65 unique groupings of age. With 10 provinces to choose from, there will be 10 groupings from provinces. Income provides 6 more categories to group by, while education provides 5. Finally, sex provides 2 more factors to group by. Overall creating $ 65 * 10 * 6 * 5 * 2 = 39000$ unique cells to group by, of course not all these cells are used, as the dataset only uses 10708 of these cells. Providing us with this post stratification equation:

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

Where $\hat{y}^{PS}$ represents the average estimate for the GSS population. N_j the total number of people who belong in the jth cell and $\hat{y}_j$ represents the estimate of that jth cell.

To simplify, we used the phone survey to see what types of people preferred which of the two major parties. Then using the census data which is far larger we found the people most alike predict how likely each type of individual was to vote for a certain party. This was done for all observations in the census data to determine which party would have more votes in the end.

All analysis for this report was programmed using `R version 4.0.2` [5].

Table 6: Summary of estimates for the logistic regression model on the survey data

term	estimate	std.error	statistic	p.value	odds
(Intercept)	-0.712	0.941	-0.756	0.449	0.491
Age	0.009	0.007	1.259	0.208	1.009
Income: $25,000 to $49,999	0.820	0.446	1.838	0.066	2.270
Income: $50,000 to $74,999	0.163	0.435	0.375	0.707	1.177
Income: $75,000 to $99,999	0.290	0.454	0.639	0.523	1.337
Income: $100,000 to $ 124,999	0.227	0.452	0.501	0.616	1.255
Income: $125,000 and more	-0.084	0.375	-0.225	0.822	0.919
Completed Highschool or Equivilent	0.523	0.579	0.904	0.366	1.687
Completed College/Trades	0.406	0.598	0.680	0.496	1.502
Completed Bachelor's degree/University certificate	0.838	0.580	1.446	0.148	2.313
Completed Above University	1.478	0.617	2.397	0.017	4.384
Prince Edward Island	-0.131	0.700	-0.187	0.851	0.877
Nova Scotia	-0.125	0.683	-0.182	0.855	0.883
New Brunswick	-0.416	0.680	-0.611	0.541	0.660
Quebec	0.411	0.561	0.732	0.464	1.508
Ontario	-0.207	0.525	-0.395	0.693	0.813
Manitoba	-1.151	0.677	-1.701	0.089	0.316
Saskatchewan	-1.709	0.679	-2.516	0.012	0.181
Alberta	-2.281	0.662	-3.446	0.001	0.102
British Columbia	-0.721	0.537	-1.343	0.179	0.486
Sex: Male	-0.313	0.239	-1.308	0.191	0.731

Results

Summary statistics calcuated using r's `summary()` and cleaned using the `broom` package [5,9].

The Model

From the summary plot we see that there are missing values, such as Income: less than $25,000, this is because that value is chosen as the baseline. Meaning keeping all else constant and changing the income category to $25,000 to $45,000 will have 2.27x the odds of voting for the liberal party. We can observe that also increases the odds of voting for liberals by 0.09x for each increase in age. From the odds of income, we can observe that most income brackets are more likely to vote liberals than the baseline which is an income of less than 25000 dollars. Only those who make more than 125,000 dollars are less likely to vote for the liberal party. However, the p-values of these are all rather large apart from those who make 25000 to 74999 dollars per year. Education seems to show an increase in odds of voting for the liberal party as the level of education increases with those who complete levels of education above university having 4.4x the odds of voting for the liberal party in comparison to the baseline of less than high school. The p values are still rather large for the high school graduates and the college/trades category, indicating this increase may just be due to variance. However, the above university category shows high significance, and those who completed a bachelors also show a somewhat moderate significance. Using Newfoundland and Labrador as the baseline many of the provinces show differences in opinion for preferred party. Prince Edward Island has 0.877x the odds of voting liberal in comparison. Nova Scotia and New Brunswick have 0.883x and 0.66x the odds of voting liberal compared to Newfoundland and Labrador, respectively. Similarly, Ontario also shows lower odds of voting for liberals at 0.813x. Quebec is the only province that is more likely to vote for liberals by odds of 1.508x. Manitoba, Saskatchewan, and British Columbia show much lower odds of voting liberal at 0.316x, 0.181x and 0.486x respectively. Alberta shows the lowest odds of voting for liberals in comparison to Newfoundland and Labrador with odds of 0.102x. The p-values of the provinces mostly do not show any significance; however, Alberta, Manitoba and Saskatchewan do show that they are significantly less likely to vote for liberals. Compared the females, males are less likely to vote for liberals in comparison. Overall, we

Table 7: Probability of winning for each party

Party	Probability
Liberal	0.519
Conservative	0.481

do see difference between categories, however the p-values of these are lacking, and many of the differences may simply be due to variance.

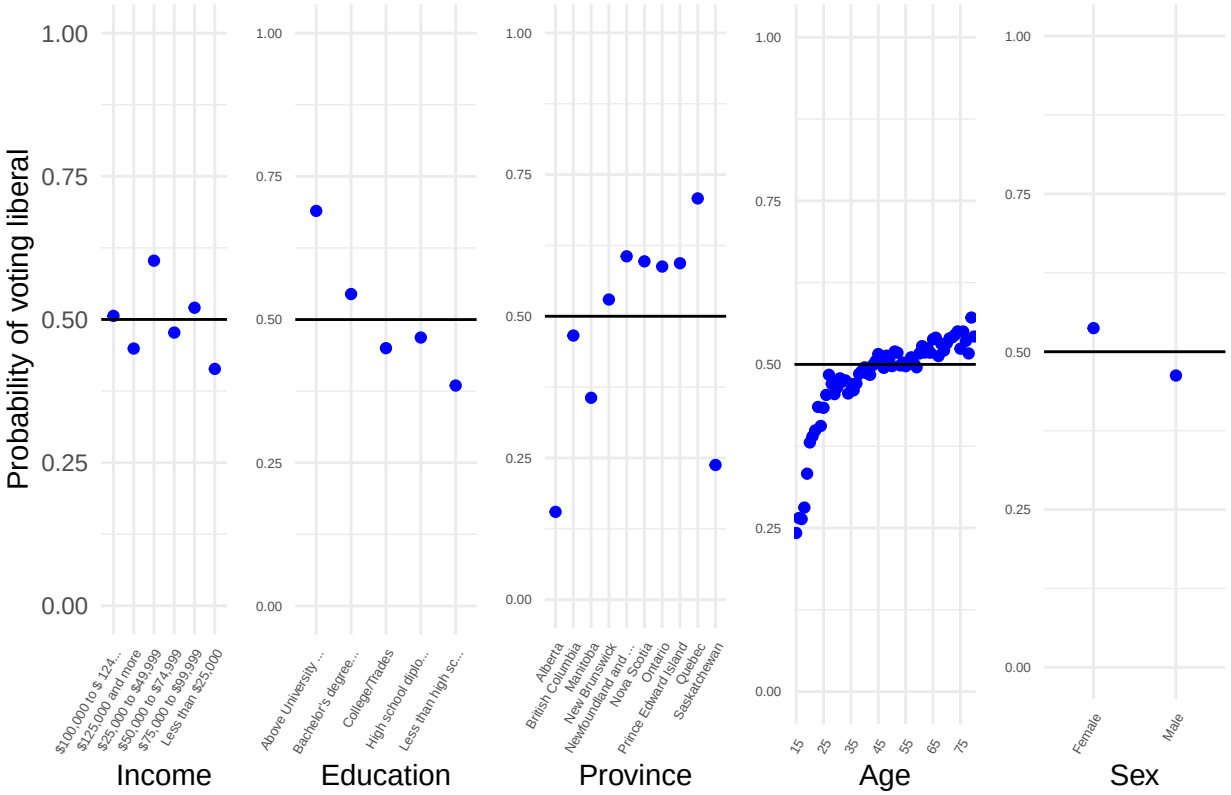

Figure 5: Probability of voting liberal throughout all predictor values for census data

Combined plot created using the `cowplot` package [10].

Prediction

From table 7 we can see that that the liberals are likely to win this election with the census data population voting liberal with a probability of 52%, and the conservative party at 48%, this result does make sense due to the frequency of people who favor the liberals. To begin, from figure 5 we can see that females are more likely to vote for liberals, and from table 5 we see that there are more females in the census data at 54%. Furthermore, with the mean age of the census population being 50.86 as seen in table 1 and judging from figure 5, most of the data points seem to favor the liberals more. Furthermore, from table 3 we can see that although more of the census falls under the line at 0.5, at 61%, those who fall under the college/trades and high school category are not greatly under, while those with above and university degree are far more likely to vote liberal. Provinces seems to disagree with the current consensus, where although 6 of 10 countries are more likely to vote liberal (figure 4). Of the 4 countries more likely to vote conservative, makes up 55% of the census population. Finally, income tends to also favor conservatives where the categories that favor conservatives make up 67% of the census population as shown by figure 5 and table 2. Overall, as the predictors are mixed, but more of the figures tends to favor the liberals, the predicted result does seem fair.

Conclusions

To conclude, we have created a reasonable prediction using post-stratification with a logistic regression model on the election using the General Social Survey of Canada and the Canadian Election Study. From the data we were determined to see if certain predictors which affect Canadian life such as sex, age, income, education, and province affected their preferred choice of federal party. From the results we do see differences between categories of these variables, and through the post stratification we have predicted liberals to win with a probability of 52%. This, difference is indeed small, which falls in line with what is found with the individual variables, where the population and proportion of those who favor certain parties tend to be equal, while slightly favoring the liberals. This study is not without limitations however as certain populations in the census data are overrepresented, irrespective of the actual population. For example, while Ontario has the largest population in all of Canada, it only represents 7% of the census data [11]. Next steps for future analysis would be to use a dataset that is more representative of the actual population to create a more accurate prediction.

Bibliography

1. Cohn, N. (2015, December 19). Why voter data is important to campaigns. Retrieved May 31, 2021, from https://www.seattletimes.com/nation-world/why-voter-data-is-important-to-campaigns/
2. Parliament of Canada. Political Parties and Leaders. (2021). Retrieved May 31, 2021, from https://lop.parl.ca/sites/ParlInfo/default/en_CA/Parties/politicalPartiesLeaders
3. Government of Canada, Statistics Canada. (2019). General Social Survey: An Overview, 2019. Retrieved May 31, 2021, from https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2019001-eng.htm
4. Stephenson et al. (2020). 2019 Canadian Election Study - Online Survey, https://doi.org/10.7910/DVN/DUS88V, Harvard Dataverse, V1
5. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
6. Wickham et al. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
7. Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. http://haozhu233.github.io/kableExtra/, https://github.com/haozhu233/kableExtra.
8. Alexander, R. (2021). Telling Stories With Data. Retrieved May 31, 2021, from https://www.tellingstorieswithdata.com/multilevel-regression-with-post-stratification.html
9. David Robinson, Alex Hayes and Simon Couch (2021). broom: Convert Statistical Objects into Tidy Tibbles. https://broom.tidymodels.org/, https://github.com/tidymodels/broom.
10. Claus O. Wilke (2020). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. R package version 1.1.1. https://wilkelab.org/cowplot/
11. Government of Canada, Statistics Canada. (2021). Population estimates, quarterly. Retrieved May 31, 2021, from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000901