Will the Conservatives win the 2023 Canadian Federal Election?

predictions made according to 2017 census and 2019 survey data

Authors: Hedi Wang and other 3 group members

May 28, 2021

Introduction

Canada has five major political parties: Bloc Québécois, Conservative, Green, Liberal, and New Democratic (NDP) (Elections Canada). The results of the 2019 election show that the Liberal Party and the Conservative Party were very close in the number of voters they secured: Liberal-5911588 vs. Conservative-6150177 (Grenier 2019). However, the Liberals won more seats since the majority of their supporters were from Canada's two largest provinces: Ontario and Quebec; even though they had fewer votes than the Conservative Party. In a recent CBC report, a poll analyst stated that 35.6% of the voters would vote for the Liberals and 29.8% would vote for the Conservatives (Grenier 2021).

Table 1 is a summary of the 2019 Canadian Federal Election results. The first row lists all five major political parties in descending order according to the number of seats won. The rest of the rows lists the information corresponding to the specific party. For instance, row two lists the leader of each party, and row three and four display the number of seats and votes they won.

2019 Canadian Federal Election Results

	Liberal	Conservative	Bloc Québécois	New Democratic (NDP)	Green
Leader	Justin Trudeau	Erin O'Toole	Yves-François Blanchet	Jagmeet Singh	Annamie Paul
Seats	157	121	32	24	3
Votes	5911588	6150177	1377234	2845949	1160694

Our report analyzes the association between the three covariates: age, family income, educational level, and the response variable: the political party that a voter will vote for; based on the 2019 Canadian Election Study Phone Survey and 2017 Census data. By doing this, we will be able to estimate the proportion of votes for each major party at the provincial level and ultimately predict the winner of the 2023 federal election.

Before conducting the analysis, We predict that the winner of the 2023 election will be decided between the Conservative Party and the Liberal Party. Furthermore, the Conservatives will lose to the Liberal party even though the Conservatives gained the most votes compared to the 2015 election amongst all other parties in 2019. The reason for this prediction is that Ontario and Quebec have the most seats in the House of Commons, and Liberals win more of the votes than the Conservatives for these two provinces (Grenier 2019).

Data

In this report, we use the General Social Survey (GSS) data and the Canadian Election Study (CES) data to predict the proportion of the population in favor of the Conservative Party of Canada. In 2017, the GSS, targeted people over the age of 14 living in the 10 provinces of Canada, conducted a survey to observe the changes in the lives of Canadians. In 2019, the CES data was collected from Canadian citizens and permanent residents who were 18 or older through a telephone survey, to understand the Canadian election better by gathering their opinions.

We are interested in using this data to build a model that predicts whether or not a person will vote for a specific party. To build the model, we need variables that would affect a person's choice. The variables we use in our model are *age*, *education*, and *family income*. Another variable we are interested in is *province* because it is an important factor needed for grouping our data and performing post-stratification. Here's a list of the ten provinces we will be grouping by 1) Newfoundland and Labrador, 2) Prince Edward Island, 3) Nova Scotia, 4) New Brunswick, 5) Quebec, 6) Ontario, 7) Manitoba, 8) Saskatchewan, 9) Alberta, and 10) British Columbia.

To build the model that predicts whether or not a person will vote for a certain party, we need a response variable indicating parties respondents are willing to vote for. However, since we are interested in the preference of one party at a time, we create five new variables that each indicates whether a respondent will vote for the Liberal party, the Conservative Party, NDP, Bloc Québécois, and the Green Party.

Even though every information we need to build the model is given in both datasets, some variables exist in different forms depending on the datasets.

For example in the GSS data, *the year a respondent was born* is provided instead of *the age of a respondent* (in the CES data). So, we create the *age* variable in the CES data by estimating the age of a respondent. We can estimate the age by subtracting the year the respondent was born from 2019 because the CES data was collected in 2019.

Since *education* is another variable that exists in different forms in both data sets, we clean the variable by rearranging every response into six categories: 1) Less than high school, 2) High school or its equivalent, 3)Post-secondary education excluding bachelor's degree, 4) Bachelor's degree, 5) Above the bachelor's degree, and 6) Unspecified or currently in education.

In the CES data, *family income* data was collected as numerical values while they were recorded as categorical values in the GSS data. Since we don't know the exact amount of family income recorded as a categorical value, we clean the numerical values in the CES data by transforming the values into categorical values with the same format as the data in the GSS data. Here is a list of six family income categories: 1) Less than $25,000, 2) $25,000 to $49,999, 3) $50,000 to $74,999, 4) $75,000 to $99,999, 5) $100,000 to $124,999, and 6) $125,000 and more.

We add one more column for cell proportions. For our calculation purpose, we first divide the cleaned data after previous cleaning steps into subset data frames by province and calculated cell proportion within each province. Then we merge all subsets together to get our completely cleaned data.

The followings are the variables considered to be important in this report.

Categorical Variables: educational level, province, political party, and family income.

Numerical Variables: age.

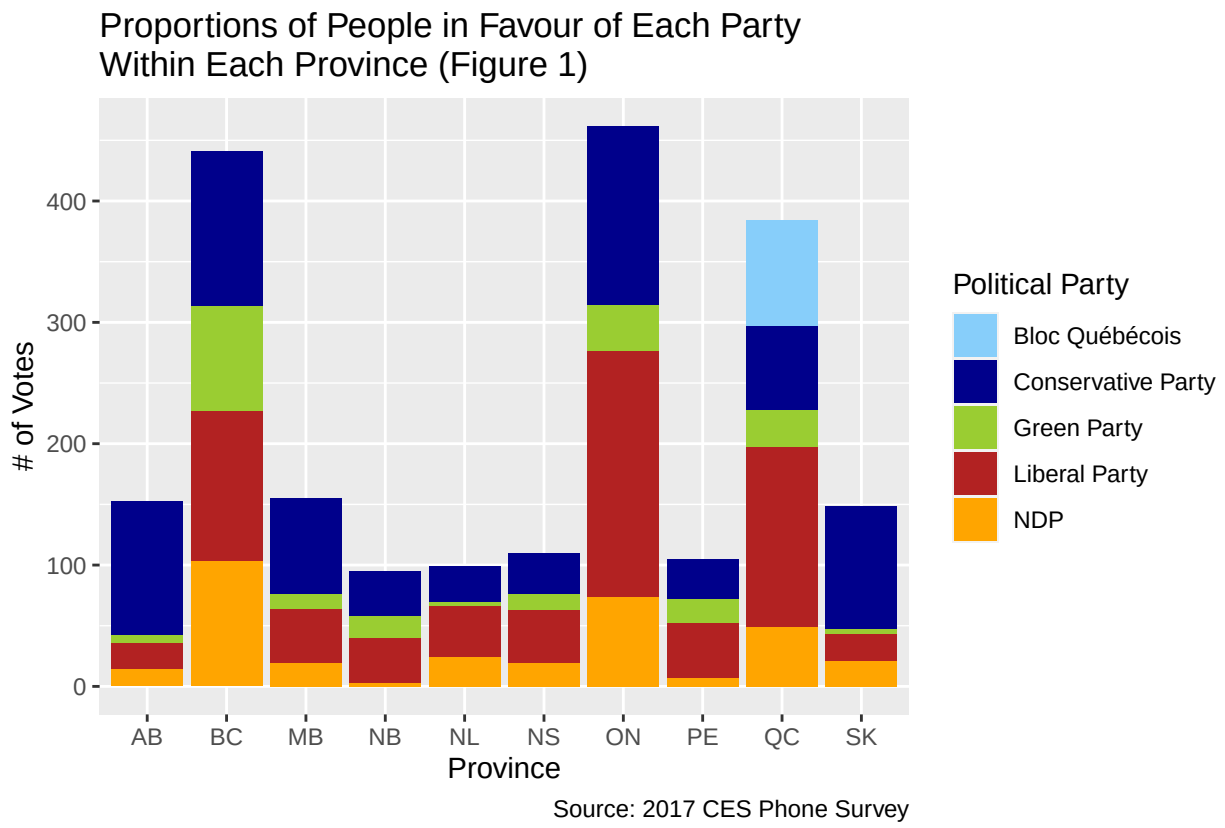

Figure 1 is showing the proportions of votes each party got for each province. The colours used in the histogram pertain to the colours of the logos of the five parties. Liberal-red, Conservative-dark blue, Bloc Québécois-light blue, Green Party-green, New Democratic (NDP)-yellow. The order of the provinces displayed in the histogram follows alphabetic order. According to Figure 1, the Conservative party has about 93 more votes than the Liberal party in Alberta (AB), 31 more in Manitoba (MB), and 78 more in Saskatchewan (SK). However, the Liberal party has about 6 more votes than the Conservative party in Newfoundland and Labrador (NL), 12 more in Nova Scotia (NS), 50 more in Ontario (ON), 12 more in Prince Edward Island (PE), and 81 more in Quebec (QC). They have almost the same number of votes in New Brunswick (NB) and British Columbia (BC). The Green Party, Liberal Party, NDP, and Bloc Québécois have a small portion of votes in each province. Notably, Bloc Québécois has the most votes in QC and no votes in other provinces.

Figure 2: Number of Votes by Party based on CES data

Party	Number of Votes
Bloc Québécois	87
Conservative Party	771
Green Party	231
Liberal Party	731
NDP	333
Other	869

According to Figure 2, The Conservative Party has the most votes (771), with the Liberal Party closely following in second place (731). The Bloc Québécois Party has the least amount of votes (87). The other row shows people who were unsure or did not want to vote for a party.

Figure 3: Frequency of Income level based on CES data

Income Level	Frequency
$100,000 to $ 124,999	392
$125,000 and more	892
$25,000 to $49,999	449
$50,000 to $74,999	538
$75,000 to $99,999	403
Less than $25,000	348

According to Figure 3, most people surveyed for the CES data had an income level greater than $125,000. This tells us that a majority of people surveyed were wealthy. The smallest number of people surveyed had an income level of less than $25,000.

Figure 4: Frequency of Education level on CES data

Education Level	Frequency
Above the bachelor's degree	497
Bachelor's degree	850
High school or its equivalent	402
Less than high school	29
Post-secondary education excluding bachelor's degree	851
Unspecified or currently in education	393

According to Figure 4, most of the people surveyed had a Bachelor's degree(850) or went to Post-secondary education excluding a bachelor's degree(851). A small minority of people had an educational level less than High School.

Figure 5: Numerical summary of Age from CES data

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max
Value	18	38	50	50.32	63	95

According to Figure 5, the average person surveyed was 50 years old. The youngest person was 18 years old and the oldest person was 95 years old.

Method

To arrive at our estimated proportion of votes for each major party in Canada, we conducted logistic regression followed by post-stratification. Logistic regression involves building a regression model to predict a binary dependent variable (Brannick). Post-stratification refers to a data analysis process that assigns observations in a given sample into homogeneous groups after consideration, then uses the calculated weights and predicted values of the small groups to extrapolate the population (Smith, 1991). Please note that our post-stratification goal is to estimate the proportion of votes for each major party in each province, therefore the post-stratification was conducted within each province and then summarized by each province.

Model Specifics

We will be using a logistic regression model to model the proportion of voters who will vote for a particular party at the provincial level to extrapolate the results for the next Federal Election in 2023. We applied the same model for all five major parties individually. More precisely, we only changed the response variable to the binary variable associated with voting the particular party each time without changing other inputs such as the covariates and the additive relation between them. Even though there are five different models for each party, they were all derived from one model. Eventually, the only difference between the models for each party is their coefficients/betas. Therefore, our model can be summarized as the following:

$$\begin{aligned} log(\frac{\hat{p}}{1-\hat{p}}) &= \beta_0 + \beta_1 x_{age} \\ &+ \beta_2 x_{\text{income level 2}} + \beta_3 x_{\text{income level 3}} + \beta_4 x_{\text{income level 4}} + \beta_5 x_{\text{income level 5}} + \beta_6 x_{\text{income level 6}} \\ &+ \beta_7 x_{\text{education level 2}} + \beta_8 x_{\text{education level 3}} + \beta_9 x_{\text{education level 4}} + \beta_{10} x_{\text{education level 5}} + \beta_{11} x_{\text{education level 6}} \end{aligned}$$

Where $log(\frac{\hat{p}}{1-\hat{p}})$ represents the log odds, $\hat{p}$ represents the estimated proportion for voting for a specific party at the provincial level, β_0 represents the intercept of the model. β_1 to β_{11} are coefficients for the covariates in the model. And due to the fact that R selects the base level based on the alphabetic and numerical order, order of levels for categorical variables were changed. Specifically, for family income, the income base level/level 1 is \$100,000 to \$124,999, income level 2 is \$125,000 and more, income level 3 is \$25,000 to \$49,999, income level 4 is \$50,000 to \$74,999, income level 5 is \$75,000 to \$99,999 and income level 6 is Less than \$25,000. As for the education level, the education base level/level 1 is above the bachelor's degree, the education level 2 is bachelor's degree, the education level 3 is high school or its equivalent, the education level 4 is less than high school, the education level 5 is post-secondary education excluding bachelor's degree, the education level 6 is unspecified or currently in education.

Post-Stratification

In order to estimate the proportion of voters who will vote for a particular party at the provincial level, we performed post-stratification. Firstly, we needed the estimated proportion of votes for a given party using the regression models we built. Given that our model is a logistic model, the predicted values directly from the R output were log odds instead of the estimated probabilities, therefore we had to convert the log odds into proper estimates using the formula below :

$$\hat{p} = e^{log(\hat{p}/1-\hat{p})}/(1 + e^{log(\hat{p}/1-\hat{p})})$$

where $log(\hat{p}/1-\hat{p})$ is the predicted log odds by the logistic model and $\hat{p}$ is the estimated probability for voting a specific party.

And again, since we are predicting and summarizing the predicted outcomes in each province instead of the entire country, we calculated the cells proportions within each province rather than proportions out of entire Canada. This is achieved by R in the data cleaning section. With all the elements gathered, we calculated the post-stratification estimated proportion by the following function:

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

where $\hat{y}_j$ is the estimate in each cell, N_j is the population size of the jth cell in the province and $\hat{y}_j^{ps}$ is the post-stratification estimated proportion for a specific province.

Lastly, for each one of the five major parties, we created charts of their predicted proportions by province respectively, which are based on the results of our post-stratification process.

Results

Models for Different Parties

Model for Conservative Party:

$$\begin{aligned} log(\frac{\hat{p}}{1-\hat{p}}) = & -2.2375 + 0.00093x_{age} \\ & + 0.4077x_{\text{income level 2}} - 0.3599x_{\text{income level 3}} - 0.0551x_{\text{income level 4}} \\ & - 0.0722x_{\text{income level 5}} - 0.3075x_{\text{income level 6}} \\ & + 0.4043x_{\text{education level 2}} + 1.0713x_{\text{education level 3}} - 0.3058x_{\text{education level 4}} \\ & + 0.8859x_{\text{education level 5}} + 1.0422x_{\text{education level 6}} \end{aligned}$$

Model for Liberal Party:

$$\begin{aligned} log(\frac{\hat{p}}{1-\hat{p}}) = & -1.3288 + 0.0105x_{age} \\ & + 0.1248x_{\text{income level 2}} + 0.1191x_{\text{income level 3}} + 0.1380x_{\text{income level 4}} \\ & + 0.2420x_{\text{income level 5}} - 0.0736x_{\text{income level 6}} \\ & - 0.2532x_{\text{education level 2}} - 0.6515x_{\text{education level 3}} - 0.7021x_{\text{education level 4}} \\ & - 0.8310x_{\text{education level 5}} - 0.6311x_{\text{education level 6}} \end{aligned}$$

Model for NDP:

$$\begin{aligned} log(\frac{\hat{p}}{1-\hat{p}}) = & -0.5632 - 0.0279x_{age} \\ & - 0.3664x_{\text{income level 2}} + 0.3793x_{\text{income level 3}} + 0.1404x_{\text{income level 4}} \\ & + 0.0617x_{\text{income level 5}} + 0.1968x_{\text{income level 6}} \\ & - 0.1996x_{\text{education level 2}} - 0.3427x_{\text{education level 3}} - 0.7409x_{\text{education level 4}} \\ & - 0.2363x_{\text{education level 5}} - 0.3434x_{\text{education level 6}} \end{aligned}$$

Model for Green Party:

$$\begin{aligned} log(\frac{\hat{p}}{1-\hat{p}}) = & -1.9858 - 0.0054x_{age} \\ & + 0.0760x_{\text{income level 2}} + 0.2303x_{\text{income level 3}} + 0.4564x_{\text{income level 4}} \\ & + 0.1225x_{\text{income level 5}} + 0.4155x_{\text{income level 6}} \\ & - 0.4621x_{\text{education level 2}} - 0.7942x_{\text{education level 3}} - 14.5746x_{\text{education level 4}} \\ & - 0.6643x_{\text{education level 5}} - 0.2645x_{\text{education level 6}} \end{aligned}$$

Model for Bloc Québécois:

$$\begin{aligned} log(\frac{\hat{p}}{1-\hat{p}}) = & -3.6305 + 0.0089x_{age} \\ & - 0.6283x_{\text{income level 2}} - 0.0592x_{\text{income level 3}} - 0.3640x_{\text{income level 4}} \\ & - 0.3942x_{\text{income level 5}} - 0.1676x_{\text{income level 6}} \\ & + 0.1884x_{\text{education level 2}} - 0.1464x_{\text{education level 3}} + 1.2087x_{\text{education level 4}} \\ & - 0.0416x_{\text{education level 5}} - 0.9679x_{\text{education level 6}} \end{aligned}$$

As the beta values/coefficients of the model show, different levels of categorical variables will have a different impact on different parties. For voters whose income level is less than $25,000, adding one more unit of this level will give a positive impact on the NDP and the Green party, whereas adding it will result in a negative impact on the Liberal, the Conservative and the Bloc Québécois. This result could imply that voters whose income is very low would prefer to vote for the NDP and the Green party. Similarly, the results imply that voters with a lower education level, i.e. below high school level, will be in favour of the Bloc Québécois the most.

Post-stratification and Prediction for Different Parties

Post-stratification and Prediction for Conservative Party

Figure 6: Proportion of Votes for the Conservative Party in Each Province

Province	Proportion of Votes
Alberta	0.2595103
British Columbia	0.2511652
Manitoba	0.2439344
New Brunswick	0.2418001
Newfoundland and Labrador	0.2424360
Nova Scotia	0.2378645
Ontario	0.2483047
Prince Edward Island	0.2440517
Quebec	0.2306444
Saskatchewan	0.2522533

The estimated proportion of votes for the Conservative party is approximately ranged from 23% to 25% in different provinces. The Conservative party is in first place amongst all major parties in Canada, slightly ahead of the Liberal party. The Conservative party is 4% ahead of the Liberal party in Alberta and has a 1% to 2% advantage against the Liberal party for other provinces. We cannot tell who will likely be the winner between the Liberal party and the Conservative party since even though the Conservative party has the highest estimated votes, its advantage is minor.

Post-stratification and Prediction for Liberal Party

Figure 7: Proportion of Votes for the Liberal Party in Each Province

Province	Proportion of Votes
Alberta	0.2198824
British Columbia	0.2312395
Manitoba	0.2215557
New Brunswick	0.2238314
Newfoundland and Labrador	0.2188073
Nova Scotia	0.2286015
Ontario	0.2318382
Prince Edward Island	0.2226707
Quebec	0.2218889
Saskatchewan	0.2163543

According to our model and post-stratification result, the estimated proportion of votes for the Liberal party is approximately ranged from 22% to 23% in different provinces. The Liberal party is in second place among all major parties in Canada, slightly behind the Conservative party.

Post-stratification and Prediction for NDP

Figure 8: Proportion of Votes for NDP in Each Province

Province	Proportion of Votes
Alberta	0.1086972
British Columbia	0.1029379
Manitoba	0.1022497
New Brunswick	0.0992501
Newfoundland and Labrador	0.0978209
Nova Scotia	0.1021771
Ontario	0.1038672
Prince Edward Island	0.1031829
Quebec	0.1063676
Saskatchewan	0.1045232

The estimated proportion of votes for NDP is around 10% in different provinces. NDP is in third place among all major parties in Canada, more than 10% behind the Conservative party and the Liberal party.

Post-stratification and Prediction for Green Party

Figure 9: Proportion of Votes for the Green Party in Each Province

Province	Proportion of Votes
Alberta	0.0612609
British Columbia	0.0620343
Manitoba	0.0551193
New Brunswick	0.0562325
Newfoundland and Labrador	0.0530908

Province	Proportion of Votes
Nova Scotia	0.0581266
Ontario	0.0622009
Prince Edward Island	0.0581186
Quebec	0.0572012
Saskatchewan	0.0554991

The estimated proportion of votes for the Green party is approximately ranged from 5% to 6% in different provinces. The Green party is in fourth place among all major parties in Canada.

Post-stratification and Prediction Bloc Québécois

Figure 10: Proportion of Votes for the Bloc Québécois Party in Each Province

Province	Proportion of Votes
Alberta	0.0363161
British Columbia	0.0387111
Manitoba	0.0446560
New Brunswick	0.0453799
Newfoundland and Labrador	0.0458858
Nova Scotia	0.0441567
Ontario	0.0386844
Prince Edward Island	0.0425044
Quebec	0.0458733
Saskatchewan	0.0416842

The estimated proportion of votes for Bloc Québécois is approximately ranged from 3% to 4% in different provinces, slightly behind the Green party. The NDP party is in last place among all major parties in Canada.

Our results from logistic regression and post-stratification lead us to some meaningful assumptions. The coefficients of our logistic models imply that there is a positive correlation between age and voting for the Conservative party, the Liberal Party, and Bloc Québécois, whereas there is a negative correlation between age and voting for the Green party and NDP. However, the impact is minor since the coefficients are extremely small. Unlike age, the absolute value of coefficients is pretty large for education and family income levels, which indicates that these two factors have a greater impact on voting proportions for each party. For example, voters whose income is very low would prefer to vote for the NDP and the Green party, and for voters with lower education levels, i.e. below high school level, they will be in favor of Bloc Québécois the most.

Overall, based on our estimated proportions voting at a provincial level for five major parties in Canada by post-stratification, we can conclude that the Conservative party and the Liberal party are two titans among the five major parties with around a 25% and 23% proportion of votes in each province respectively. However, even though the Liberal Party is behind the Conservative party in votes, the difference is approximately 2% at the most, which is not to an extent that we can tell which one of these two parties will be the winner for the next federal election.

Conclusions

Although we hypothesized that the Liberal Party and the Conservative Party will likely be in first and second place respectively in the 2023 Canadian Federal Election, the Conservative party will have fewer seats than

the Liberal because the Liberals would secure the most votes in Ontario and Quebec, the two provinces with the two highest populations in Canada.

Our results reveal that the Conservatives could secure more of the votes in Ontario and Quebec than the Liberals in 2023. This goes against our original hypothesis. However, the difference in votes secured at the provincial level between the Conservative party (Approximately 25%) and Liberal Party(Approximately 23%) is marginal. Additionally, our results show that the other 3 parties are far behind the Conservative and Liberal Party (NDP: 10%, Green Party: 5-6%, Bloc Québécois: 3-4%).

To arrive at our estimated proportion of votes for each major party in Canada, we conducted logistic regression followed by post-stratification. Our Logistic regression involved building a regression model to predict whether an individual votes for each of the 5 political parties based on their age, income level, and education level. Our Post-stratification process assigned each observation in our data sample into homogeneous cells which share the same characteristics(age, income level, and education level) within each province. We then calculated the post-stratification estimate for the proportion of votes for each party at the provincial level. We did our post-stratification at the provincial level because we were specifically interested in the proportion of votes in Ontario and Quebec since they have the most seats.

The main weakness of our analysis is the data sets that were used for the post-stratification process. Our census data was from 2017 while our Phone survey data was from 2019. There are three problems with the age of the data sets. The first problem is that our data sets for the census and the phone survey are from two different years, which makes it so that our model is built on data from 2019, but the demographic characteristics end up being from 2017, making our prediction inaccurate. The second problem is that both data sets are outdated by 4-6 years to predict the 2023 election. Ideally, we would have census and phone survey data for 2022 to have a more accurate prediction. Because demographic characteristics change over time, sample data closer to the 2023 election would have better represented the current population. Using data that represents the current population well, is crucial for the post-stratification process. The third problem is that since the model is built using the phone survey data, the age of the phone survey data affects the accuracy of the model to predict results in 2023. It would be better to use newer phone survey data because peoples' opinions of each political party may have changed over time.

A future analysis would require collecting updated data for both the census data and phone survey data. Further analysis could also look at more covariates other than age, income level, and education level. Additionally, we could dig deeper into the mathematical relation between the covariates and the response variable. For instance, our model was an additive model, nonetheless, an exponential, polynomial, quadratic and other models could possibly better fit our data.

Bibliography

1. Brannick, M. (n.d.). *Logistic Regression*. University of Florida. Usf.Edu. Retrieved May 31, 2021, from http://faculty.cas.usf.edu/mbrannick/regression/Logistic.html

2. Canada, E. (n.d.). *Registered Political Parties and Parties Eligible for Registration*. Elections Canada. https://www.elections.ca/content.aspx?section=pol&dir=par&document=index&lang=e#Liberal.

3. Ebner, J. (2021, March 17). *How to Use the case_when Function in R*. Sharp Sight. https://www.sharpsightlabs.com/blog/case-when-r/. (Last Accessed: May 28, 2021)

4. Grenier, É. (2021, May 26). *CBC News Canada Poll Tracker. CBC news*. https://newsinteractives.cbc.ca/elections/poll-tracker/canada/.

5. Grenier, É. (2019). *Federal election 2019 live results*. CBC news. https://newsinteractives.cbc.ca/elections/federal/2019/results/.

6. Statistics Canada. (2020, April). *2017 GSS: Families User Guide for the Public Use Microdata File*. Ottawa.

7. Stephenson, L. B., Harell, A., Rubenson, D., & Loewen, P. J. (2020, August 17). *Canadian Election Study, 2019, Phone Survey*. Ontario Council of University Libraries.

8. Smith, T. (1991). *Post-Stratification*. Journal of the Royal Statistical Society. Series D (The Statistician), 40(3), 315-323. doi:10.2307/2348284