Application of Post-Stratification: Predicting the Support Rate of Popular Parties in the 45th Canadian Federal Election

STA304 - Assignment 3

Group 41: Shuo Wang, Zhiheng Zhou

November 5, 2021

Introduction

The 2019 Canadian federal election was held on October 21, 2019. The Liberal Party led by Justin Trudeau won the election with a 33.12% (Joseph, 2019) vote in the 2019 election. The 2019 election had the second-lowest share for a party to form a single-party **minority government** (Joseph, 2019) with the 2021 federal election raking the first. The 2021 federal election was held on September 20, 2021. Justin Trudeau won a third term as prime minister. However, in both years, the Liberal Party lost the popular vote to the Conservative Party. In these years, the gap between the two parties has been reduced a lot, and they are having a similar support rate that would make the next election more competitive.

Based on past data, the Liberal Party and the Conservative Party would be the most popular parties. On the other side, the New Democratic Party and the Green Party are also ambitious. The next federal election is expected to be more competitive and ambiguous. Recently, the COVID-19 pandemic has had a great impact on public health and the economy, so it may change people's responses and attitudes towards the federal election. (Gillies, 2021) COVID-19 as a great exogenous effect would cause the election to be more unpredictable and suspenseful. Thus, in this project, we will estimate the voting probability of the four parties mentioned and thus make a simple prediction for the 2025 federal election.

Our research primarily is interested in ** What are the potential support rates of the popular parties in the 45th Canadian Federal Election? **

Before the study, we predicted that people would have different voting decisions due to the provinces in which they are living. We suggested that there would be a correlation between people's voting decisions and their provinces. For the 45th Cfanadian federal election, we predict that the Liberal Party would win based on the empirical evidence.

In this project, there will be a **Data** section that introduces the data collecting and cleaning process as well as provides a basic overview of the important variables. Then in the **Method** section, we will introduce the method used to construct the linear model, and the model will be presented. The **Post-Stratification** section will contain the process and the aim of post-Stratification. All the results will go to **Results** and we will finally wrap up and give **Conclusions** to the project.

Terminology

Minority government: A minority government is a government formed by a political party that does not have an overall majority of MPs (Members of Parliament)in the House of Commons. (UK Parliament, 2021)

Post-Stratification: Poststratification is a technique for adjusting a non-representative sample (i.e., a convenience sample or other observational data) for which there are demographic predictors characterizing the strata.(Stan, 2021)

Generalized linear model (glm): It is a generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution like Gaussian distribution. (DataCamp, 2020)

Logistic regression: Logistic regression models the probabilities for classification problems with two possible outcomes. It's an extension of the linear regression model for classification problems. (Molnar, 2021)

Log-odds: It is equal to the logarithm of the odds ratio $\frac{p}{1-p}$, where p is probability.

Data

Survey Data

This data set contained sample records of telephone interviews completed by 4021 Canadian citizens during the 43rd Canadian General Election. All participants of the telephone interviews took part in the following two phases:

- The process of the Campaign-Period Survey (CPS) was completed totally over the phone. Among the households with the number of adult Canadian citizens greater than one, the phone numbers were determined employing a modified random digital dialing (RDD) procedure. When six attempts were made but failed to get an eligible respondent for a specific sample record, a new sample record would be called in place of the previous one. The CPS process terminated once the number of qualified records reached the expected quantity of respondents. All the participants in the first process have the right to decide whether to At the end of the first process, the information willingness and the preferred methods of continuing to participate of the respondents were gathered.
- 72% of respondents from the CPS process consent to go through the second phase called Post-Election Survey (PES). According to the preferences recorded, 2067 (72%) were contacted over the phone and the rest 28% were offered a link to the survey through email.

Admittedly, there was no gainsaying that this data collection strategy was maneuverable and of high efficiency. However, some inherent weaknesses cannot be ignored in this case. The phone numbers used in the first phase Campaign-Period Survey (CPS) were randomly selected by the algorithm, from which the manual intervention cannot be conducted. Therefore, we cannot determine the composition of the sample records during the data collection process in order to fit the actual population.

In our scenario, the distribution of voters coming from different provinces tends to diverge from the demographics due to the random phone numbers selection. More importantly, the variety of voters' provinces affects the parties they were willing to vote for. Nearly 50% of voters in Alberta would like to vote for the Conservatives Party, but in Quebec, only 17.7% of voters supported the Conservatives. And for the Liberal Party, both Ontario and Atlantic Canada had relatively high supporting rates (around 40%), while it only gained 17% in Alberta. (Grenier, 2021) As a result, the province could be a reasonable variable to divide our census data into different cells, which is a foundation for correcting the model estimates for the inconformity between survey data and the target population. (discussed in **Post-Stratification** section)

Census Data

The census dataset was from the 2017 General Social Survey (GSS) conducting telephone surveys across the ten provinces. The demographics including all non-institutionalized persons 15 years of age and older are provided. With the aim to monitor Canadians living conditions and offer information on specific social policy issues, the census dataset contains core content and classification variables. The core content measured changes in society while the classification variables, such as age, gender, and province, help to divide people into different groups. (CHASS, 2017)

In our research, we would like to use one of the classification variables, which is the provinces people currently lived in, to categorize our target population that is the same as the population of interest in the GSS survey.

Data Cleaning

For the 2019 survey data, we first looked at the variable "q11", which refers to the survey question "Which party will you likely to vote for" and dropped the undecided, refused, and "will spoil ballet" responses, which may cause bias to our results (and there is no skipped response in the original data).

Then we added several columns `support_l`, `support_c`, `support_n`, `support_g` to the copy of the original survey data. Each represents whether the respondent would like to vote for the specific party (Liberal, Conservative, New democratic, Green correspondingly) in the upcoming federal election. Thus, `support_l` is a binary variable that equals 1 when the respondent was likely to vote for the Liberal Party and 0 otherwise.

The other predictors all have similar interpretations. We then dropped useless variables and kept only `q4` and the support variables we have created. We finally renamed the column `q4` as `province` as well as converted the province names to everyday English rather than the index for clarification. According to the codebook:

1: Newfoundland and Labrador

2: Prince Edward Island

3: Nova Scotia

4: New Brunswick

5: Quebec

6: Ontario

7: Manitoba

8: Saskatchewan

9: Alberta

10: British Columbia

For the 2017 census data, we kept only the variable "province" corresponding to the 2019 survey data. Since there is no missing data in "province", we do not worry about missing value

Introduction to Important Variables

The following is the table of the description of the important variables in our sample dataset. The rounded comprehension of the variables selected is a foundational part of data manipulation and further analysis. It is consistently associated with the statistical and mathematical tools utilized to analyze our survey data. For instance, we choose a proper regression model based on the types and properties of variables included.

Table 1: Introduction to Important Variables of survey data

Variable	Type	Feature
age	num	A number of the Age of the person, calculated by subtracting 2019 by the year that person borned in. (rounded to whole number)
province	chr	A name of province or territory that person is currently living
support_c	num	A binary variable stroing whether the person decides to vote for Conservatives (Tory, PCs, Conservative Party of Canada) or not.
support_l	num	A binary variable stroing whether the person decides to vote for Liberal (Grits) or not.
support_n	num	A binary variable stroing whether the person decides to vote for NDP (New Democratic Party, New Democrats, NDPers) or not.
support_g	num	A binary variable stroing whether the person decides to vote for Green Party (Greens) or not.

Numerical Summary

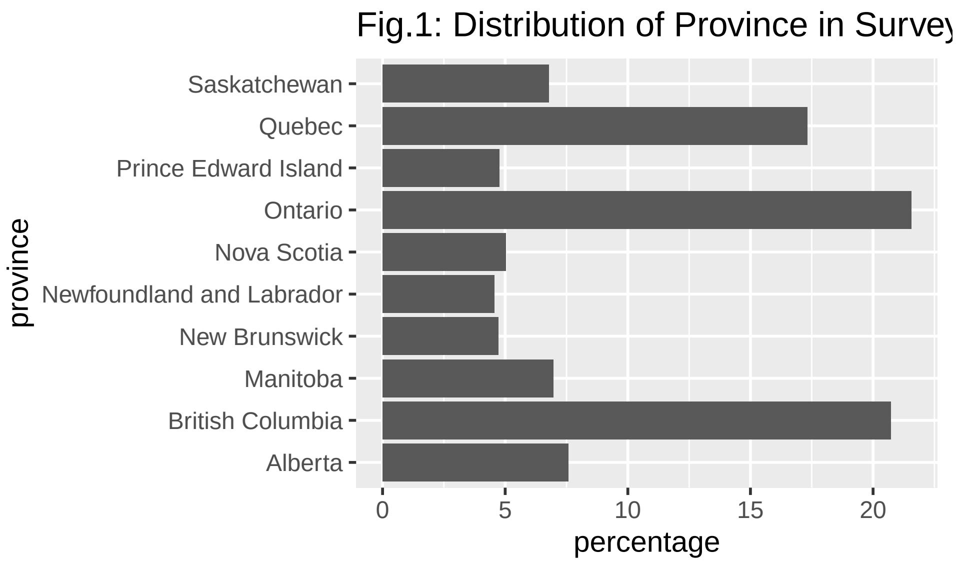
Table 2: Support Rate in survey data

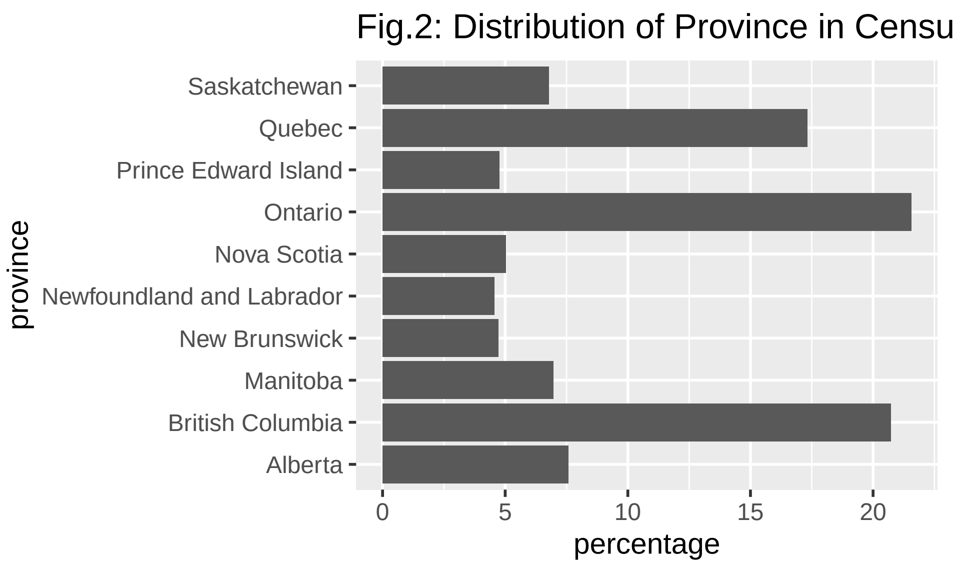
Conservatives	Liberal	NDP	Green Party
0.35	0.33	0.15	0.1

The table output above showcases the support rates of four competitive parties of interest in survey data. It is obviously indicated that the Conservative Party and the Liberal were both popular with approaching support rates over 0.3, while the NDP only gained 15% that is approximately half of their support rate, and the support rate for the Green Party was even lower.

The support rate in survey data 2019 provides the motivation that how the support rate of these four parties performs among our target population, even though our survey sample might be biased relative to the whole Canadian voters. It is worth mentioning that the prerequisite of predicting the supporting rates in 2025 is that the number of eligible voters in each province remains almost unchanged relative to it in 2019.

Graphical Summary





It shows clearly in Fig.1 and Fig.2 that there are exist some discrepancies in the distribution of voters by provinces between survey data and census data. In survey data, more than 20% of voters lived in British Columbia, while less than 10% of eligible voters are from this province based on our demographics data. The differences were inevitable to some extent since the phone numbers through the survey data collection process were randomly selected. However, the difference in the voters' distribution among each province between survey and census data could influence the outcome estimates, which are the supporting rates of interest. Hence, using the survey data to estimate the supporting rate for a particular party is not appropriate, where the Post-Stratification is required to adjust the difference.

Methods

We are interested in the relationship between the likelihood of voting for a specific party and the province of an individual. Obviously, we have a binary dependent variable, thus we need to construct a generalized linear model (glm) with a binomial distribution with a single predictor to run logistic regression. Logistic regression can model the probabilities for classification problems that have only two outcomes such as true/false, included/excluded, and so forth. With logistic regression, we can explore the probability that an individual would like to vote for a specific party based on personal information, especially the province according to our theme.

We selected our model based on the p-values. The p-value is commonly used for checking the significance of the parameters. (Dekking, 2005) Small p-values mean that there would be unlikely to have the extreme outcome under the null hypothesis. Therefore, if the p-value of a parameter is small enough, that is, smaller than the significance level ($p < 0.05$), then we can reject the null hypothesis and conclude that there exists a significant relationship between the predictor and the response, that is, the coefficient of the predictor is not zero($\beta_j \neq 0$). Hence, we checked the p-values of each province for each party. If they are small enough, then we may conclude that our predictor is appropriate for the model of that party.

Model Specifics

$$y = \beta_0 + \beta_1 x_{AB} + \beta_2 x_{BC} + \beta_3 x_{MB} + \beta_4 x_{NB} + \beta_5 x_{NL} + \beta_6 x_{NS} + \beta_7 x_{ON} + \beta_8 x_{PE} + \beta_9 x_{QC} + \beta_{10} x_{SK} + \epsilon$$

To predict the supporting rates for four competitive parties selected based on the provinces of voters, four logistic regression models are required. We set the outcome to be the binary variables, indicating the votes' attitudes of supporting or not on the Conservatives, the Liberal, the NDP, and the Greens respectively.

Since the interpretation of these four models is all similar, we take one of the models that estimate the supporting rate of the Conservative Party as an example.

In our regression model shown above, there is one predictor that introduces the 10 provinces of Canada each as a category and response variable representing if an individual would vote for the Conservative Party in the next(based on any specified year) federal election. x_{AB}, x_{BC},..., x_{SK} are all binary variables, recording the province that a specific voter lived in. β_0 is the intercept and ϵ is the random error.

From the model above, y represents the **log-odds** of whether the individual would vote for the Conservative Party. And the mathematical relationship between y and the probability that the individual would like to vote for the particular in the next federal election is $y = log(\frac{p}{1-p})$.

β_0 represents the intercept of the model, β_1, β_2,..., β_{10} represent the increase in likelihood depending on the province of the individual correspondingly, and ϵ is the random error of the model.

Since we have constructed a logistic regression model, we double-checked that the dependent variable has binary values. However, since we only have one predictor in the linear model, the model would be incomplete and insufficient and may violate the assumption of linearity. On the other hand, we predicted in **Hypotheses** that there would be a correlation between the voting probability (response) and the province (predictor), thus, it may violate the assumption of independence. Fortunately, there aren't any influential outliers and the categories are independent of each other.

Post-Stratification

In the reality, a survey sometimes aimed to investigate the overall population within a wide geographical boundary such as a whole nation or state, but it is not convenient to collect a survey sample that was on behalf of the whole population we were interested in due to some impractical reasons, such as the limited capital support, etc. Therefore, in order to decrease the bias produced by the difference between the sample survey population and the targeted population, post-stratification is applied in our research. It is a statistical technique to reweight the survey results after observation of the data to adjust the inconsistency between survey data and the target population.

Our sample data was gathered from telephone interviews completed by 4021 Canadian citizens during the 43rd Canadian General Election. To explore the willingness of voting for the four competitive parties, i.e.the Conservative Party, the Liberal, NDP, and the Green Party, the variable province was chosen to divide the voters into different cells. Recall from the previous party that voters from different provinces tend to have different attitudes on their opinion of voting for a particular party.

Since the telephone numbers were randomly dialed within the whole country, it was virtually impossible to stratify the sample records by the province before those samples were attained in our dataset successfully. Thus, the sample records in the survey dataset might fail to be representative of the Canadian voters as a whole. That means the distribution of voters recorded in the survey data might not be in line with it in the census data. However, our targeted population, which is all the eligible Canadian voters in 2019, is too large to access through the common data collection process and it will also consume a large amount of money and time.

Our research is primarily focused on predicting the ratio of supporting the different parties in 2025 based on the voter distribution in each province in Canada. Four simple logistic regression models were created (in the "Model Specifics" Section) to estimate the supporting rate based on province for four competitive parties respectively. Hence, for each party, a specific Post-Stratification procedure is needed and two key steps for each Post-Stratification are as followed:(take the Conservatives as an example)

- Employing the relationship between voting willingness and voters characteristics (currently living province in our research) modeled by simple logistic regression above, estimating the preference in individual level, to predict the logarithmic odds of voting for the Conservative Party for each cell. Note that the variable province that has 10 categories partitions the voters in census data into 10 cells. That is to say, voters who lived in a specific province were all allocated into the same cell.
- Reweighting each cell by the relative proportion of voters in each province according to the census data and aggregating the estimates on a voting willingness on voting for the Conservatives in the cell up to the national level. Then the logarithmic odds of voting for the Conservatives among all Canadian voters were attained after correcting the differences in province distribution between the sample population and our target population.

Mathematical Techniques

$$\hat{y_j}^{PS} = \frac{\sum_{i=1}^{10} N_i \hat{y_i}}{\sum_{i=1}^{10} N_i}$$

The formula shown above is the post-stratification measure utilized to speculate the logarithmic odds of voting for each party in Canada. Thus, after four independent processes of Post-Stratification based on four regression models for each party respectively, we will end up having four $\hat{y_j}^{PS}$, where j represents four competitive parties of interest. Let's look at the procedure of calculating $\hat{y}^{PS}_{conservatives}$: $\hat{y_i}$ is the estimate in each cell, i.e., the model estimates in the province level, which represents the logarithmic odds of supporting the Conservatives for voters living in the same province. And the values of N_i, recording the number of voters in a particular cell, are used to weight the estimates in each cell in terms of the demographics. Notice that we have 10 cells in total, thus the accumulation count i is in the range [1, 10]. This fraction is essentially a weighted average on willingness to vote for Conservatives based on the different number of voters in each province, which provides more precise insights to the overall public opinions based on the relatively biased phone survey dataset.

$$log\frac{p_j}{1-p_j} = \hat{y_j}^{PS}$$

To facilitate our interpretation, we could use the above mathematical relationship between the logarithmic odds and the actual probability of voting for a particular party to transform the results produced from the post-stratification process. To be specific, p_j is the corresponding proportion of being willing to vote for the Conservatives, which is the decent to be showcased in the outcome.

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

Tables for analysis in this research are made using the `knitr` 1.33 package (Xie et al., 2021), `kableExtra` 1.3.4 package (Zhu, 2021) and `vtable` 1.3.3 package (Huntington-Klein, 2021).

Results

Table 3: Table of Model summary (For Conservatives)

	Estimate Value	P-value
(Intercept)	1.01	9.011000e-11
as.factor(province)British Columbia	-1.79	0.000000e+00
as.factor(province)Manitoba	-1.00	2.406776e-06
as.factor(province)New Brunswick	-1.53	1.590910e-10
as.factor(province)Newfoundland and Labrador	-2.05	1.000000e-15
as.factor(province)Nova Scotia	-1.88	9.000000e-15
as.factor(province)Ontario	-1.82	0.000000e+00
as.factor(province)Prince Edward Island	-1.84	5.600000e-14
as.factor(province)Quebec	-2.53	0.000000e+00
as.factor(province)Saskatchewan	-0.53	1.351294e-02

model for the Conservative Party:

$$y = 1.01 - 1.80x_{BC} - 1.00x_{MB} - 1.53x_{NB} - 2.05x_{NL} - 1.88x_{NS} - 1.82x_{ON} - 1.84x_{PE} - 2.53x_{QC} - 0.53x_{SK} + \epsilon$$

Note: All the results in the **Results** section would be under the assumption of satisfaction of all the assumptions

Based on our logistic regression models, we can decide the value of the coefficients, that is, the βs. For example, the "estimate" column in the table above represents the coefficients, which refers to the likelihood that people from each province are voting for the Conservative Party in the next election (the 2021 election in this case). The reason why Alberta isn't on the list is that R automatically treats Alberta as a reference, that is, the estimates of the provinces are the results after comparing with the result of Alberta. For instance, people in BC are less likely to vote for the Conservative Party **than people in Alberta** since the estimated coefficient is negative -1.795. Similarly, people in Quebec are much less likely to vote for the Conservative Party than those in Alberta since the estimated coefficient of Quebec has a greater negative value -2.534. Based on this information, we noticed that the Conservative Party would be the most popular in Alberta since all the estimated coefficients have a negative value.

Numerically, a voter living in Alberta has a log-odds of 1.012 to vote for the Conservative Party, while a voter living in BC has a log-odds of -0.783 to vote. The rest provinces all have similar interpretations.

Plus, we notice that all the p-values are smaller than the 0.05 significance level, thus our results are significant, and can reflect the behavior of the real population.

The estimation of the other three parties also has identical processes and similar interpretations with only the differences in the values of the estimation.

Table 4: Supporting Rate Predicted

Conservatives	Liberal	NDP	Greens
0.34	0.33	0.13	0.08

Provided that the immigration among provinces in Canada is not obvious, the census data in 2017 could be considered to be the demographics in 2025. Thus, after building the logistic regression model to estimate the outcome for each province-based cell, we attain the relative proportion in population from census data and aggregate the province-level estimates up to the supporting rate at the national level. The predicted values of the supporting rate are presented above in the table.

It is clearly indicated that the predicted rate of supporting the Conservatives is the highest, but with a proportion slightly higher than the runner-up, the Liberal Party. And the supporting rates for NDP (New Democratic Party, New Democrats, NDPers) and the Green Party are 0.13 and 0.08 respectively, which were at a lower level relative to the voting rates for the first two parties. It is not a surprise that the Conservative Party and the Liberal Party will have the approaching supporting rates, which implies the continued competitive relationship between these two parties in 2025.

Recall from the numerical summaries made in the Data section, the supporting rates of these parties in survey data do not greatly vary from the final predicted rates. The possible reason might be the difference in the distribution of province for voters is not obvious between the survey sample and the census data. In this case, it seems rational to look at the supporting rates of interest simply based on the sample record collected by phone in 2019. Nonetheless, this situation could be a coincidence. The randomness of the selection method of phone numbers is inveterate so that the uncertainty in the sample record exists inherently. We cannot guarantee the survey data relied on this random selection method could capture the variety of the population comprehensively and represent all Canadian voters. Therefore, the Post-Stratification procedure is always necessary and vital when it comes to the known difference between a sample population and the target population in our scenario.

Conclusions

According to our observations, we found that the Conservative Party would probably gain a slightly higher support rate than the Liberal Party. Although the result contradicts our hypothesis, it makes sense in practice as we can see a trend that the Conservative Party in the last few years kept reducing the gap between itself and the Liberal Party. COVID-19 may play an important role in this condition. Since the beginning of the pandemic, vaccinating or not has always been a well-concerned social problem. The leader of the Conservative Party Erin O'Toole describes vaccination as a personal and optional health decision, of which the idea was highly welcomed at the beginning of the pandemic. Nevertheless, as more and more people get vaccinated, there may be more supporters of the Liberal Party as the leader Justin Trudeau asks for mandatory vaccinations (Gillies, 2021). Vaccination is only one exogenous element that may affect the election, and there may actually be a bunch of factors such as speech content that would impact the results. Therefore, there may be more ambiguousness in the next federal election as the pandemic has been lasting for a long time. Any important announcement and behavior of the parties can vary the support rate quickly. Therefore, the results still seem to be unpredictable though we have estimated a slightly higher support rate of the Conservative Party.

Weaknesses

In our research, we made the assumption that the census data, especially the number of voters through ten provinces, does not change dramatically so that the census data 2017 could be the estimated demographics in 2025 and used to predict the supporting rate for each party. However, changes in demographics are inevitable even though no apparent immigration is monitored between each province. Therefore, updating the census data timely is of great significance to make a more precise prediction of the voting rate.

Besides, when estimating the outcome at the individual level, we choose the logistic regression model with simple variable, but hardly can one deny the fact that the supporting rate for a specific party is a complicated variable which is associated with a variety of predictors, not only including demographic information but some social factors such as the financial situation, feelings of well-being and the education level, etc.

In addition, we mentioned in the Data Cleaning section that the group of voters who did not know or have not decided who to vote for was filtered out of our survey data. The non-response bias occurs in this case since the differences between the non-respondents and the respondents influence our survey results in a meaningful way. (Wu & Thompson, 2020) Those who were not included in our survey sample could be the determining factor of the final results because the competition between the Conservatives and the Liberal is intense according to our research outcome, thus the final decision of the non-respondents could be important to the voting result.

Next Step

Equipped with more advanced mathematical and statistical analysis methods, searching for the most suitable model to estimate the supporting rate based on the personal characteristic is worth putting the effort into. Furthermore, it is admitted that the willingness of voting does not present the whole story and the opinions of the people are too labile to be actually captured. So given the sufficiency of time and funds, monitoring the changing tendency on the willingness of the voters through a regular interview can be an effective way to track the fluctuation of the opinions in the target population, meanwhile, narrow the difference between the attitudes shown on the poll and the candidates the voters eventually vote for.

Bibliography

- Grolemund, G. (2014, July 16) *Introduction to R Markdown.* R Studio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)
- Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.
- Allaire, J.J., et. el. *References: Introduction to R Markdown.* R Studio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)
- 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, UNF:6:eyR28qaoYlHj9qwPWZmmVQ== [fileUNF]
- Wu, C., & Thompson, M. E. (2020). *Sampling theory and practice.* Springer International Publishing.
- Brean, Joseph (2019, October 22). "All-Time Low Share of Popular Vote Is Enough for Liberals to Win Power." nationalpost. National Post, October 22, 2019. https://nationalpost.com/news/politics/election-2019/canadian-federal-election-2019-liberals-justin-trudeau-win.
- "Minority Government." UK Parliament.2021. https://www.parliament.uk/site-information/glossary/minority-government/. (Last Accessed November 1, 2021)
- Stan Development Team. "Stan User's Guide." Stan. 2021. https://mc-stan.org/docs/2_23/stan-users-guide/poststratification.html. (Last Accessed November 1, 2021)
- DataCamp Team. "GLM in R: Generalized Linear Model Tutorial." DataCamp Community, 2020. https://www.datacamp.com/community/tutorials/generalized-linear-models.
- Molnar, Christoph. "Interpretable Machine Learning." 5.2 Logistic Regression, November 1, 2021. https://christophm.github.io/interpretable-ml-book/logistic.html.
- Grenier, Éric. "CBC News Canada Poll Tracker." CBCnews. CBC/Radio Canada, 2021. https://newsinteractives.cbc.ca/elections/poll-tracker/canada/.
- Philippe J. Fournier. (2021, Jan 21) "338Canada: Who Leads among Young Voters? the NDP." Macleans.ca. https://www.macleans.ca/politics/ottawa/338canada-who-leads-among-young-voters-the-ndp/
- Uoftlibraries. (n.d.). University of Toronto Libraries. my.access - University of Toronto Libraries. Retrieved November 5, 2021, from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm.
- Gillies, Rob. "Trudeau's Liberals Win Canada Election, but Miss Majority." AP NEWS. Associated Press, September 21, 2021. https://apnews.com/article/canada-health-elections-toronto-coronavirus-pandemic-ddfc4d35fb45559f096bc76f7106949b.