Statistical Analysis on the Potential Implications of the Increased Voter Turnout for the 2019 Canadian Federal Election

Deyu Meng 1004739991 Code and data supporting this analysis is available at: https://github.com/Deyu-Meng/STA304-Final-Project

18/12/2020

Contents

1 Abstract

Plenty of times, citizens have witnessed that having more vote share is simply not enough for candidates to win an election. Election forecasts have become more and more prevalent together with the emerging issue of low voter turnout rate in democratic countries. To investigate the political implications of voter turnout, we have adopted the methodology of establishing a logistic regression model with the post-stratification technique on census data and survey data to predict the proportion of votes for the two major parties of our interest, the Liberal and the Conservative, for the 2019 Canadian federal election in the case of more voter turnout. We have estimated that the Liberal party will win 23.03% of total votes while the Conservative party will win 30.05% of total votes. By comparing results, we have concluded that voter turnout has significant political implications and would potentially make a difference if "everyone" has voted.

2 Keywords

2019 Canadian Federal Election, Liberal party, Conservative party, Post-stratification, Logistic Regression Model, Voter Turnout

3 Introduction

The applications of statistical techniques and analyses have an indispensable role in many aspects of political science. In particular, predicting the election results has become increasingly prevalent as more and more data being attainable. One of the major issues regarding federal elections for many countries with high democracy index, including Canada[1], is the voter turnout rate. This is illustrated by not all voters who are at eligible voting age or registered participated in the voting procedure (DeSilver, 2020) and may potentially lead to the under-representativeness of the data. This not only poses a challenge for the policy makers, but more importantly, it leaves the citizens the question of "what if everyone voted, would the outcome change?".

One of the statistical techniques — the multilevel regression model with poststratification (MRP), is highly useful and cost-effective in terms of correcting for samples that are not really representative for election forecasts and ultimately generating valid and accurate results. (Wang, Rothschild, Goel, & Gelman, 2014) A similar approach to Wang et al.'s through employing the MRP model will be taken for this report to predict the vote proportions of both the Liberal and the Conservative Party for the 2019 Canadian federal election in the case of more voter turnout. In this way, the political impact of turnout would be examined.

In the 2019 Canadian federal election, the Conservative party had the most vote share of 34.4% among all the candidates whereas the Liberal party had a vote share of 33.1%. (Haggarty et al., 2019) However, given the electoral system in Canada, the number of seats elected in the House of Commons is the true determining factor of whether winning the election or not. (Elections Canada, 2020) Thus, the 2019 Canadian federal election resulted in the Liberal party outcompeted the Conservative party by 36 more seats in the House of Commons even though the Conservative party had advantage in the vote share. (Haggarty et al., 2019) The facts that the Liberal party eventually won the election and the voter turnout rate was 67% make the investigation of whether increased voter turnout would yield different outcomes even more fascinating. (Elections Canada, 2020)

Throughout this report, two datasets obtained from the Canadian Election Study, 2019, Phone Survey and the 2017 General Social Survey: Family will be used to investigate the potential outcomes with increased turnout. The data processing procedures and the logistic regression model with post-stratification technique will be discussed in the Methodology section. In the Results section that comes after, the election forecasts when voter turnout is increased will be displayed. Lastly, the statistical inferences, conclusions as well as limitations and further improvements will be included in the Discussion section.

[1] Canada has a score of 9.22 and a rank of the 7th, according to the Economists Intelligence Unit. (2020)

4 Methodology

In this report, we adopt the methodology of the MRP model and perform it on two datasets in order to generate accurate predictions for the two major parties of our interest, the Liberal and the Conservative, for the 2019 Canadian federal election in the case of increased voter turnout.

4.1 Data Collection

Based on the main objective and the methodology of this report, we have retrieved two data sets — Canadian Election Study, 2019, Phone Survey (CES 2019 Phone) via the cesR package and the 2017 General Social Survey: Families (GSS) via the CHASS data center.

- CES 2019 Phone — Survey Data

The CES phone survey was conducted and distributed to potential eligible voters amid and after the 2019 Canadian federal election period through wireless and landline telephone interviews. The target population of this study would be all Canadian citizens and permanent residents at the age of 18 or older. The frame population would be those who are eligible and whose telephone numbers are on the list by the data collector, the Advanis Inc. The sample population would be those respondents who answered the telephone calls. This survey data has a strong focus on the 2019 Canadian federal election which yields high relevance and timeliness. This study employs a "two-wave panel with a modified rolling-cross section" during and after the election as the sampling procedure. There is a specific weighting policy for non-response adjustments that ensures the representativeness of the data. The survey data contains 278 variables and 4021 observations in total, and it consists of the attitudes and information from respondents with respect to the election. It has provided us with data that is rich enough to facilitate our statistical analysis on predicting electoral behaviour. More generally, the survey data is very useful for scholars in many different fields including political science.

- GSS — Census Data

The GSS was conducted from February 2 to November 30 in 2017 and distributed to the respondents through landline and cellular telephone interviews. It serves as the post-stratification dataset and the census data for this report. The target population would be people who are non-institutionalized and aged 15 years old and older from ten provinces in Canada. The frame population would be those who are eligible and whose telephone numbers are on the list accessible to Statistics Canada and the Address Register. The sample population would be those eligible respondents who answered the telephone calls. This study employs stratified sampling together with simple random sampling without replacement as its sampling procedures. There is a specific weighting policy for non-response adjustments that ensures the representativeness of the data. The data contains 461 variables and 20602 observations and it is composed of core content and classification variables. Both of the components target to achieve the purpose of predicting social trends such that the changes in the living conditions and well-being of Canadians will be kept track of. At the same time, they help to identify and give information about certain issues regarding social policies of either existing or emerging interests. This dataset has provided us with demographic information of the respondents.

4.2 Data Cleaning

- Raw Survey Data

Based on the data collecting process of the phone survey, most answers to the questions are recorded with numeric values that have labels to each. This implies that there may be a lot of categorical variables. For instance, "(-9)", "(-8)" and "(-7)" in almost all questions refers to "Don't know", "Refused" and "Skipped", respectively. **Referring to Figure 1 in Appendix**, we see a large portion of the observed data have the classes of haven labelled, vectors and double, some fraction of character vectors and a relatively small portion is numeric. It is noteworthy that there exists "NA"s for the most part of the data except for character and numeric vectors. Thus, we need to select relevant variables and filter out those "NA"s for smoother analysis.

- Cleaned Census Data

Since the raw GSS data contains a fairly large number of variables and observations which could be difficult to deal with, we have first cleaned it using the "gss_cleaning.R" codes provided by Professor Rohan Alexander and Professor Sam Caetano created on October 7, 2020. (Alexander & Caetano, 2020)The total variables and observations are down to 81 and 20602 respectively. **Referring to Figure 2 in Appendix**, we see that a large part of the observed values for the census data are factors, some are integers and a small proportion is numeric. This implies that we may have a lot of categorical variables. It is also worth noting that NAs take up a relatively large portion of the overall data. Therefore, more data cleaning is needed to select only the relevant data.

4.3 Variable Selection and Data Manipulation

Since the objective of this report is to predict the vote share of the Liberal party and the Conservative party for the 2019 Canadian federal election under the situation of more voter turnout, we will specifically look into factors that represent the socioeconomic status of a respondent which could potentially have strong association with voting — age, sex, education and (pre-tax) family income. These will be our independent variables for our model. Our dependent variables in this case would be the vote choice. Given the nature of the MRP model and the large amount of irrelevant data to our report, we will need the data cleaning and manipulation procedures to ensure there is consistency between the two datasets.

In the following subsections, we will discuss our choices of variables and the data manipulation procedures performed.

- age

Since the legal voting age is 18 years old, age is a critical factor that has impacts on voting. (Statistics Canada, 2020) In the survey data, q2 asks "In what year were you born?" and we know that all the respondents are eligible voters, to get their ages, we will have their year of birth subtracted from 2019 which is the year that the survey is conducted. In the census data, the respondents are at least 15 years old and the recorded ages are with decimals. Hence, we will round them first and then filter out those which are less than 18 for data consistency. Ultimately, age is a numeric predictor variable and the minimum age in both datasets are now 18.

- sex

Research has shown that sex is highly related to voting behvaiour in terms of men tend to lean towards the Liberal party. (Goodyear-Grant & Bittner, 2020) In the census data, there are only two categories for sex which are male and female. However, in the survey data, only "Gender" asked in q3 with multiple labels is recorded. We will select only responses "(1)Male" and "(2)Female" and transform them into the variable sex to correspond with the one in the census data. Hence, sex is a categorical variable with two catergories.

- education

Generally speaking, when a person achieves more education, it is more likely that he/she is going to vote. (Feess, 2001) The education factor has different categories in the census dataset and q61[2] in the survey dataset. For matching and simplicity purposes, we combine and recategorize the categories into four new levels of education. Therefore, education has become a categorical variable with four categories.

- (pre-tax) family income

In general, family income has a positive impact on voting behaviour in terms of the more family income, the more likely of voting. (Akee, 2020) In the census data, the variable income_family has 6 new levels. Whereas in the q69[3] of the survey data, responses are collected in specific amounts instead of income ranges. Hence, we combine the survey data and re-categorize them into the income levels that correspond to the census data. Thus, family income is now a categorical variable with six categories.

- vote choice (Liberal or Conservative)

[2] q61: What is the highest level of education that you have completed?
[3] q69: Could you please tell me your total household income before taxes?

Since we are interested in the predicted outcomes of the Liberal and Conservative Party with increased voter turnout for 2019 Canadian federal election, we will rely on the q11 of the survey data which asks the respondents " Which party will you likely to vote for". Hence, we create dependent variables for each party for further analysis and comparison in the following sections of this report. In this case, our response variables will be binary.

4.4 Model

To achieve our goal of this report, we have constructed a logistic regression model with the post-stratification technique on our datasets in R. In general, we first created the logistic regression model based on the survey data. Then, the post-stratification technique is employed on the census data which yields us the final results for the election prediction.

- Logistic Regression Model

The reasoning for constructing a logistic regression model is that our response variable is binary which logistic regression model is considered to be a good fit. Moreover, another reason is that the logistic regression model allows predictor variables to be both categorical and numeric which fits our variables well. In the report, the frequentist method, as in non-bayesian method, is applied.

The following equation in mathematical notations is what we have generated in R for our logistic regression model. This equation is relatively general and can be applied to predicting the probabilities of outcomes for both parties. Hence, there will be a Liberal Model and a Conservative Model.

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_{age} X_{age} + \beta_{sex} X_{sex} + \beta_{education} X_{education} + \beta_{income\_family} X_{income\_family}$$

p_i: the probability of the respondent voting for the Liberal party or the Conservative party in the 2019 Canadian federal election.

β_0: the intercept term. It represents the log odds of a female respondent who has achieved elementary school diploma or less and has family income of \$100,000 to \$124,999. The previously mentioned would be the reference categories for variables sex, education and family income in the case of X_{sex}, $X_{education}$ and $X_{income\_family} = 0$.

β_{age}: the slope coefficient for age. It represents the change in log odds for each unit of increase in age.

X_{age}: the age of the respondent.

β_{sex}: the slope coefficients for sex. It represents the difference in the change in log odds between female and male respondents.

X_{sex}: $= 1$ if the respondent is identified as a male.

$\beta_{education}$: the slope coefficients for education. It represents the difference in the change in log odds between respondents who achieved elementary school diploma or less and respondents who has other three levels of education.

$X_{education}$: $= 1$ if the respondent's level of education is not "Achieved Elementary School Diploma or Less" but any other three categories.

β_{income_family}: the slope coefficients for family income. It represents the difference in the change in log odds between respondents with family income from \$100,000 to \$124,999 and respondents with other five ranges of family income.

X_{income_family}: $= 1$ if the respondent's family income is not from \$100,000 to \$124,999, but any other five categories.

Based on our logistic model, we are now able to make predictions about the probabilities of an eligible voter voting for each party.

During the data manipulation process, in order to have data consistency in both datasets, we have made some adjustments on our four predictor variables. In particular, the variable education from both datasets has undergone the most variation and alteration. We have checked to ensure that this would not cause potential errors that might mislead our predictions by comparing the Akaike's information criterion (AIC) for the current model and the model without the education variable. *Referring to the Table in Appendix*, the current model yields a lower AIC value which means we should keep the education variable.

- Model Diagnostics - Multicollinearity

In this section, we have checked whether our four predictor variables are highly correlated via the Variance Inflation Factors (VIF). Referring to *Table*, the VIF values are all below the cut-off threshold of 5 which means that there is no indication of multicollinearity in our model and all the predictor variables can be kept.

- Alternative Model

An alternative model that would be suitable in this case would be the multilevel regression model since some variables are group-level data.

4.5 Post-Stratification

By applying the post-stratification technique, the census data are partitioned into many demographic cells based on our four predictor variables. We first estimate the probability of voting for the Liberal or the Conservative party in each demographic cell. Then, we continue to extrapolate out to the entire population through weighting each demographic cell based on the population size. The estimate for the proportion of voters voting for each party at the entire population level is carried out. The post-stratification technique helps us to make a relatively more accurate inference based on non-probability based sampling where there is non-response bias.

5 Results

Referring to Table 1, we estimate that the Liberal party will win 23.03% of total votes while the Conservative party will win 30.05% of total votes. Essentially, the predicted results we have arrived at are on the basis of the post-stratification analysis modelled by the logistic regression model which takes into account four predictors variables — age, sex, education and family income.

Table 1: Predicted Results: Proportion of Voters for Liberal and Conservative Party

Liberal	Conservative
0.2303	0.3005

The results are consistent with the survey results from the phone survey datasets and the actual election outcomes in terms of vote share, *referring to Table.*

Table 2: Survey Results: Proportion of Voters for Liberal and Conservative Party

Liberal	Conservative
0.2309	0.2489

Table 3: Actual Election Outcomes: Proportion of Voters for Liberal and Conservative Party

Liberal	Conservative
0.331	0.344

Moreover, it is worth to note that the variable age and the category of "Achieved University Degree, Diploma or Certificate At or Above Bachelor's Level" under the variable education for the Liberal Model are considered to be statistically significant given their small p-values, *referring to Table*. For the Conservative Model, in addition to the previously mentioned, the variables age and sex, as well as four categories under the variable of family income are also statistically significant for their small p-values, *referring to Table*. This finding echoes with our reasoning for choosing these variables that represent the socioeconomic status of the respondents mentioned in the *Variable Selection section*.

6 Discussion

6.1 Summary

The objectives of this report include predicting the proportion of votes for the two major parties of our interest, the Liberal and the Conservative, for the 2019 Canadian federal election in the case of more voter turnout; and examining the political implications of the voter turnout subsequently. In order to meet our objectives, we have obtained two datasets, the CES 2019 Phone and the GSS datasets, and performed data cleaning, variable selection and data manipulation procedures to arrive at four predictor variables (age, sex, education and family income) and our reponse variables of vote proportions. Then, we have constructed a logistic regression model with the post-stratification technique which yields us the prediction results of the Liberal party gaining 23.03% vote share while the Conservative party gaining 30.05% vote share which is consistent with the survey results and the actual election results.

6.2 Conclusion

In light of the statistical analysis conducted throughout this report, it is estimated that there will be 23.03% of total votes in favor of the Liberal Party while there will be 30.05% of total votes in favor of the Conservative party, referring to *Table*. By comparing with the survey results and the actual election outcomes, referring to *Table*, we find that those results are consistent in terms of voters being in favor of the Conservative. The inference could be drawn from this finding is that with increased voter turnout, voters would lean towards the Conservative party. However, according to the Canadian electoral system, one party wins the election by winning more House of Commons sets of a total of 338. In the 2019 federal election, even though the Conservative party gained the most vote share, the Liberal party still won the election by having 36 more seats which resulted in Justin Trudeau leading a minority government. (Haggarty et al., 2019) With the voter turnout for the last federal election being 67% according to Elections Canada, it is possible that the election winner may no longer be the Liberal party. (Elections Canada, 2020)

Turnout rate is crucial for political events like the federal elections, especially in democratic countries. Since different parties have their own ways of governing and sets of policies that are intended to benefit the country and the citizens, if eligible citizens do not participate to vote, then those governing methods and public policies would not be able to play their roles.

6.3 Weaknesses and Next Step

The two datasets that serve as the census data and the survey data may be subjected to both sampling and non-sampling errors. To elaborate, respondents may miscomprehend some questions which may lead to non-responses or answers that fail to meet the objectives of the designed questions. Moreover, data may be recorded incorrectly into the system. These all could potentially lead to increased bias. Generally, the sampling errors are considered to be unavoidable. Moreover, prior to constructing our model, we selected the variables mainly subjectively while we could have omitted many other potential factors that also have a strong association to our response variable. Another weakness of our model is that we only estimated the vote share of the two parties, however, what really determines the election result is the electoral districts, as in how many seats each party has in the House of Commons. (Elections Canada, 2020)

We could gather more data and information and make more comprehensive and accurate prediction results. This means that we will take into account both the vote share and electoral districts given the electoral system in Canada. Also, we can include more variables in our model that could potentially have influence over the predictions and apply the cross validation step or more variable selection methods and criteria such as stepwise regression with Bayesian Information Criterion (BIC). We can even conduct the alternative model of the multilevel regression model mentioned in the Model section. Then, we can compare the results with the actual outcome and draw statistical inferences that would lead to better and more accurate predictions for future elections.

7 Appendix

Figure 1: Raw Survey Data

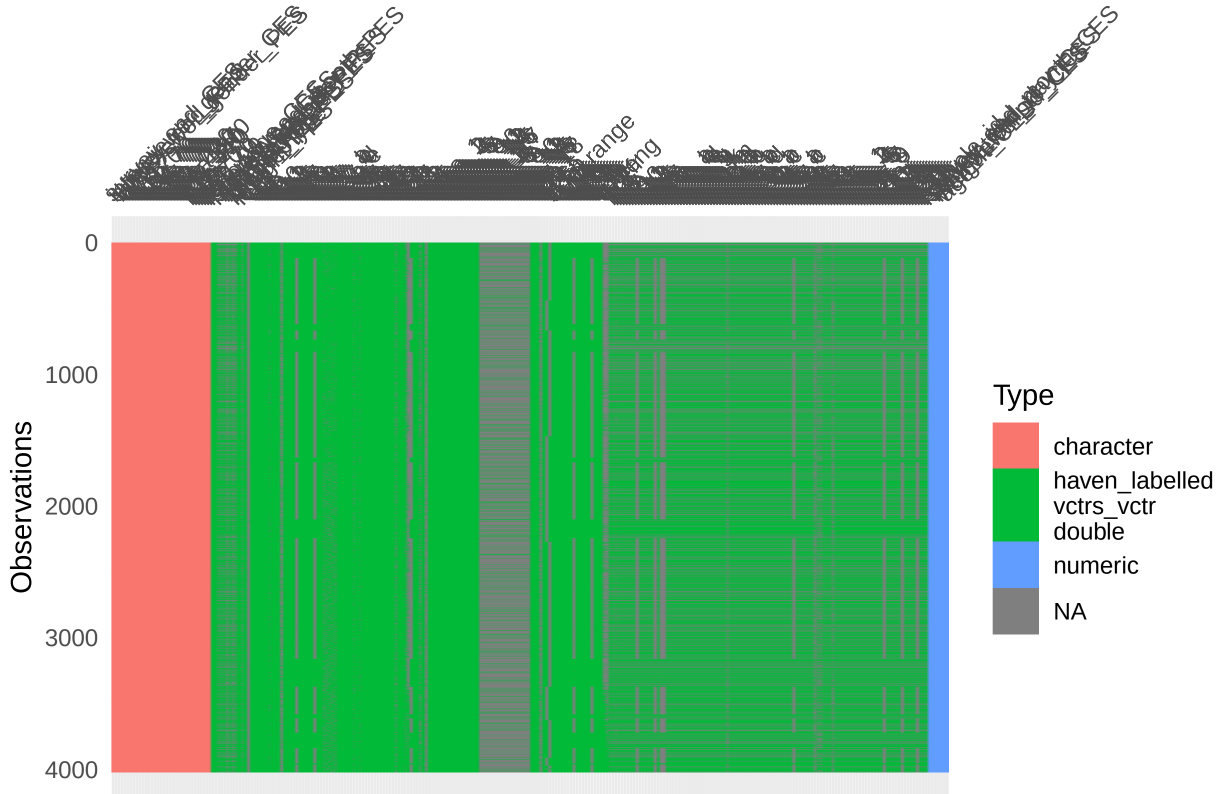

Figure 2: Raw Census Data

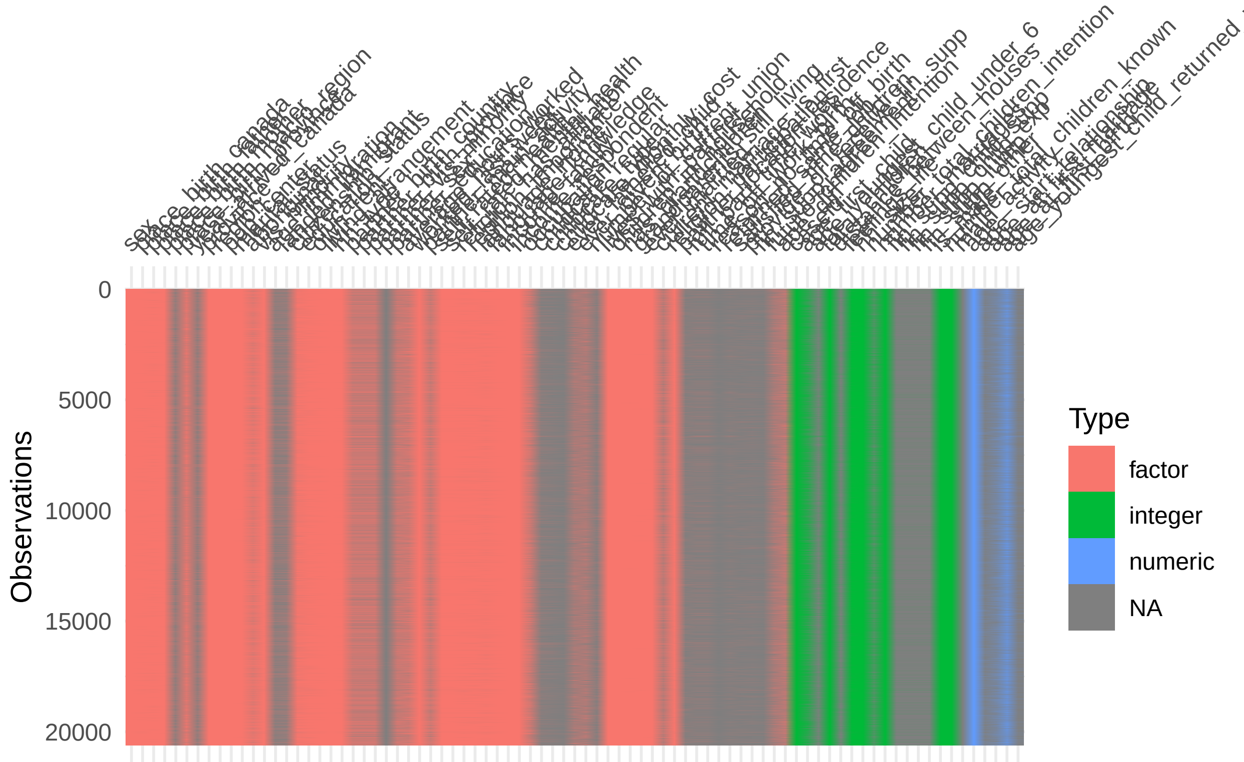

Table 4: AIC for Liberal Models

Include education	Exclude education
3291.2	3332.6

Table 5: VIF Values for Predictor Vairables

	GVIF	Df	GVIF^(1/(2*Df))
age	1.02	1	1.01
sex	1.02	1	1.01
education	1.09	3	1.01
income_family	1.11	5	1.01

Table 6: Summary of Logistic Regression for Liberal Party

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.1519	0.2748	-7.8322	0.0000
age	0.0113	0.0027	4.1289	0.0000
sexMale	-0.0375	0.0879	-0.4273	0.6692
educationAchieved High School Diploma or Equivalent	0.3340	0.2296	1.4549	0.1457
educationAchieved University Degree, Diploma or Certificate At or Above Bachelor's Level	0.7490	0.2212	3.3866	0.0007
educationAchieved University/Non-University Certificate, Diploma or Equivalent Below Bachelor's Level	0.0706	0.2290	0.3085	0.7577
income_family$125,000 and more	0.0627	1.1676	0.0537	0.9572
income_family$25,000 to $49,999	0.0063	0.1360	0.0461	0.9632
income_family$50,000 to $74,999	0.0330	0.1228	0.2690	0.7880
income_family$75,000 to $99,999	0.1463	0.1311	1.1163	0.2643
income_familyLess than $25,000	-0.1669	0.1526	-1.0935	0.2742

Table 7: Summary of Logit Regression for Conservative Party

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.2149	0.2439	-4.9815	0.0000
age	0.0081	0.0027	2.9925	0.0028
sexMale	0.6033	0.0908	6.6411	0.0000
educationAchieved High School Diploma or Equivalent	-0.1632	0.1903	-0.8576	0.3911
educationAchieved University Degree, Diploma or Certificate At or Above Bachelor's Level	-0.7996	0.1871	-4.2740	0.0000
educationAchieved University/Non-University Certificate, Diploma or Equivalent Below Bachelor's Level	-0.2280	0.1875	-1.2156	0.2241
income_family$125,000 and more	0.8914	1.0063	0.8858	0.3757
income_family$25,000 to $49,999	-0.5319	0.1368	-3.8875	0.0001
income_family$50,000 to $74,999	-0.2515	0.1215	-2.0711	0.0384
income_family$75,000 to $99,999	-0.3127	0.1367	-2.2868	0.0222
income_familyLess than $25,000	-0.4580	0.1500	-3.0530	0.0023

8 References

1. Akee, R. (2020, September 24). Voting and Income. Retrieved December 22, 2020, from https://econofact.org/voting-and-income
2. Alexander Rohan, Sam Caetano (2020, October 7). gss_cleaning.R. Retrieved December 22, 2020.
3. DeSilver, D. (2020, November 13). In past elections, U.S. trailed most developed countries in voter turnout. Retrieved December 9, 2020, from https://www.pewresearch.org/fact-tank/2020/11/03/in-past-elections-u-s-trailed-most-developed-countries-in-voter-turnout/
4. Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu and Shannon Ellis (2020). skimr: Compact and Flexible Summaries of Data. R package version 2.1.2. https://CRAN.R-project.org/package=skimr
5. Feess, S. (2001). Does education influence voter turnout? Retrieved December 22, 2020, from https://www.grin.com/document/101356
6. Global democracy has another bad year. (2020, January 22). Retrieved December 9, 2020, from https://www.economist.com/graphic-detail/2020/01/22/global-democracy-has-another-bad-year
7. Goodyear-Grant, E., & Bittner, A. (2020, November 20). How sex and gender influence how we vote. Retrieved December 22, 2020, from https://theconversation.com/how-sex-and-gender-influence-how-we-vote-106676
8. Government of Canada, S. (2020, February 26). Reasons for not voting in the federal election, October 21, 2019. Retrieved December 21, 2020, from https://www150.statcan.gc.ca/n1/daily-quotidien/200226/dq200226b-eng.htm
9. Hadley Wickham and Evan Miller (2020). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 2.3.1. https://CRAN.R-project.org/package=haven
10. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr
11. Haggarty, E., Grasley, R., Kindrachuk, T., Crider, M., Ryan, A., & Sardesai, M. (2019). Federal election 2019 live results. Retrieved December 22, 2020, from https://newsinteractives.cbc.ca/elections/federal/2019/results/
12. John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/
13. Paul A. Hodgetts and Rohan Alexander (2020). cesR: Access the CES Datasets a Little Easier.. R package version 0.1.0.
14. Stephenson, Laura, Allison Harrel, Daniel Rubenson and Peter Loewen. Forthcoming. 'Measuring Preferences and Behaviour in the 2019 Canadian Election Study,' Canadian Journal of Political Science. LINK: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/8RHLG1
15. 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]
16. Technology, Advancing Knowledge through. Computing in the Humanities and Social Sciences. www.chass.utoronto.ca/.
17. Tierney N (2017). "visdat: Visualising Whole Data Frames." *JOSS*, *2*(16), 355. doi: 10.21105/joss.00355 (URL: https://doi.org/10.21105/joss.00355), <URL: http://dx.doi.org/10.21105/joss.00355>.
18. Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. International Journal of Forecasting, 31(3), 980-991.

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

20. Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.

21. Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

22. Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595