Difference in 2019 Canadian Federal Election if Everyone had Voted

Qiyue Zhang

17 December 2020

Abstract

In the last Canadian Federal Election, the Liberal Party narrowly won the most of the seats in the parliament and its biggest adversary the Conservative Party has gained more seats than the last election. In this paper, we aim to identify how the 2019 Canadian Federal Election would have been different if "everyone" had voted. We ran a multilevel logistic regression model using observations from the results of online survey data provided by the 2019 Canadian Election Study (CES). To provide more robust conclusion, we utilize the regression with post-stratification method using 2017 General Social Survey (GSS) data provided by Statistics Canada. Our research finds out that Conservative actually wins the popular vote. Hence, our finding shows the importance of converting people to your cause, getting them to turn out to vote, and how the two are linked.

Keywords: 2019 Canadian Federal Election; Liberal; Conservative; Multilevel regression with post-stratification; Voter turnout

1 Introduction

Voter turnout has been declining in recent Canadian federal elections[Ouellet, 2019]. Just over three-quarters (77.1%) of Canadians reported voting in the 2019 Federal Election. Statistics Canada conducted a supplement to the November 2019 Labour Force Survey that asked respondents about the voting in the October 21, 2019, federal election. Among the 22.9% of Canadians who did not vote in 2019 Federal Election, the main 18 different reasons collected were grouped into four categories: Everyday life reasons (45.9%), Political reasons (41.9%), Electoral process reasons (5.4%), All other reasons (6.8%)[Government of Canada, S. C., 2020, February 26]. For the nation's democracy to function properly and for the government to provide fair representation, all eligible Canadians must have the opportunity to vote and be encouraged to do so.

Even though the final result of the Canadian federal elections involves a stable percentage of non-voting, does it really have an effect on the election outcome? We are interested in identifying how the 2019 Canadian Federal Election might be changed if "everyone" had voted. To accomplish this, we construct a multilevel logistic regression model based on the party preference of the respondents from the survey data as the dependent variable, and their demographic background as explanatory variables, where data is provided by the 2019 Canadian Election Study - Online Survey. We then obtain the fitted estimations by utilizing the model with post-stratification method using the 2017 General Social Survey (GSS) data provided by Statistics Canada.

This paper is structured in the following manner. The two datasets that we used for the model will be discussed in Section 2 along with the description of the data cleaning process. Section 3 introduces the model we choose to construct, that is the methodology of the multilevel regression with post-stratification. Section 4 presents our result on the estimated 2019 Canadian Federal Election outcome. Lastly, we comment on our findings, address limitations, and suggestions for future work in Section 5.

2 Data

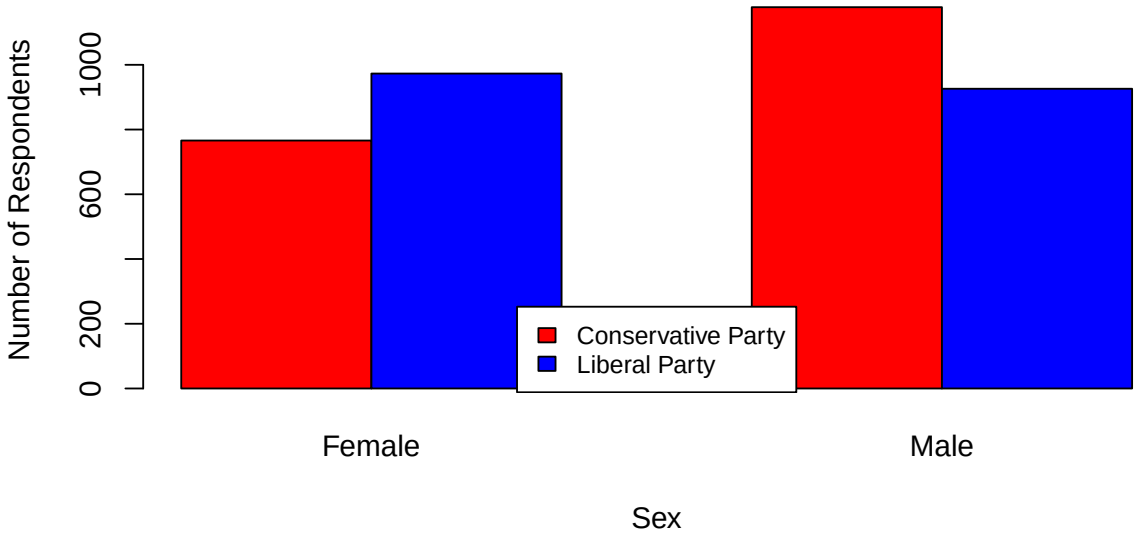
2.1 CES Survey Data

The survey data used to train our model for this report is the subset data obtained from the 37,822 observations from the results of online survey data provided by the 2019 Canadian Election Study (CES). We do not consider the surveys conducted in the previous years since we want to focus on the 2019 Canadian Federal Election specifically. 2019 CES is a survey designed to document the attitudes of Canadians during and after the 2019 election. It provides data on a variety of political and social topics, such as views of democracy, political interest, opinions of current leaders and parties, and topics of interest in the corresponding election cycle. In 2019, the CES is conducted via online and phone-based surveys. In this paper, we will examine the data resulting from the online CES, partially on the factors of sex, education, employment, province, born in Canada, and age group that can have an impact on the vote for the Liberal Party or the Conservative Party.

The target population of the survey includes all Canadian citizens and permanent residents, aged 18 or older, which is exactly the desired population we are interested in. The Campaign Period Survey (CPS) held from September 13th to October 21st, 2019, produce an online sample of 37,822 members of the Canadian general population through Qualtrics platform, with targets stratified by region within Canada and balanced on gender and age within each region with an aim for 50% men and 50% women and an aim for 28% respondents aged 18-34, 33% aged 35-54 and 39% aged 55 and higher[Stephenson et al. 2020]. To be noticed, the survey target was increased during the last five days of the campaign for increasing the total number of respondents. The the Post-Election Survey (PES) held from October 24th to November 11th, 2019, that re-contacted 10,340 respondents from the CPS for a follow-up survey. And, the survey instrument was also presented on the Qualtrics online platform. The weights of the survey sample have been created for the dataset using an iterative "raking" process, as provided by the ipfraking command in STATA15[Stephenson et al. 2020].

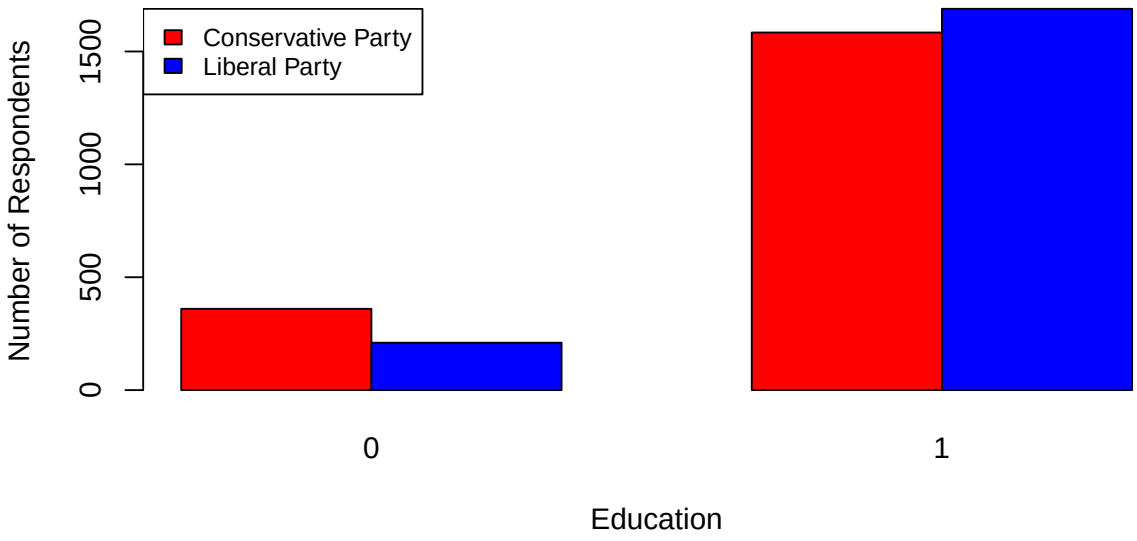
A subset of the online 2019 Canadian Election Study (CES) dataset is selected and retrieved for this report. We use the subset data to build our model based on the respondents' vote choice as the dependent variable and their demographic background as explanatory variables. Key demographic divisions reflected in the contemporary Canadian political landscape are what we are focusing on in our model. The subset is cleaned by firstly selecting interested variables, and then renaming some variables to keep consistency with the census data. We also reconstruct many variables into binary or categorical variables for a more convenient model building later. Vote choice is the variable that only exists in the survey dataset indicating the respondent's vote for Liberal Party or Conservative Party only, which are the main two competitive parties we are focusing on. The other six explanatory variables are included in both survey data and post-stratification data. Figures 1-6 display the distribution of the six reconstructed variables from survey data by the party preference.

Figure 1: Respondent Vote Choice in 2019 CES by Sex



Sex is a binary variable by "Male" and "Female". According to Figure 1, females prefer Liberal better than Conservative, but males prefer the other way and support the Conservative.

Figure 2: Respondent Vote Choice in 2019 CES by Education



Education variable is a binary variable containing two groups of respondents: people who have received some college education(1) and those who have not(0). Similar to sex, these two groups of people also prefer different parties, that people with a higher educational level vote for Liberal more with a small difference via Figure 2.

Figure 3: Respondent Vote Choice in 2019 CES by Employment

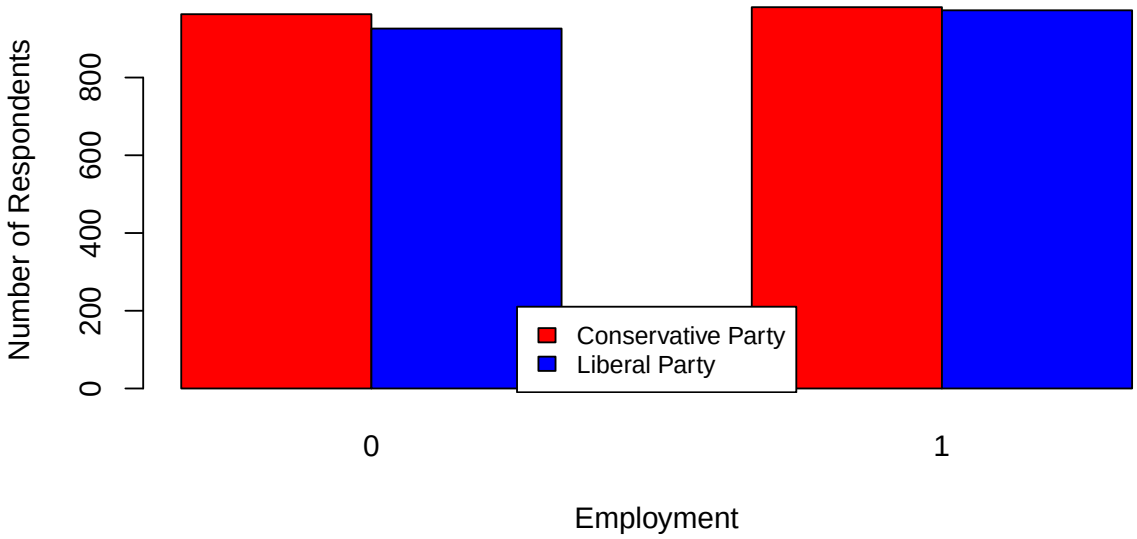
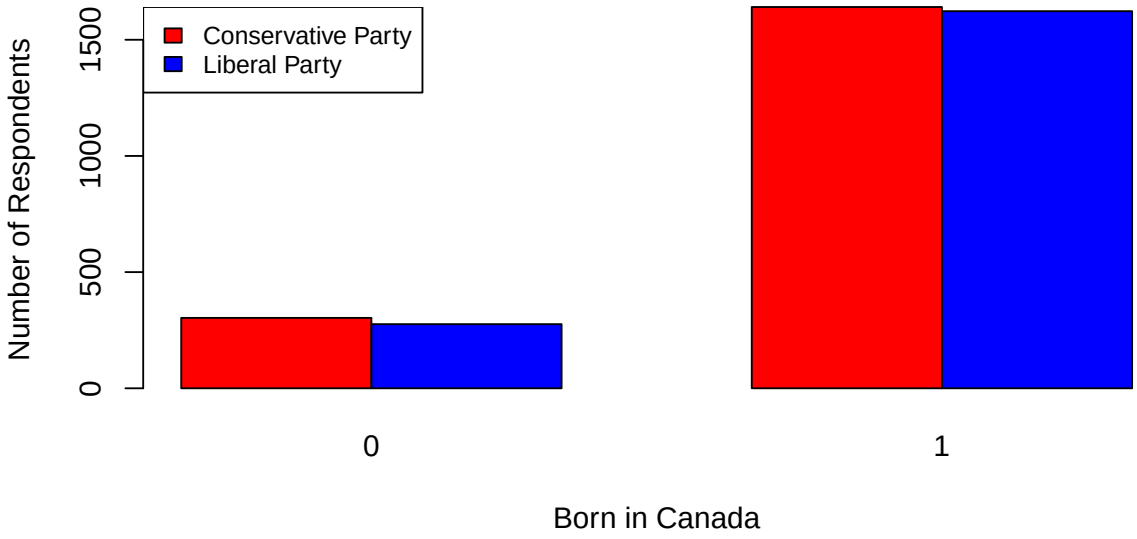
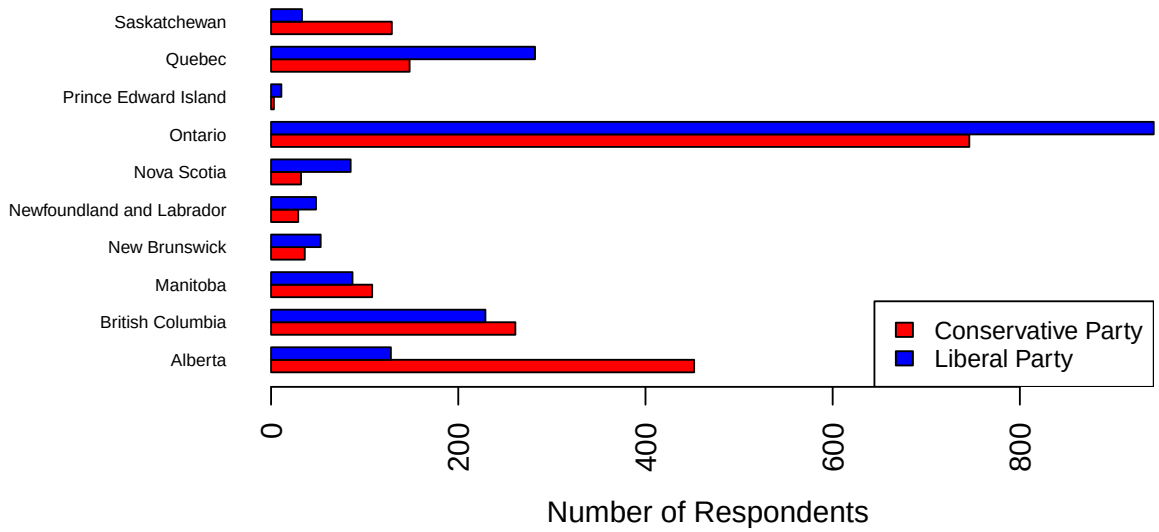


Figure 4: Respondent Vote Choice in 2019 CES by Born in Canada



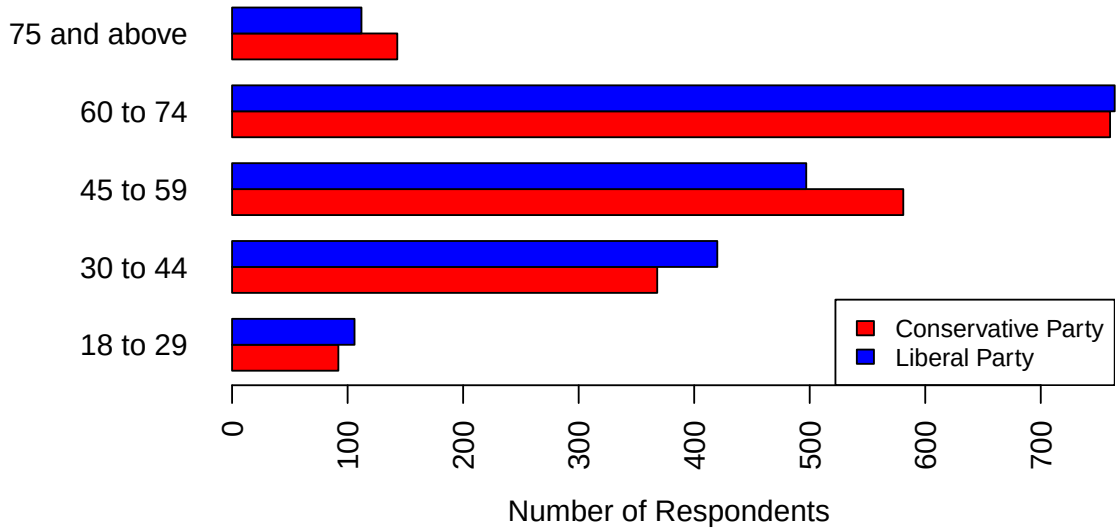
Moreover, employment and born in Canada are two variables also constructed into binary variables. For employment, 1 represents respondents who are employed for pay, and 0 otherwise. For born in Canada, 1 indicates "yes", and 0 indicates "no". Interestingly, there is not much difference in the preference of party between the two groups for both variables, that a very little more vote goes for Conservative Party in all four groups.

Figure 5: Respondent Vote Choice in 2019 CES by Provinces



Province has also shown to be associated with voter preference with our survey data. According to Figure 5, more than half of people from Alberta and Saskatchewan more likely to vote for the Conservative, but more than half of people from Quebec, Prince Edward Island, and Nova Scotia more likely to vote for the Liberal. And, province Ontario with its most population also prefers the Liberal.

Figure 6: Respondent Vote Choice in 2019 CES by Age Groups



Younger voters under 44 years old prefer the Liberal more. Respondents over this age prefer the Conservative more. However, an exception occurs that there is a tie for the 60 to 74 age group.

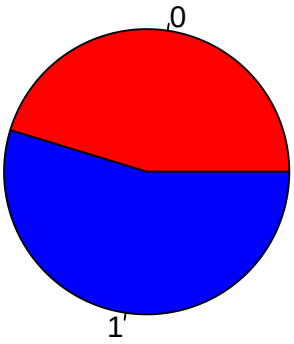
2.2 GSS Census Data

The census data used for post-stratification is the data collected by the 2017 General Social Survey (GSS) on the Family. We use the data from the closest year available to 2019, which is 2017, because we want to focus on the 2019 Canadian Federal Election specifically. GSS is a survey done by the Social and Aboriginal Statistics Division of Statistics Canada, and it is objective to gather data on social trends to monitor changes in the living conditions and well-being of Canadians over time, and then to provide information on specific current social policy issues or emerging interest. In this paper, we will examine the data resulting from the 2017 GSS, partially on the factors of sex, education, employment, province, born in Canada, and age group that can have an impact on the vote for the Liberal Party or the Conservative Party.

The data for the GSS is collected from February 2 to November 30, 2017[GSS 2017]. The target population is non-institutionalized people aged 15 and older, living in the ten provinces[GSS 2017]. To be noticed, people from Yukon, Northwest Territories, and Nunavut are excluded from the dataset, and this can cause bias in the final results. Moreover, this survey is further reduced by it's chosen frame. It uses a new frame that combines telephone numbers (landline and cellular) with Statistics Canada's Address Register, and collects data via telephone[GSS 2017], which excludes people who do not have phones or do not have a stable address to be sampled. The frame is sampled by cross-sectional design, where geographical regions are split into strata. Each strata was then randomly sampled without replacement. However, non-sampling errors occur as a weakness of the survey, when the respondent did not understand or misinterpreted a question, refused to answer a question, or could not recall the requested information. Finally, the target sample size is 20,000 while the actual number of respondents is 20,602. The overall response rate for the 2017 GSS was 52.4%, and non-response was handled by adjusting the weight of households who responded to the survey to compensate for those who did not respond[GSS 2017].

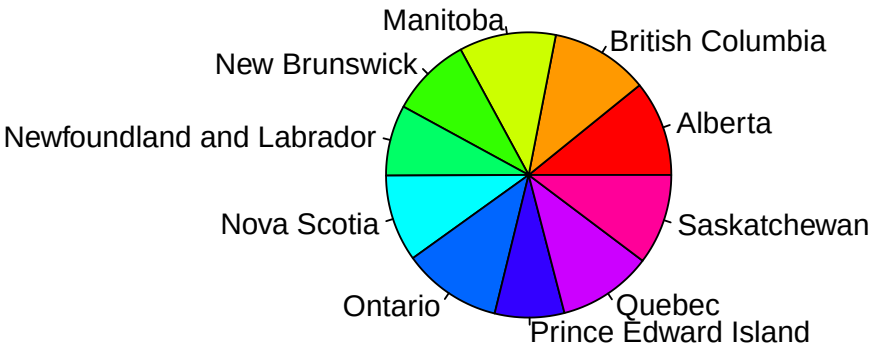
A subset of the 2017 General Social Survey (GSS) on the Family dataset is obtained for this report, which includes selected demographics and retrieved to post-stratify. The subset is cleaned by firstly selecting interested variables that match with the selected variables from survey data, then renaming some variables to keep consistency with the survey data. The selected variables contain sex, education, employment, province, born in Canada, and age group, which are all the factors that can be used to determine the vote for the Liberal Party or the Conservative Party. Similar to survey data, for the census data, we also reconstruct many variables into binary or categorical variables to fit the model later. Figures 7-12 display the distribution of the six reconstructed variables from census data. Since the graphs of binary variables sex, education, and employment show close proportion, please refers to appendix part 2 for Figures 7-9.

Figure 10: Born in Canada Proportion in 2017 GSS



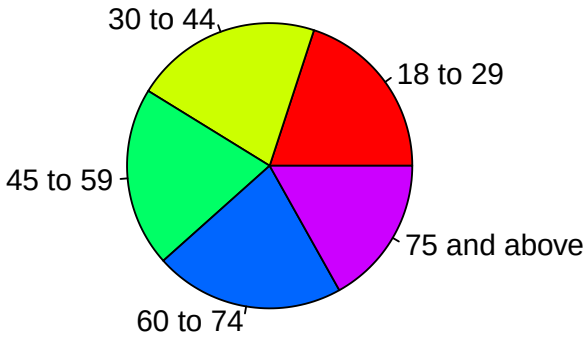
Born in Canada is a binary variable where 1 indicates "yes", and 0 indicates "no". According to Figure 10, there are many more people who are born in Canada. However, refer to Figure 4, we know both groups show equal like for both parties. Then this large difference proportion may not have any strong effect on the vote outcome.

Figure 11: Province Proportion in 2017 GSS



By the province distribution plot, we can see people from Ontario and Quebec take up the most proportion(nearly half), then people from other provinces split the rest of the half proportion, and people from Prince Edward Island take up the least.

Figure 12: Age Proportion in 2017 GSS



As shown by the pie plot of age groups, people aged between 30 to 74 take up over 3/4 of the entire respondents and they are equally divided. Then the rest of the respondents come from the youngest people aged between 18 to 29, and the eldest people aged 75 and above, and they are also equally divided. Hence, the main component for the final result is the votes from the middle ages.

3 Model

3.1 Multilevel Logistic Regression Model

We are interested in determining the popular vote outcome for the Conservative Party for the 2019 Canadian Federal Election by using a multilevel logistic regression based on the factors of sex, education, employment, born in Canada, and age group for different intercept for each province.

Multilevel logistic regression model:

$$\log\left(\frac{\hat{p}_{ij}}{1-\hat{p}_{ij}}\right) = \beta_0 + \beta_1 x_{ij}^{sexMale} + \beta_2 x_{ij}^{educationed} + \beta_3 x_{ij}^{employed} + \beta_4 x_{ij}^{bornCA}$$

$$+\beta_5 x_{ij}^{age30-44} + \beta_6 x_{ij}^{age45-59} + \beta_7 x_{ij}^{age60-74} + \beta_8 x_{ij}^{age75above}$$

where $i = 1, ..., 711$(number of observations), $j = 1, ..., 10$(number of provinces),

p represents the probability of voting for the Conservative Party,

$\frac{p}{1-p}$ represents the odds of voting for the Conservative Party,

β_1 coefficient represents the change in log odds for male respondents,

β_2 coefficient represents the change in log odds for respondent who has received some college education,

β_3 coefficient represents the change in log odds for respondent who is employed for pay,

β_4 coefficient represents the change in log odds for respondent who was born in Canada,

β_5 coefficient represents the change in log odds for respondent age between 30 to 44,

β_6 coefficient represents the change in log odds for respondent age between 45 to 59,

β_7 coefficient represents the change in log odds for respondent age between 60 to 74,

β_8 coefficient represents the change in log odds for respondent age from 75 and above,

x represents each factor respectively.

The general aim of multilevel logistic regression is to estimate the odds that an event will occur (the yes/no outcome) while taking the dependency of data into account. Practically, it will allow you to estimate such odds as a function of lower level variables, higher level variables, and the way they are interrelated (cross-level interactions). Specifically, a multilevel logistic regression can be used when the outcome variable describes the presence/absence of an event or a behavior[Sommet 2017].

We considered using a linear regression model, but finally we kept with logistic regression model. As in our situation, the response variable, whether vote for the Conservative Party, is a binary variable where 1 represent vote for the Conservative and 0 represent vote for the Liberal. Thus, it is more appropriate to use a logistic regression model in predicting the popular vote outcome for the Conservative Party for the 2019 Canadian Federal Election. Moreover, layers were added to our model, so we can get different intercepts for each state in order to make a preciser determination.

Whereas linear regression gives the predicted mean value of an outcome variable at a particular value of a predictor variable, logistic regression gives the conditional probability that an outcome variable equals one at a particular value of a predictor variable (e.g. the likelihood of vote for the Conservative for a male respondent aged at 35 who has received some college education, and he was born in Canada, and also he is employed for pay).

The expression on the left hand side of the equation is often called logit function, it is used to predict such a probability. In our model, it describes the relationship between a series of explanatory variables(sex, education, employment, born in Canada, and age group) and the conditional probability that an outcome variable Y_i equals one(vote for the Conservative). Also, a multilevel regression is used to smooth noisy estimates in the cells with too little data by using overall or nearby averages[Multilevel regression with poststratification 2020].

We use `glmer()` from `lme4` package in `R` to fit the model to our data. We use `as.factor()` for age variable, because even though age is numerical in the original dataset, but it becomes categorical as we group them into age groups during the data cleaning process. For each categorical variable(sex, education, employment, born in Canada, and age group) with n levels, we need $n-1$ dummy variables to fully study its influence on our response variable(vote for the Conservative).

3.2 Post-stratification Calculation

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

where $\hat{y_j}$ is the estimate in each cell and $\sum N_j$ is the population size of j^{th} cell based off demographics.

In order to estimate the proportion of voters who will vote for the Conservative Party, we performed a post-stratification analysis. We make predictions using our model above with census data from 2017 GSS, specifically estimate y from each cell using our multilevel model, meaning use demographics to extrapolate how entire population will vote.

Response recorded basic demographics: age(5 categories), gender(2 categories), education(2 categories), born in Canada(2 categories), employment(2 categories), thus partitioning the data into 80 cells.

We weight each proportion estimate by the respective population size and sum those values and divide by the entire population size. The post-stratification weights are a sophisticated weighting strategy that help to reduce sampling error and potential non-response bias[ESS Methodology. n.d.].

4 Results

4.1 Model Result

Multilevel logistic regression model estimates by interpreting regression coefficients:

$$\log\left(\frac{\hat{p}_{ij}}{1-\hat{p}_{ij}}\right) = 0.37 + 0.57x_{ij}^{sexMale} - 0.66x_{ij}^{educationed} + 0.01x_{ij}^{employed} - 0.24x_{ij}^{bornCA}$$

$$-0.05x_{ij}^{age30-44} + 0.24x_{ij}^{age45-59} - 0.02x_{ij}^{age60-74} + 0.14x_{ij}^{age75above}$$

Table 1: Summary of Model Estimates

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.3663	0.3307	1.1077	0.2680
sexMale	0.5684	0.0708	8.0252	0.0000
edu	-0.6606	0.1006	-6.5689	0.0000
employed	0.0103	0.0835	0.1229	0.9022
bornin_canada	-0.2385	0.0967	-2.4663	0.0137
as.factor(age)30 to 44	-0.0488	0.1696	-0.2877	0.7736
as.factor(age)45 to 59	0.2416	0.1654	1.4610	0.1440
as.factor(age)60 to 74	-0.0184	0.1686	-0.1092	0.9131
as.factor(age)75 and above	0.1446	0.2122	0.6813	0.4957

4.2 Post-stratification Calculation Result

Errors exist in the code for getting vote outcome. To be fixed.

5 Discussion, Limitations, and Future Work

Voter participation rates are symbolic of the health of a democracy. Research shows that when communities have strong associations, they are in fact more likely to participate in national and local politics.[How can we encourage more Canadians to vote? n.d.] For a country as diverse as Canada, that's a sure win for everyone.

Appendix

1. Github link which contains all the code, dataset(except for original 2019 CES and 2017 GSS data, method to download is attached in readme.md), and report for the project: https://github.com/ma521yyy/Difference-in-2019-Canadian-Federal-Election-if-Everyone-had-Vote
2. Figures 7-9 display the distribution for binary variables of sex(femle, male), education(1 indicates some college education, 0 otherwise), and employment(1 indicates employed for pay, 0 otherwise) from census data.

Figure 7: Sex Proportion in 2017 GSS

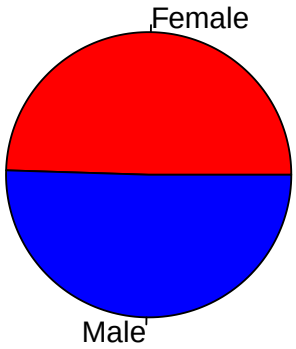


Figure 8: Education Proportion in 2017 GSS

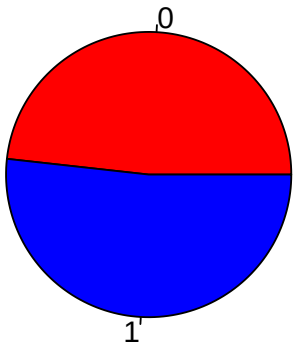
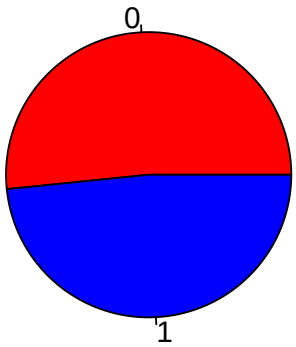


Figure 9: Employment Proportion in 2017 GSS



References

Depner, W. (2020, March 03). Political apathy main reason for not voting in 2019 Canadian federal election. Retrieved from https://www.vicnews.com/news/political-apathy-main-reason-for-not-voting-in-2019-canadian-federal-election/

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

ESS Methodology. (n.d.). Retrieved from https://www.europeansocialsurvey.org/methodology/ess_methodology/data_processing_archiving/weighting.html#:~:text=Post-stratification weights are a,gender, education, and region.

GSS. (2017). General Social Survey – Family (GSS). Retrieved from https://www.statcan.gc.ca/eng/survey/household/4501

Government of Canada, S. C. (2020, February 26). Reasons for not voting in the federal election, October 21, 2019. Retrieved from https://www150.statcan.gc.ca/n1/daily-quotidien/200226/dq200226b-eng.htm

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

Hadley Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016, https://ggplot2.tidyverse.org

Hadley Wickham, Jim Hester and Romain Francois (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr

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

Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. http://haozhu233.github.io/kableExtra/, https://github.com/haozhu233/kableExtra.

Home. (1970, October 01). Retrieved from https://www.elections.ca/content.aspx?section=res&dir=rec/eval/pes2019/lfs&document=index&lang=e

How can we encourage more Canadians to vote? (n.d.). Retrieved from https://cnmc.ca/how-can-we-encourage-more-canadians-to-vote/

Multilevel regression with poststratification. (2020, October 14). Retrieved from https://en.wikipedia.org/wiki/Multilevel_regression_with_poststratification

Ouellet, Andre Real, "The Democracy Defibrillator: The Decline of Canadian Voter Turnout in Federal Elections, and Suggestions for Revitalisation" (2019). Major Papers. 77. Retrieved from https://scholar.uwindsor.ca/major-papers/77

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

Rohan Alexander and Sam Caetano (2020). "GSS_Cleaning" Retreived from https://q.utoronto.ca/courses/184062

Sommet, N., & Morselli, D. (2017). Keep Calm and Learn Multilevel Logistic Modeling: A Simplified Three-Step Procedure Using Stata, R, Mplus, and SPSS. International Review of Social Psychology, 30(1), 203–218. DOI: http://doi.org/10.5334/irsp.90

Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, "2019 Canadian Election Study - Online Survey", https://doi.org/10.7910/DVN/DUS88V, Harvard Dataverse, V1

Wikipedia contributors. (2020, December 8). 2019 Canadian federal election. In Wikipedia, The Free Encyclopedia. Retrieved 15:18, December 9, 2020, from https://en.wikipedia.org/w/index.php?title=2019_Canadian_federal_election&oldid=993089856

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30, https://yihui.org/knitr/