Data Analysis for Predicting the next Canada Election

STA304 - Assignment 2

GROUP 50: Yaqian Han, Henry Lu, William Wu, Yiyao Zhang

November 24, 2022

Introduction

In this research, the goal is that we aim to predict the proportion of Liberal Party, Conservative party and New Democratic party's votes in the next election using a census data and a survey data. The 2017 GSS census data 'gss_clean' [6] and 2019 CES phone survey data 'ces2019-phone_clean' [12] include numerical and categorical variables that describe the essential information about the voters in the election, such as the conjugal and parental history, family origins, and other socioeconomic characteristics of all non-institutionalized people with over 15 years of age, residing in the 10 provinces of Canada [16]. The analysis will focus on four certain predictor variables "birth place", "education level", "age", and "household size" to investigate and predict the response variable "vote liberal party" based on the survey data. The variables the analysis includes have demonstrated basic characteristics of the voters in the election. Therefore, the result of this analysis can be applied to predict the future election based on the four variables.

The analysis is unique in the categories of the variables and modeling. The variables "education level" and "birth place" were first collected from the original survey and census dataset. Then our team has transformed their categories into new ones in order to model and calculate the result by poststratification.

Today, the public opinion of Justin Trudeau has worsened as Canada struggles with record high inflation [3]. This is also the main problem of this research. The disapproval rate of the government in Canada has reached its highest rate of 51% since Prime Minister Justin Trudeau was elected in 2015 [3]. With this ongoing trend, there is a high probability that the Liberal Party will not be re-elected in 2025. Since different political parties have very different social and economic views, the result of the next Canadian election will likely have important political, economical and social implications, such as universal healthcare and income tax. [3] There are five major political parties in Canada: Liberal Party, Conservative Party, New Democratic Party(NDP), Bloc Quebecois, and Green Party of Canada. The Liberal Party, Conservative Party and the New Democratic Party are the focus of this research.

We selected 5 predictor variables to estimate the log proportion of the Liberal Party votes: voter age, voter household size, voter birthplace, voter education level. We selected these predictors because previous research has shown found relationship between these variables and their voting behavior. According to research, age has an impact on voter behavior because as voter ages, they are more likely to have higher income, causing them to vote in favor of the Conservative Party, which sets lower tax for high-income groups [2]. Household size and birthplace influence voting behavior as voters with larger household and votes that are immigrants have a lower voting rate [5] . Voters that are more education tend to lean more towards the Liberal Party [13].

We will be using post-stratification to adjust the weights of each sub-population so that the overall sample is more representative of the true distribution of the sub-populations in the target population.

The research question is "What are the proportions of the Liberal Party, Conservative Party, and New Democratic Party votes in the next election?". With the current results of the seat projections, it is predicted that the Conservatives will win 108 seats if the election was held today as opposed to 106 seats for the Liberals [3]. Our hypothesis is that the Conservatives will have the majority of votes in the next election.

Data

Our census data, 2017 GSS data 'gss_clean', is originated from CHASS [16], a collection of online databases managed by the University of Toronto. The 2017 GSS data collects the information in 2017 by telephone interview on the conjugal and parental history, family origins, and other socioeconomic characteristics of all non-institutionalized people with over 15 years of age, residing in the 10 provinces of Canada [6].

Our survey data, 2019 CES phone survey data 'ces2019-phone_clean', was performed to gather the opinions of 641 Canadian voters after the 2019 federal election. It is part of a series of Canadian Election Studies that started in 1965 [12]. We loaded the data from the cesR package after installing the devtools package directly in Rstudio. We selected 5 variables from the dataset for our research: q2(vote age), q61(educational level), q64(voter birthplace), q70 (voter household income), and q71 (vote household size).

Since the common variables between the two datasets do not have exactly the same outcomes, we adjusted the outcomes of the variables in each dataset so that they are comparable between the two datasets. The steps are as follows: -We first mutated the name of column q64 in the survey data (CES) to 'birth_place'. We adjusted the outcome so that when original oucome is 'Canada', the new outcome becomes '1'; when the original outcome is not 'Canada', the new outcome becomes '0'. -Next, we mutated the column q61 in the CES to 'education' . It has the outcomes of "1" corresponding to the original outcome of "9" in column q61, "2" corresponding to the original outcome of "6" and "7" in column q61, "3" corresponding to the original outcome of "4" and "5" in column q61, "4" corresponding to the original outcome of "1", "2" and "3" in column q61, "5" corresponding to the original outcome of "8" in column q61, and "6" corresponding to the original outcome of "10" and "11" in column q61. -Further, we mutated two variables 'age' and 'hh_size'. 'age' collects the data where we use 2019 to minus the data in column q2. 'hh_size' collects the data from column q71. -At last, we select these variables and remove all the missing values.

In the census dataset: -we first create a new variable called 'birth_place'. It collects the data in column place_birth_canada in census data and has the outcome of "1" when the outcome of place_birth is "Born in Canada", and has the outcome of "0" when the outcome of place_birth is not "Born in Canada". -Next, we have mutated a variable called education. It collects the data in column education and has the outcomes of "1" corresponding to the original outcome of "Bachelor's degree (e.g. B.A., B.Sc., LL.B.)" in column education, "2" corresponding to the original outcome of "College, CEGEP or other non-university certificate or di. . . " in column education, "3" corresponding to the original outcome of "High school diploma or a high school equivalency certificate" in column education, "4" corresponding to the original outcome of "Less than high school diploma or its equivalent" in column education, "5" corresponding to the original outcome of "University certificate or diploma below the bachelor's level" in column education, and "6" corresponding to the original outcome of "Trade certificate or diploma" in column education. -Furthermore, we mutated the variable called age. age collects the data where we round up the data in column age. -At last, we select these variables and remove all the missing values.

The model has three response variables: vote_lib, vote_con and vote_ndp. All the three variables have the same outcome meanings: if the outcome is "1", it means that the voter has voted for the party; if the outcome is "0", it means that the voter has not voted for the party.

The model has four important predictor variables: birth_place, education, age, and hh_size. The variables are common in both census data and survey data. In the census dataset, the data was collected in 2017, while in the survey dataset, the data was collected in 2019.

Birth place represents the country where the respondent was born and its outcome is categorical. The outcome is binary where "1" means the respondent was born in Canada and "0" means he/she was born somewhere else. Education represents to what extent the respondent was educated, where its outcome has six categories. "1" means Bachelor's degree, "2" means College degree, "3" means High School degree, "4" means less than a High School degree, "5" means University degree and "6" means others. Age represents the age of the respondent, whose outcomes are numerical. And the hh_size represents the household size of the respondents, which is the number of members in the household that the respondent was in, whose outcomes are also numerical.

Table 1: Table 1: Numerical summary

Variables	Min	Max	Mean	sd	Q_1	Q_3
age census	15.00	80.00	52.23	17.9602	37.00	67.00
hh_size census	1.000	6.000	2.336	1.2524	1.000	3.000
age survey	16.00	98.00	49.81	19.0084	35.00	65.00
hh_size survey	1.000	7.000	2.638	1.3613	2.000	3.000

In Table 1, we compare the age and household size in census data and the survey data. First, the minimum age is almost the same in both sets of data. However, the maximum age has a significant difference: the maximum age in the census data is 80 and the maximum of age in survey data is 98. The mean of age in the census data is 52.23 which is a little bit larger than 49.81 in the survey data. There is also a certain difference in standard deviation of age, 17.9602 in census data and 19.0084 in the survey data. Furthermore, we compare the household size of these two data. The minimum household size is the same and the maximum household size in the survey data is 7 which is larger than 6 in the census data. The means of the household sizes are similar in census data and survey data, with 2.336 and 2.638 respectively. The standard deviations of household size are similar, 1.2524 in the census data and 1.3613 in the survey data.

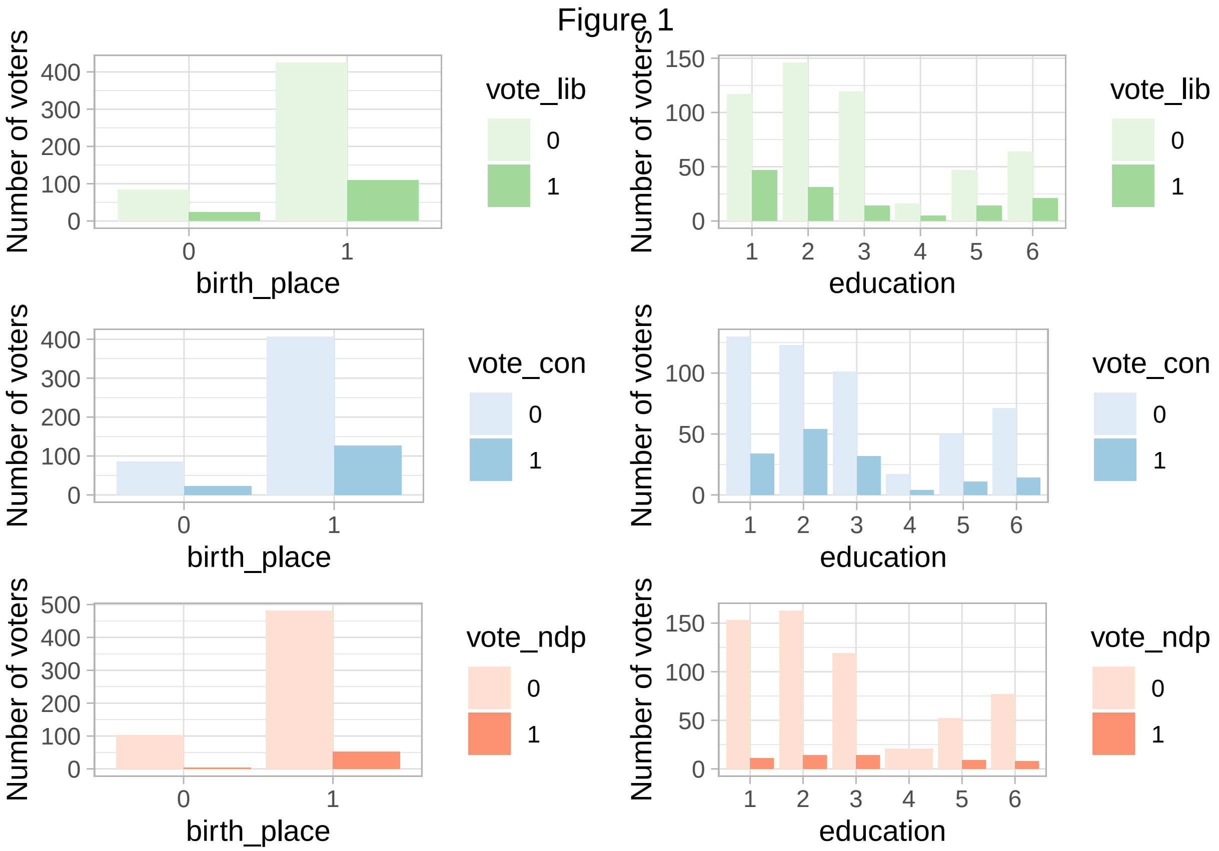
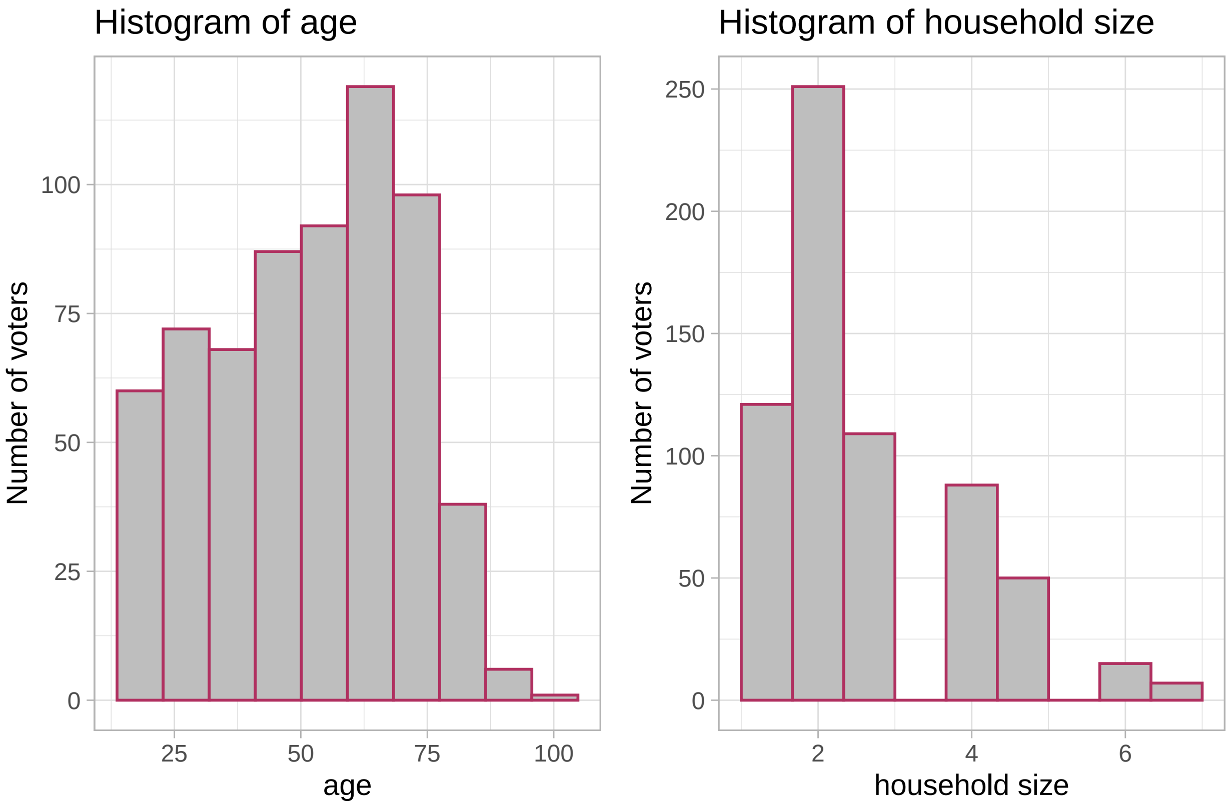


Figure 1 has demonstrated three pairs of histograms displaying birth_place and education. The first column of histograms display the distribution of birth_place versus vote_lib, vote_con and vote_ndp, implying the number of voters that were born or not born in Canada. The second column of histograms display the distribution of education versus vote_lib, vote_con and vote_ndp, implying the number of voters' education level.

Figure 2


Histogram of age

Histogram of household size

Figure 2 has demonstrated two histograms for checking the normality of the two variables age and hh_size. The left histogram has displayed the distribution of the age among the voters. Among 641 observations, the range of age is from 15 to 100. The histogram is approximately left-skewed and the mode is around 65, the median is around 50 and the mean is around 60 by only observing the graph. It is roughly symmetric and not significantly normally distributed. There is one extreme value at 100, which needs more analysis to determine if the outlier is rational or needs to be eliminated. It can be deduced that most of the voters have aged approximately from 50 to 70. The right histogram has displayed the distribution of the household size among the voters. Among 641 observations, the range of size is from 0 to 7. The histogram is approximately right-skewed and the mode is around 3, the median is around 4 and the mean is around 5 by only observing the graph. It is not significantly normally distributed or symmetric. There is one extreme value at 7, which needs more analysis to determine if the outlier is rational or needs to be eliminated. It can be deduced that most of the voters have household size approximately from 2-5.

Methods

The goal of this research is to find the proportion of the Liberal Party, Conservative Party, and New Democratic Party votes in the next election, to achieve this, we are going to use a combination of multi-logistic regression model and post-stratification. Using the logistic regression model we will find the proportion of survey participants that votes for the party of interest for every bin/cell of the sample. Once we have our estimated proportions, we will use post-stratification which is a technique that will improve the efficiency of our estimators. Through post-stratification, the survey weight will adjust the estimated numbers of participants in each of a set of estimation cells to be equal to the known population totals from our census data.

Model Specifics

The research will be using three separate multi-logistic regression model to model the proportion of survey participants that will vote for each of the following political parties: The Liberal Party, The Conservative Party, and The New Democratic Party. Each of the three models will have a different response variable that correspond to the party of interest, but they will all share the same predictor variables. We chose to use the multi-logistic regression model because the predicted outcome is a dichotomous response variable(vote_lib, vote_con, or vote_ndp). The outcome of the following model is the exprected log odd of the party being voted, in which the predictor variable used are two categorical variables(birth_place and education) and two numeric variable(age and hh_size). The multi-logistic regression model we are using is:

$$log(\frac{p}{1-p}) = \beta_0+\beta_1 x_{age}+\beta_2 x_{hh.size}+\beta_3 x_{birthplace1}+\beta_4 x_{education2}+\beta_5 x_{education3}+\beta_6 x_{education4}+\beta_7 x_{education5}+\beta_8 x_{education6}$$

p is the expected probability that the outcome where 1 indicates that the outcome of interest is present, and 0 indicates that the outcome of interest is absent. For our model, the probability would represent the proportion of voters that voted for the party of interest.

For predictors x, each numeric predictors x is the numeric value of the variable, and each categorical predictors x are used as an indicator variables(also known as dummy variable) which takes the value of 0 or 1 to indicate the presence or absence of some categorical effect that may be expected to affect the outcome.

For the coefficients, β_0 represents the expected log proportion of each party's votes when all x's are 0. β_1 represents the expected change in log proportion of each party's votes relative to a one unit change in participant's age,holding all other predictors constant. β_2 represents the expected change in log proportion of each party's votes relative to a one unit change in participant's house hold size,holding all other predictors constant. β_3 represents the expected change in log proportion of each party's votes if the participant is born in Canada, holding all other predictors constant. β_4 represents the expected change in log proportion of each party's votes if the participants' highest education is some/completed community college, holding all other predictors constant. β_5 represents the expected change in log proportion of each party's votes if the participants' highest education is some/completed high school, holding all other predictors constant. β_6 represents the expected change in log proportion of each party's votes if the participants' highest education is no/some/completed elementary school, holding all other predictors constant. β_7 represents the expected change in log proportion of each party's votes if the participants' highest education is some university, holding all other predictors constant. β_8 represents the expected change in log proportion of each party votes if the participants' highest education is master degree/professional degree/doctorate, holding all other predictors constant.

```
# Creating the Model
model_lib <- glm(as.factor(vote_lib) ~ age + hh_size + as.factor(birth_place) + as.factor(education),
                     data = survey_data,
                     family=binomial)

summary(model_lib)
```

```
##
## Call:
## glm(formula = as.factor(vote_lib) ~ age + hh_size + as.factor(birth_place) +
##     as.factor(education), family = binomial, data = survey_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9182  -0.7552  -0.6141  -0.4403   2.2351
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.082164   0.541609  -1.998 0.045711 *
## age                      0.004855   0.005960   0.815 0.415291
## hh_size                 -0.046992   0.083964  -0.560 0.575704
## as.factor(birth_place)1  0.051375   0.265658   0.193 0.846656
## as.factor(education)2   -0.638813   0.264786  -2.413 0.015841 *
## as.factor(education)3   -1.229278   0.332858  -3.693 0.000222 ***
## as.factor(education)4   -0.369309   0.549440  -0.672 0.501485
## as.factor(education)5   -0.299648   0.352312  -0.851 0.395036
## as.factor(education)6   -0.210552   0.305890  -0.688 0.491247
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 651.91  on 640  degrees of freedom
## Residual deviance: 632.34  on 632  degrees of freedom
## AIC: 650.34
##
## Number of Fisher Scoring iterations: 4
```

```
model_con <- glm(as.factor(vote_con) ~ age + hh_size + as.factor(birth_place) + as.factor(education),
                     data = survey_data,
                     family=binomial)

summary(model_con)
```

```
##
## Call:
## glm(formula = as.factor(vote_con) ~ age + hh_size + as.factor(birth_place) +
##     as.factor(education), family = binomial, data = survey_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3038  -0.7708  -0.6342  -0.4068   2.4007
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -3.474292   0.598730  -5.803 6.52e-09 ***
## age                      0.026703   0.006223   4.291 1.78e-05 ***
## hh_size                  0.196041   0.082671   2.371   0.0177 *
## as.factor(birth_place)1  0.179416   0.269619   0.665   0.5058
## as.factor(education)2    0.621886   0.259275   2.399   0.0165 *
## as.factor(education)3    0.235732   0.285682   0.825   0.4093
## as.factor(education)4   -0.259355   0.604315  -0.429   0.6678
## as.factor(education)5   -0.134291   0.392081  -0.343   0.7320
## as.factor(education)6   -0.339988   0.354810  -0.958   0.3379
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 695.12  on 640  degrees of freedom
## Residual deviance: 665.63  on 632  degrees of freedom
## AIC: 683.63
##
## Number of Fisher Scoring iterations: 4
```

```
model_ndp <- glm(as.factor(vote_ndp) ~ age + hh_size + as.factor(birth_place) + as.factor(education),
                     data = survey_data,
                     family=binomial)

summary(model_ndp)
```

```
##
## Call:
## glm(formula = as.factor(vote_ndp) ~ age + hh_size + as.factor(birth_place) +
##     as.factor(education), family = binomial, data = survey_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9179  -0.4597  -0.3326  -0.2346   3.1316
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.541496   0.823038  -1.873   0.0611 .
## age                      -0.040670   0.008835  -4.603 4.16e-06 ***
## hh_size                   0.002410   0.107354   0.022   0.9821
## as.factor(birth_place)1   0.993907   0.543095   1.830   0.0672 .
## as.factor(education)2    -0.024456   0.429680  -0.057   0.9546
## as.factor(education)3     0.238047   0.436586   0.545   0.5856
## as.factor(education)4   -14.353665 820.376294  -0.017   0.9860
## as.factor(education)5     0.623444   0.498446   1.251   0.2110
## as.factor(education)6     0.523310   0.498197   1.050   0.2935
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 379.98  on 640  degrees of freedom
## Residual deviance: 340.71  on 632  degrees of freedom
## AIC: 358.71
##
## Number of Fisher Scoring iterations: 16
```

Checking Logistic Regression Model Assumptions

In this research, we will be using logistic regression because the response variables ‘vote_lib’, ‘vote_con’, and ‘vote_ndp’, which represent whether the voter voted for the Liberal Party, Conservative Party, or the New Democratic Party, for our models have binary outcomes (1 or 0).

The logistic regression model has 4 assumptions to be satisfied: outcome is binary, linearity between continuous predictor variables with the response variable, absence of multicollinearity between predictors, and absence extreme outliers.

Assumption 1: Binary response variable The first assumption is satisfied as the response variables - vote_lib, vote_con, and vote_ndp - in each of our three models have binary outcomes ‘1’ or ‘0’.

Assumption 2: Linear relationship between logit of the outcome and each predictor variables The second assumption ‘linearity between continuous predictors variables with the response variable’ is satisfied and does not need checking because we will be working with only discrete predictor variables in our models.

Assumption 3: Multicollinearity between each predictor variable

The third assumption ‘absence of multicollinearity between predictors’ is satisfied as all predictors in each model have a Variance Inflations Factor (VIF) lower than 5.

Assumption 4: Absence of extreme outliers

The fourth assumption ‘absence of extreme outliers’ is satisfied in the models ‘model_lib’ and ‘model_con’. ‘model_ndp ‘ contains one outlier. This is checked by identifying the influential points by visualizing the Cook’s distance values. The top four largest values in each model are labelled. Then, we used the standardized residuals to determine if the data points are outliers. Data points with a standardized residual larger than 3 represent potential extreme outliers. ‘Model_ndp’ contains one data point with a standardized residual of 3.137, which may potential be an extreme outlier.

Post-Stratification

First, we used the survey data to construct the logistic regression model as above. Then, we completed the random data sampling and stratified the data (in accordance with their birth place, education level, age, and household size) to make sure that the data is representative of the population that aims at conducting research and concluding from. The combination of predictor variables is called “cell”. Next, to estimate response variables for each cell, we applied this logistics regression model into cells to predict the proportion of survey participants that will vote for each of the following political parties: The Liberal Party, The Conservative Party, and The New Democratic Party. The estimate in each cell is denoted as $\hat{y}_j$. In our dataset, $\hat{y}_j$ for The Liberal Party is estimate_lib, $\hat{y}_j$ for The Conservative Party is estimate_con, and $\hat{y}_j$ for The New Democratic Party is estimate_ndp. For these three parties, they have the same size of population, denoted as N. Then, we calculated the weighted average: combine the estimated proportion for each cell-level up to a population level estimate by weighting each cell so that the weighted totals within mutually exclusive cells equal to the known population.[11] Finally, we got the adjusted estimates $\hat{y}^{PS}$ for these three parties. Here, in our dataset, $\hat{y}^{PS}$ for The Liberal Party is denoted by result_lib, $\hat{y}^{PS}$ for The Conservative Party is denoted by result_con, and $\hat{y}^{PS}$ for The New Democratic Party is denoted by result_ndp.

<In order to estimate the proportion of voters, we ave used this mathematical formula,>

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

<Where $\hat{y}_j$ represents the estimate in each cell, N_j represents the population size of the jth cell based off demographics.>

Stratification is used to classify populations into different groups based on their characteristics or some features before doing analysis [10]. People in the same groups have the same or similar characteristics in the variables we’ll study. However, in some specific situations, for example, when the variables utilized to stratify the population can only be observed after sampling, we’ll choose to use an alternative method, called post stratification.[10] Post stratification is the process of doing a proportional (weighted) adjustment based on the combination of predictor variables. The census data and survey data provides us the population weights. And then we can re-weight the data based on the variables we will analyze in this study: age, household size, whether the participant is born in Canada, the highest education of participants (community college, high school, elementary school, university, and master degree/professional degree/doctorate).

Since the background of every voter is different, sometimes we’ll be affected by the estimates of relatively large population. For example, in Figure 2, the histograms of age and household size for voters imply that they have different distributions. People aged 45-75 and household size of 2 occupy the majority of voters. Furthermore, in Figure 1, it shows that the number of voters has significant difference in whether voters were born in Canada and the education level for these three parties. This means that if we don’t take this effect into consideration, then the estimate we get may cause bias. By doing post stratification, we can reweight each cell by its relative proportion in the population we study to do adjustments. Post stratification can efficiently reduce the influence of strata with a large proportion of population on the overall prediction. It’s useful to correct our sampling and increase the accuracy of estimation for our research question.

<As part of the poststratification technique you should also describe the cell/bin splits that you will display/implement in the Results, based on the sample data. Here you should briefly recall the variables that you are using to create the cells (again, the full description of these should be in the Data section). You can briefly justify the choice to include or exclude certain variables when creating the cells/bins. (For example, choosing “province” because it is likely to influence voter outcome because of. . . , or not including “eye colour” because it is not available in the census data).>

In this research question, we chose The Liberal Party, The Conservative Party, and The New Democratic Party as our main studying objects. Recall the variables that we used to create cells, age, whether their birth places are in Canada or not, education level, and household size. To estimate the proportional of survey participants that will vote for these three parties, we first considered the birth places of voters. If voters are born in Canada, intuitively speaking they will be likely more interested in these three major political parties in Canada compared to voters born elsewhere. For these two numerical variables: age and household size, each voter will make different choices at different ages and household size may refer to the happiness index of each family that possibly affect the outcome to some extent. Therefore, these are also two aspects that will influence the outcome and worth to studying. The distribution of the number of voters of each age group and household size can be conducted separately in the histogram. Finally, we used the education level as the last variable to create the cells since educational qualifications may have a potential impact on elections. For example, population with higher academic qualifications will consider more factors to vote. These above variables are selected when creating the cells and used to do further estimation.

<An explanation of the method for a general science reader (i.e., not a statistician).>

<A description of why the method is appropriate (based off assumptions, variable types and practical rationale).>

<If you want to include some additional analysis (e.g., standard error, poststratification by province, etc.) then you should describe your methodology here. Additionally, if you do this be sure to include any citations/references that may be needed by the reader.>

Results

Hypothesis Testing: Hypothesis testing is used to test what possibility our hypothesis would occur. Thus, We state our null hypothesis and alternative hypothesis, which is the there is no logistic relationship between the response variable and the predictor variables; the alternative hypothesis is that at least one predictor variables is logistically related to the response variable. Then, we operate 2 tail-test to calculate the p-value. P-value is the possibility of obtaining a result, given that the null hypothesis is true. Therefore, if the p-value that we calculate is less than 0.05, we have strong evidence against our null hypothesis and thus the null hypothesis should be rejected.

$$H_0 : \beta_j \ = 0$$
$$H_a : \beta_j \neq 0$$

For the model of the Liberal party, the p value for level 2 of education is 0.015841 and is less than 0.05 and the p value for level 3 of the education level is 0.000222 and is less than 0.01. This means that we have significant evidence to reject the null hypothesis which states that there is no logistic relationship. Thus, there is a logistic relationship between the two levels of education and votes for the Liberal party. For the model of the Conservative party, the p value for age is 1.78e-05 is less than 0.01. The p value for household size is 0.0177 and is less than 0.05. The p value for level 2 of the education level is 0.0165 and is less than 0.05. This means that we have significant evidence to reject the null hypothesis. Thus, there is a logistic relationship between the age, household size and level 2 of education and votes for the Liberal party. For the model of the New Democratic party, the p value for age is 4.16e-06 and is less than 0.01. This means that we have significant evidence to reject the null hypothesis. Thus, there is a logistic relationship between the age and votes for the Liberal party.

Table 2: Table 2: Predicted result summary

Party	Predicted_result
Liberal	0.197041
Conservative	0.220784
New democratic	0.06847475

Table 2 has displayed the results of the predictions from each model. The first column describes each party and the second column describes the predicted probability of the outcome for each party. The result for predicting the probability of the Liberal party winning the election is 0.197041045822495; the result for predicting the probability of the Conservative party winning the election is 0.220783965886662; the result for predicting the probability of the New Democratic party winning the election is 0.0684747538599982. The result means predictions about the probability of the outcome. Therefore, the Conservative party has the highest probability of winning the election, the Liberal party has the second highest probability and the New Democratic party has the lowest probability of winning. Our research suggests that the Conservatives and the Liberals have roughly equal chances of winning the election, with the Conservatives having a slightly higher chance of winning. This supports the predictions found in relevant researches. The seat projection from the researches estimates that the Conservatives will win 108 seats and the Liberals closely follow with 106 seats.

Conclusions

To conclude, the hypothesis is that the Conservatives will likely have the majority of votes in the next election. To prove the hypothesis is correct, we have applied a logistic regression model and poststratification. By modeling the three party using voters' age, household size, education level and birth place, we have predicted the probability of the outcome, where the Liberal party would have the probability of 0.197041 to win the election; the Conservative party would have the probability of 0.220784 and the New Democratic party would have the probability of 0.06847475. Therefore, our hypothesis is proved to be true.

There are also a few key results from the visualizations. From Figure 1, we can also tell that the Conservative party has the most voters based on the birth place and education. In particular, the Conservative party has approximately 125 voters that are born in Canada, which is the highest among the three parties; and it has approximately 85 voters that have education level of Bachelor's Degree and College Degree, which is also the highest. From the model of the Conservative Party, 3 out of 4 predictor variables have logistic correlations to the response variables. From Table 2, it states the key results of the predicted probability of the outcome, proving that the Conservative party has the highest predicted probability of winning the election.

Based on the four variables, it is possible to model and predict the outcome of the probability of winning the election. However, not all the variables can create strong logistic relationships with the response variables. And the probability is quite small, meaning that the result may fluctuate and not be inaccurate.

One of the drawbacks is categorizing the common categorical variable "education" in the survey dataset and the census dataset. In survey data, education was first categorized into 11 categories, while there were 6 categories for education in census data. Although the variable has similar categories in each dataset, the result has occurred error and inaccuracy in rearranging the outcomes into the same categories for modeling. Hence, it has affected the prediction of the probability of the outcome. In addition, as mentioned above, not all variables can create a strong correlation with the response variables. The moderate correlation may lead to errors in predicting the outcomes.

In addition, the distributions of the two numerical variables:age and household size are not perfectly normally distributed. Therefore, in further analysis, we would collect more information from the respondents and try to build a symmetric and normally distributed model for the numerical variables.

In the future analysis, we would search larger datasets that include more valid variables for survey and census, in order to include more predictor variables in modeling and predicting the results. This would eliminate some errors and support the logistic modeling to be more correlated with the response variables. Hence, the results can be more accurate in reperesenting the probability of winning the election.

Bibliography

1. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)

2. BBC. (n.d.). Age - factors influencing voting behaviour - higher modern studies revision - BBC Bitesize. BBC News. Retrieved November 25, 2022, from https://www.bbc.co.uk/bitesize/guides/zd9bd6f/revision/5

3. Conservative lead widens as unhappiness with federal government and prime minister Trudeau Grows. Abacus Data. (n.d.). Retrieved November 25, 2022, from https://abacusdata.ca/canadian-politics-july-2022-2/

4. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.

5. Factors associated with voting. Statistics Canada: Canada's national statistical agency / Statistique Canada : Organisme statistique national du Canada. (2015, November 27). Retrieved November 25, 2022, from https://www150.statcan.gc.ca/n1/pub/75-001-x/2012001/article/11629-eng.htm

6. Government of Canada, S. C. (2021, December 8). General Social Survey – Family (GSS). Government of Canada, Statistics Canada. Retrieved November 25, 2022, from https://www.statcan.gc.ca/en/survey/household/4501

7. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)

8. Nanos projections show Poilievre's Conservatives winning more seats than Trudeau's Liberals. CTVNews. (2022, October 5). Retrieved November 25, 2022, from https://www.ctvnews.ca/politics/nanos-projections-show-poilievre-s-conservatives-winning-more-seats-than-trudeau-s-liberals-1.6097326

9. Posit. (2022, November 14). Retrieved November 25, 2022, from https://posit.co/

10. Poststratification. Poststratification - an overview | ScienceDirect Topics. (n.d.). Retrieved November 25, 2022, from https://www.sciencedirect.com/topics/mathematics/poststratification

11. Post-stratification and conditional variance estimation. (n.d.). Retrieved November 26, 2022, from https://www.bls.gov/osmr/research-papers/1993/pdf/st930500.pdf

12. Stephenson, et al.(2020) *2019 Canadian Election Study (CES) - Phone Survey.* Harvard Dataverse. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/8RHLG1

13. Taube, M., McKay, J., Canseco, M., & Wise, A. (2019, August 15). Conservatives 'the party of the uneducated?'. The Orca. Retrieved November 25, 2022, from https://theorca.ca/visiting-pod/conservatives-the-party-of-the-uneducated/

14. Technology, A. K. through. (n.d.). Data Centre. CHASS Data Centre. Retrieved November 25, 2022, from https://datacentre.chass.utoronto.ca/

15. The R project for statistical computing. R. (n.d.). Retrieved November 25, 2022, from https://www.r-project.org/

16. Welcome to the 2019 Canadian election study. Canadian Election Study. (n.d.). Retrieved November 25, 2022, from http://www.ces-eec.ca/

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