2025 Canadian federal election forecasting using multiple logistic regression models and poststratification

STA304 - Assignment 2

GROUP 24: Hunter Chen, Leo Watson, TIANYI GUO, Yuka Saito

November 24, 2022

Introduction

Election forecasting intends to give the mass market predictions of political election outcomes by utilising statistical models. It initially evoked the interest of political scientists, but it is increasingly commercialised by polling companies, news organisations and betting markets. Election forecasting could also be an adjunct to developing better political science theories, such as voting behaviour, campaign effects and voters' characteristics (Bélanger and Godbout 2010). Some statistical models for estimating election results have been very successful – a study by Bélanger and Godbout (2010) revealed that their vote function model with macroeconomic factors provided an acceptable estimation to the 2008 Canadian Federal election. In contrast, in recent years, there have been several failures in election forecasting. For example, most surveys and betting markets mispredicted Brexit and the 2016 US Presidential election (Cohn 2016).

Thus election forecasting still requires abundant improvements, but one statistical approach considered to be fairly accurate in estimating the election polls is poststratification. Poststratification reweights the overrepresented or underrepresented poll data(or survey data), assuming we know the true sizes of the strata in the population, resulting in a more representative sample of the true population. Since many representative polling methods encounter high non-response rates; are costly and time consuming(Wang et al., 2014), one could employ poststratification which only required to model for a sample related to the population we are interested in(Reilly et al., 2001).

In this paper, we studied the effect of demographic variables including sex, education level, income level, age on the percentage of popular vote to explore which party will obtain the popular vote of the 2025 Canadian Federal Election. To begin our estimation, we modeled three logistic regression models each predicting the percentage of popular votes for one of the top three most popular parties from the 2019 election: the Liberal Party, positioned to the centre-left, Conservative Party, which sits to the right and New Democratic party(NDP), sitting to the left. After fitting our model on the survey data, which surveyed Canadian citizens in regards to political beliefs, behaviours, attitudes preceding the 2019 Canadian Federal Selection(Stephenson et. al 2020), we used the census data, which could describe the general characteristics of the population in Canada(Statistics Canada 2019), to poststratify. Poststratification is a relevant tool so that we can obtain the percentage of popular vote of the true population of Canada and we justify the logistic regression in the Methods section. We hypothesise that the Conseverative Party will acquire the highest percentage of the popular vote followed by the Liberal Party then the NDP, as the Conservative party had the highest percentage of popular votes for the past two elections in 2019 and 2021(Heard, n.d.).

Data

GSS Data (Census Data): The GSS Data was collected from the 2017 General Social Survey Election Survey (GSS), a nation-wide survey held every five years concerning social trends, quality of life, and living conditions of people fifteen or older living in Canada (Statistics Canada 2019). This survey was a

cross-sectional design study. Responding to the survey was voluntary with 20602 total responses gathered using a "computer-assisted telephone interviewing method" (Statistics Canada 2019).

CES Data (Survey Data): The CES Data was collected from the 2019 Canadian Election Survey (CES), a nation-wide survey of Canadian citizens held every election year concerning Canadians' political beliefs, behaviors, and attitudes (Stephenson et. al 2020). There was an online survey version and a phone survey version. We are only concerned with the phone survey version which had 4021 total responses; respondents were asked over 70 questions with a high, rich level of detail (Stephenson et. al 2020).

GES Cleaning

First, we stored the provided preliminarily cleaned GSS csv in a dataframe. We then rounded all observed age values to the nearest integer and then removed observations with an age less than 18 years old. This was done as the voting age in Canada is 18 so information about Canadians under 18 years old would not be relevant for determining the popular vote.

Next, we trimmed our data by removing all variables except for sex, age, family income, and education level. These variables were kept as they will form our predictors in our final regression models. Then, we removed observations with missingness in any of the remaining (4) variables.

We next defined a new variable (*age_group*; 4 possible levels) dependent on each observation's measured age value split into four possible levels: "ages 18 to 29", "ages 30 to 44", "ages 45 to 59", "ages 60+". This binning was done as the impact of age on voting preference is not strictly linear. Binning is appropriate to group our individuals by these age ranges as they represent individuals in different stages of life. Note that the XBox paper uses the same binning technique and ranges (Wang et. all 2015).

Similarly, we defined a new variable (*education_level*; 5 possible levels) contingent on each observation's education level. We use 5 factor levels so the census and survey education level data would match as originally they had different numbers of categories.

- "College, CEGEP or other non-university certificate or di..." AND "Trade certificate or diploma" were mapped to "College or Trade Diploma"
- "Bachelor's degree (e.g. B.A., B.Sc., LL.B.)" AND "University certificate or diploma below the bachelor's level" were mapped to "Bachelor's Degree"
- "Less than high school diploma or its equivalent" was mapped to "Less than High school Diploma"
- "High school diploma or a high school equivalency certificate" was mapped to "High school Diploma"
- "University certificate, diploma or degree above the bach..." was mapped to "Master degree or higher"

Then, we added a people variable to each observation containing a numeric value of 1. This will be used to count the number of observations in each cell later on.

We then created a copy of the dataframe we have been working with. We will use this copy for the cell creation and poststratification. In this copy, we mutated the sex, age group, family income, and education level variables to ensure each of their levels are stored as factors levels by R (as opposed to numerics, etc.).

Lastly, we group by all of these variables (& their factor levels), creating a new variable (*Number_cell*) containing the number of observations in each of the possible cells (since we have 4 variables with 2, 4, 6, and 5 levels, theoretically we should have 2*4*6*5 = 240 cells following poststratification).

CES Cleaning

First, we read in the "ces2019-phone_clean.csv" CSV file and store it in a dataframe.

We then subset our dataset by only keeping individuals where all of the following are true:

(1) Likely/certain to vote

(2) Male or female gender

(3) Responded to the survey question asking about their education level

(4) will vote for one of the Liberals, Conservatives, NDP, Bloc Québécois, Green Party, People's Party

(5) Responded to the survey question asking about their household income

Satisfying conditions (1) - (5) above corresponds to individuals in the survey satisfying all of:

(1*) Variable q10 == 1 or q10 == 2

(2*) Variable q3 == 1 or q3 == 2

(3*) Variable q61 >= 0

(4*) Variable 1 <= q11 <= 6

(5*) Variable q69 > 0

This is done to remove individuals with missingness in any of the predictors or response for our logistic regression models (**see the Appendix for more information about these variables and their possible factor levels**). We also want to make sure our model is created by those who actually plan to vote and hence robust.

We then trim our dataset by removing all the variables EXCEPT for q10 (propensity to vote or not), q3 (gender), q61 (highest education level), q11 (party the individual supports), q69 (household income), and q2 (year of birth) from all observations (See the Appendix for more detailed descriptions of these raw variables). This is because the other variables are irrelevant to forming our logistic regression models. Then, we remove observations with missingness in any of the above remaining (6) variables.

Next, we define a variable containing the participant's age (*age*), determined by the formula $2019 - q2$. 2019 and q2 are used to determine age because they are the year the survey was done and birth year of participant respectively.

Next, we define a variable describing the participant's family income (*income_family*; 6 possible levels) dependent on the observation's *q69* (survey reported numerical family income) variable value:

- q69 < 25000 is mapped to "Less than $25,000"
- 49999 >= q69 & q69 >= 25000 is mapped to "$25,000 to $49,999"
- 74999 >= q69 & q69 >=50000 is mapped to "$50,000 to $74,999"
- 99999 >= q69 & q69 >= 75000 is mapped to "$75,000 to $99,999"
- 124999 >= q69 & q69 >= 100000 is mapped to "$100,000 to $ 124,999"
- q69 >=125000 is mapped to "$125,000 and more"

These six possible levels were picked in order to mimic how family income was measured in the GSS census data, and hence we could have meaningful comparisons between the census and survey data.

Next, we define a variable describing whether the participant will vote for the Liberal Party or not (*choice1*; 2 possible levels). This variable (and *choice2*/*choice3*) is created as we want to develop a logistic regression model for whether an individual will vote Liberal or not and hence we want a binary classification (i.e. an indicator variable) response.

- If q11 == 1, *choice1* takes value "Liberal". If not, *choice1* takes value "Non-liberal"

We similarly define a variable describing whether the participant will vote for the Conservative Party or not (*choice2*; 2 possible levels).

- If q11 == 2, *choice2* takes value "Conservative". If not, *choice2* takes value "Non-Conservative"

And again we similarly define a variable describing whether the participant will vote for the NDP or not (*choice3*; 2 possible levels).

- If q11 == 3, *choice3* takes value "NDP". If not, *choice3* takes value "Non-NDP".

Mimicking the GES dataset's way of measuring age, we next define a new variable (*age_group*; 4 possible levels) describing each participant's age split into 4 possible bins:

- "ages 18 to 29", "ages 30 to 44", "ages 45 to 59", and "ages 60+"

We rename the following variables for intuition and easier understanding:

- *q10* is mapped to "vote"
- *q3* is mapped to "sex"
- *q3* took only values 0 and 1. Map 1 to "Male", 0 to "Female"
- *q11* is mapped to "vote_party"
- *q69* is mapped to "income"
- *q2* is mapped to "year_born"
- *q4* is mapped to "province"

Next, mimicking the GES dataset's way of measuring education level, we define a variable describing the participant's education level (*education_level*; 5 possible levels) dependent on the observation's *q61* (survey reported education level) variable value:

- *q61* < 3 is mapped to "Less than High school Diploma"
- $4 <=$ *q61* $<= 5$ is mapped to "High school Diploma"
- $6 <=$ *q61* $<= 7$ is mapped to "College or Trade Diploma"
- $8 <=$ *q61* $<= 9$ is mapped to "Bachelor's degree"
- $10 <=$ *q61* is mapped to "Master degree or higher"

Then, we define an indicator variable describing whether the participant will vote liberal or not (*vote_Liberal*; 2 possible levels):

- If *choice1* == "Liberal", *vote_Liberal* takes value 1. Otherwise, *vote_Liberal* takes value 0.
- We similarly define indicator variables describing whether the participant will vote Conservative (*vote_Conservative*; 2 possible levels) and whether the participant will vote NDP (*vote_NDP*, 2 possible levels).

Lastly, using this dataframe, create three separate dataframes:

- One dataset containing whether the participant will vote Liberal or not, their sex, age group, family income, and education level.
- Another dataset containing whether the participant will vote Conservative or not, their sex, age group, family income, and education level.
- One more dataset containing whether the participant will vote NDP or not, their sex, age group, family income, and education level.

Note that this final step is done so that we have the three necessary datasets for the separate logistic regression models we intend to fit for the three different binary responses: voting Liberal or not, voting Conservative or not, voting NDP or not.

Age_group: The GES dataset contains the measured age of the participant and the CES contains the participant's birth year so we can extract the participant age from both datasets. We convert this continuous variable into the categorical variable *age_group* with four bins: 18 to 29, 30 to 44, 45 to 60, and 60+. We determined age would be highly relevant in predicting voting preference as research has shown that younger people tend to be more liberal while older people tend to be more conservative (Holland 2013).

Income_family: Both the GES and CES datasets contains the participant's family income in Canadian Dollars. We convert this continuous variable into the categorical variable *Income_family* with 6 bins: less than $25,000; between $25,000 and $49,999; between $50,000 and $74,999; between $75,000 and $99,999; between $100,000 and $124,999; $125,000+. We determined family income would be highly relevant in predicting voting preference/turnout as research has shown that as income increases, a voter's likelihood of voting more conservative (right-leaning) increases (Arunachalam and Watson 2018).

Sex: The GES data directly contained information about participants' sex, while the CES data gave information about the participants' gender. Since gender has more than two categories (Kennedy et. al 2020), we decided to only consider surveyed individuals who self-identify as "Male" or "Female" gender, labelling these as the individual's *Sex*. We decided this would be appropriate as only a single participant (out of 4021) was non-binary and hence we would not lose much information nor would we have enough data to make inferences about the non-binary population. We determined sex would be highly relevant in predicting voting preference as research has shown that women are more likely to be liberal (left-leaning) compared to men in their voting preferences (Abendschön and Steinmetz 2014).

Education_level: The CES survey allowed participants to select 11 possible highest education levels while the GSS data had 7 possible levels of highest education. We decided to combine these into a single categorical variable *Education_level* with five bins: Less than High school Diploma, High school Diploma, College or Trade Diploma, Bachelor's degree, Master degree or higher. We determined education level would be highly relevant in predicting voting preference as research has shown that as one becomes more educated, they are more likely to be conservative (right-leaning) in their voting preferences (Marshall 2015).

Table 1: The numbers of individuals in each gender in the survey data

sex	Number	Proportion
Female	793	41%
Male	1138	59%

Table 2: The numbers of individuals in each gender in the census data

sex	Number	Proportion
Female	10849	54%
Male	9066	46%

Table 3: The numbers of individuals in each family income category in the survey data

income_family	Number	Proportion
$100,000 to $ 124,999	248	13%
$125,000 and more	613	32%
$25,000 to $49,999	285	15%
$50,000 to $74,999	371	19%
$75,000 to $99,999	269	14%
Less than $25,000	145	8%

Table 4: The numbers of individuals in each family income category in the census data

income_family	Number	Proportion
$100,000 to $ 124,999	2091	10%
$125,000 and more	4505	23%
$25,000 to $49,999	4223	21%
$50,000 to $74,999	3581	18%
$75,000 to $99,999	2841	14%
Less than $25,000	2674	13%

Remark: Note that all the predictors are categorical hence numerical summaries involving measures of center & spread such as mean and standard deviation wouldn't be meaningful. Hence, the summary tables indicate each factor level's count and proportion.

Table 1 and Table 2 indicate that while the proportion of females in the survey data was less than the proportion of males by about 20%, in the census data there was about 10% more females than males. This indicates that the survey sample may not correctly reflect the actual population distribution. We observe a similar phenomenom in the pairs Table 3/4, Figure 1/2, and Figure 3/4. Hence, poststratification will be helpful in improving the accuracy of our estimates and reducing variance.

Table 3 indicates that the proportion of surveyed individuals in the $50,000 to $74,999 category, $75,000 to $99,999 category, and $100,000 to $124,999 category are roughly the same. Almost a third of surveyed individuals had a family income of $125,000 or more while less than 10% had a family income less than $25,000.

Table 4 is rather different from table 3 with the income group with the smallest proportion of total participants now being the $100,000 to$124,999 category. The $125,000 or more category still comprises the largest proportion of individuals, but with only 23% of the total as opposed to 32% of the surveyed individuals. The other categories are roughly similar in terms of proportion to the survey data.

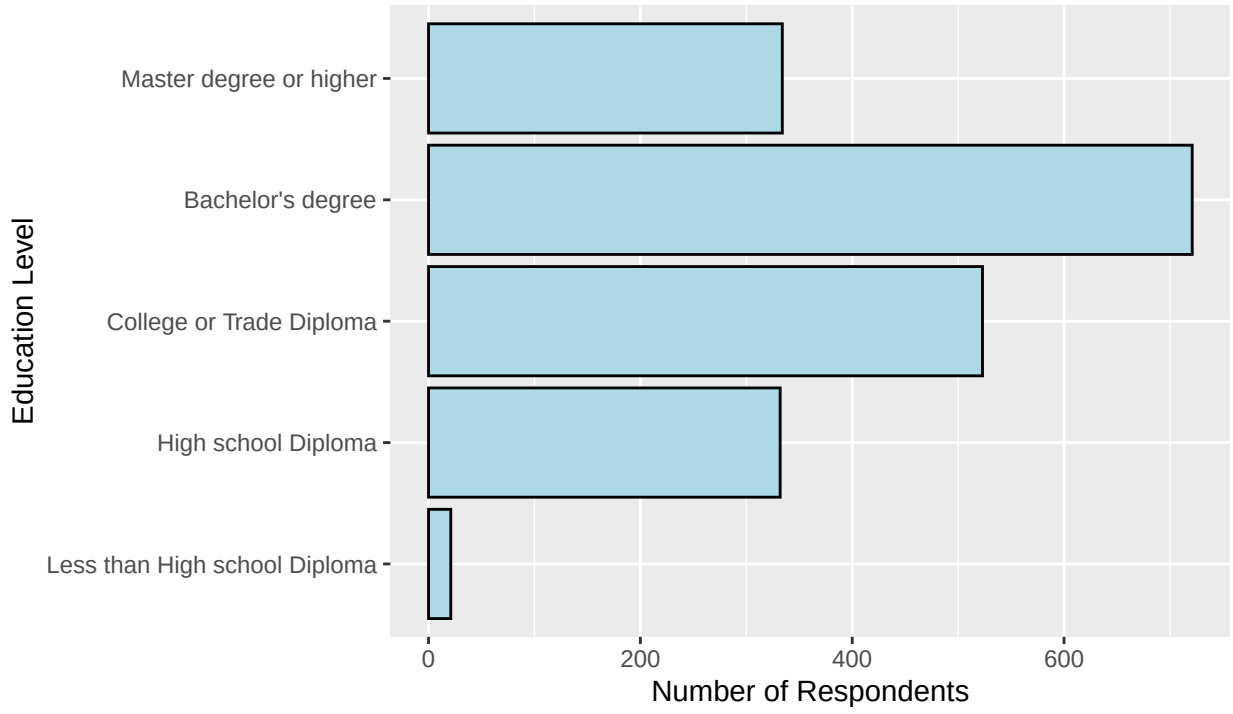


Figure 1: Survey Respondent Education Level Distribution

Figure 1 is a barplot of the highest attained education level for the CES survey individuals. Figure 1 demonstrates that almost all respondents have at least a high school diploma, with around 50% having a Bachelor's Degree or higher. The mode of this distribution is the Bachelor's degree by a significant margin, i.e. most people in the survey had a Bachelor's degree as their highest level of education.

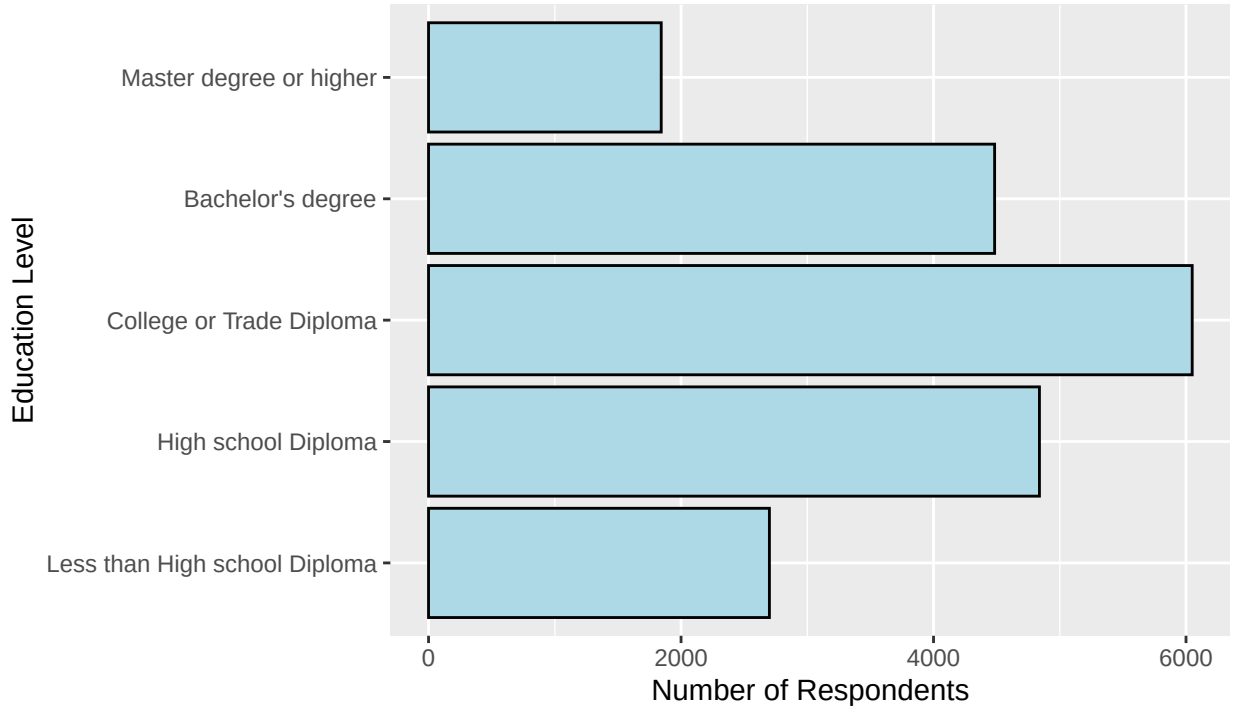


Figure 2: Census Respondent Education Level Distribution

Figure 2 is a barplot of the highest attained education level for the GSS survey individuals. Comparing Figure 2 to Figure 1, we observe what looks to be a a fairly symmetric "normal" distribution centred around the College/Trade diploma (note it makes sense to talk about a normal distribution in this context because highest education level is an ordinal variable). This contrast with Figure 1 which has greater mass towards the higher education, indicating that the survey may have sampled from a non-representative subset of the population. In Figure 2, the college/trade diploma is the most common highest level of education and the individuals without a high school diploma comprise a larger proportion of the total sample relative to the survey data.

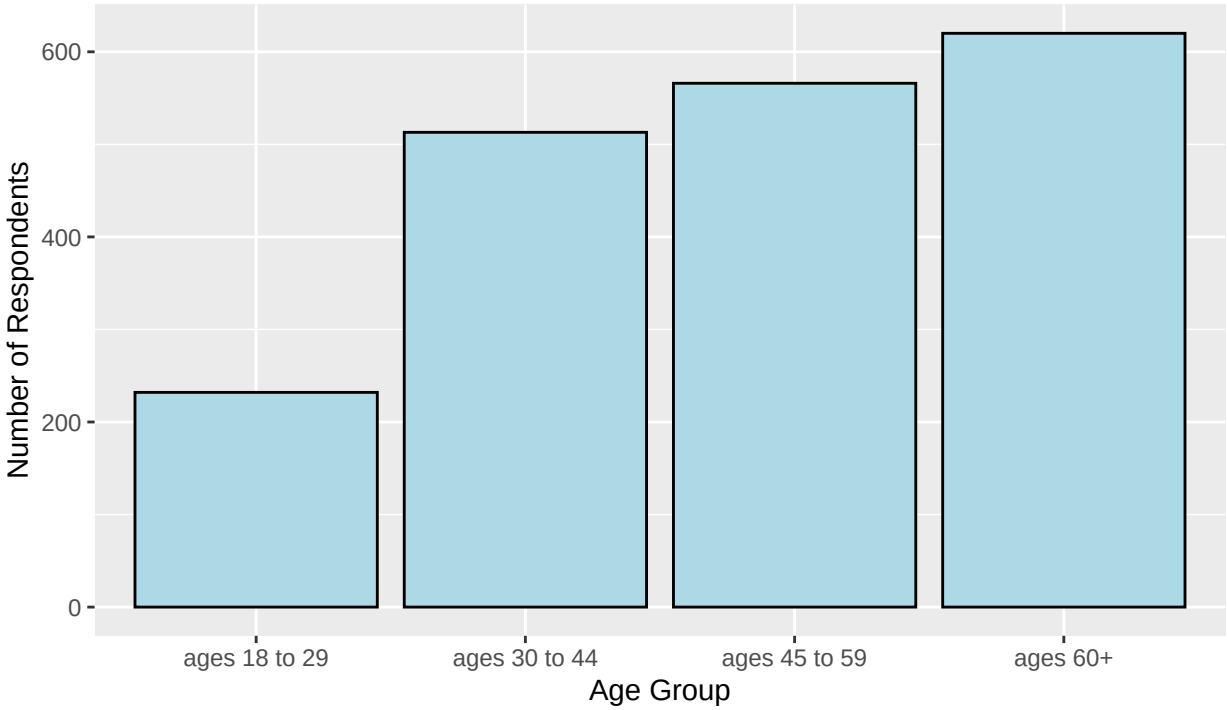


Figure 3: Survey Respondent Age Group Distribution

Figure 3 is a barplot of the age group the CES survey individuals were in at the time of the survey. We see there's more individuals in the higher age group categories compared to the lower age group categories. There's significantly less individuals 15-29 years old sampled in the survey while the number of 30-44, 45-59, and 60+ individuals is roughly the same.

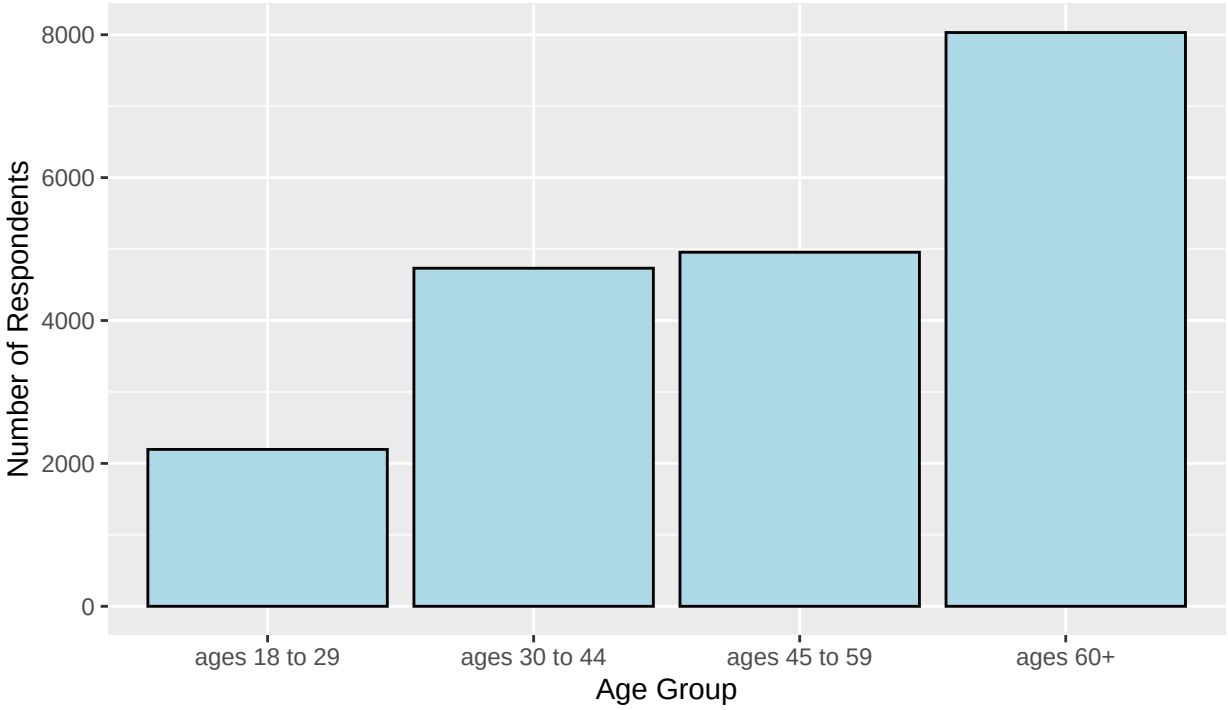


Figure 4: Census Respondent Age Group Distribution

Figure 4 is a barplot of the age group the GSS cenuss individuals were in at the time of the census. We note that similar to Figure 3, the upper age groups have greater mass. However, in Figure 4 we notice individuals 60+ form a far larger proportion of the census individuals compared to survey individuals. This again may indicate that the survey is not entirely representative of the population.

Note: `R version 4.0.2` was used to complete this section. Tables were generated using the kable() function from the knitr package. Figures were created using the ggplot() function from the Tidyverse package.

Methods

We will apply logistic regressions and poststratification to predict the percentage of the popular vote for each party. Specifically, we will mainly focus on the probability of winning for Liberal, Conservative, and NDP, as they are the three major parties in Canada. We will use data from the 2019 phone survey to fit our predictive model. Logistic regression can easily explain the relationship between a binary variable and other independent variables. Since we are interested in whether people will vote for a particular party in the next Canadian federal election, a logistic regression model is appropriate. Based on the model, we can predict voters' probability of voting for a specific party. The advantages of using multiple logistic regression include simplicity and efficiency compared to other more complicated models.

The main procedure consists of several steps. In the first step, we will use the survey data to select several significant variables and use them to construct a multiple logistic regression model. After fitting the model based on the survey data, we partition the census dataset into demographic cells. Then we apply the fitted model to the post-stratified dataset to estimate the winning probability for the specific party in each cell. Lastly, we aggregate the cell-level estimates to the population level by weighting each cell by proportion. The paper will show a more rigorous explanation with mathematical models later.

Model Specifics

The research uses logistic regression to predict the probability of individuals voting for a specific party, given their demographic characteristics. We will use three logistic regression models to estimate the probabilities of winning the election for Liberal, Conservative, and NDP.

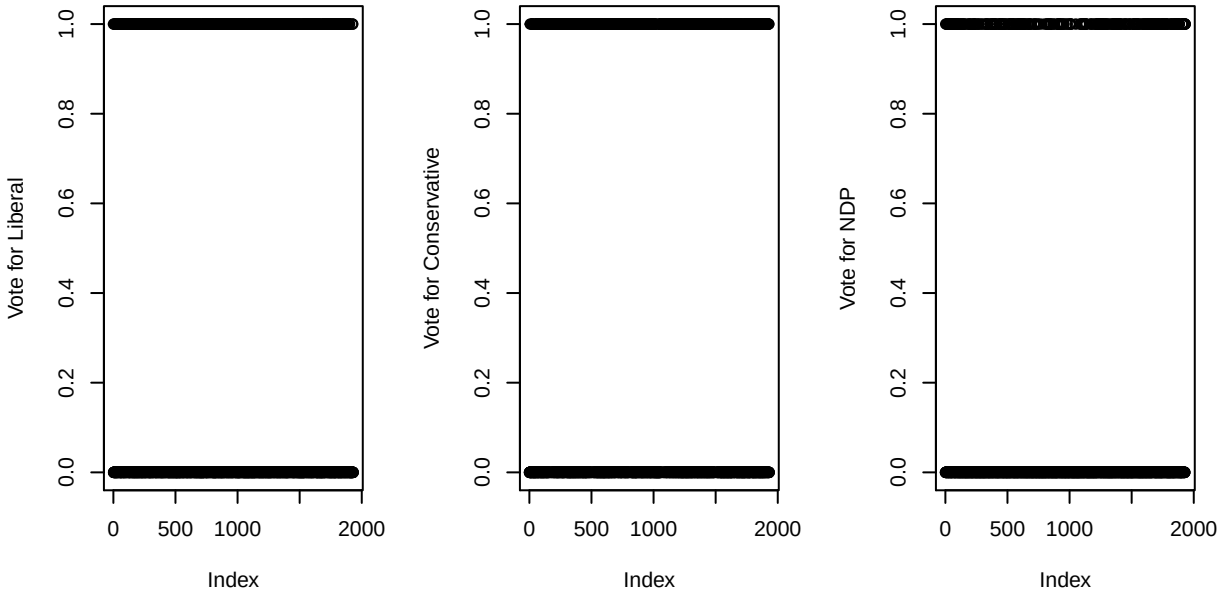
Logistic regression

Before applying the logistic regression, it is necessary to check if the data satisfy the following assumptions

- outcome variable is binary
- linearity in the logit for continuous variables
- no observation of multicollinearity
- lack of severe influential points

1. Binary outcome

Figure.5 Values of outcome variables in three logistic models



Based on our definition of outcome variables *vote_Liberal*, *vote_Conservative*, and *vote_NDP*, we know that they should all be binary. As a sanity check, these graphs illustrate that the assumption of binary outcome holds.

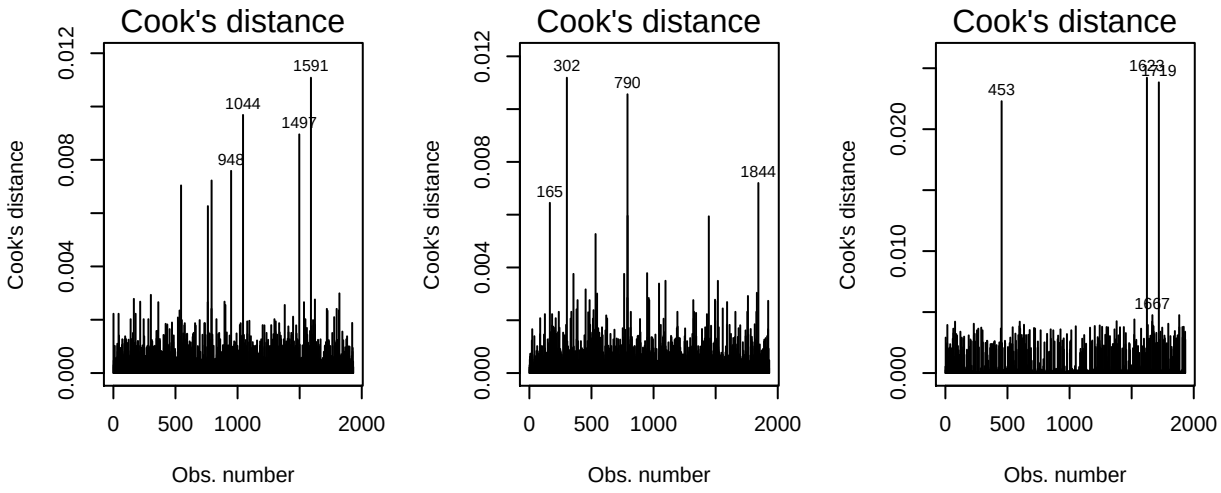
2. Checking linearity of logit

The models we construct do not have any continuous variables. The models only contain categorical variables. Checking linearity for categorical variables is beyond the scope of the research. We would assume it to be true for our analysis.

3. Checking influential values

Influential values in the survey dataset could affect the quality of the logistic regression model. We could find the most extreme values in the data by visualizing the Cook's distance values.

Figure.6 cooks distance based on three logistic models



As we can see, the three graphs indicates 4 most influential observations for each model predicting voting probability for liberal, Conservative and NDP party. Since we do not have a contextual reason to remove the problematic points. We will simply acknowledge these influential points' existence.

4. Multicollinearity

Multicollinearity refers to a situation where the data involve highly correlated predictor variables. We will compute the variance inflation factors to assess this assumption.

Table 5: VIF table

Variables	Liberal model	Conservative model	NDP model
sex	1.035589	1.022295	1.041563
age	1.141874	1.146242	1.142083
income	1.228355	1.241818	1.187990
education	1.138382	1.152742	1.131983

The above table displays the Variance Inflation Factor(VIF) for each variable under three different models. As the results show that VIF values are all below 5, which implies no severe problem with collinearity. The assumption of no multicollinearity satisfies for our models

Model Selection

We will exploit the Akaike information criterion (AIC) to identify the best-fit model. We want to estimate whether vote or not for the Liberal party from the following pool of predictors: sex, age, income and education level. These predictors are chosen since they are likely correlated with one's vote intention. In addition, they are appeared both in the census and survey data, which allows us to conduct poststratification later.

Table 6: AIC for different models

model	AIC_values
sex	2583.821
age	2579.728
income family	2589.094
education level	2555.169

model	AIC_values
sex + age	2577.592
sex + income family	2586.55
age + education level	2550.72
sex + age + income family	2577.818
sex + age + education level	2550.003
sex + income family + education level	2559.622
age + income family + education level	2555.696
sex + age + income family + education	2554.801

Note: Akaike information criterion is calculated by R function AIC() from the stats package.

After trying a few different models, the model that attains the lowest AIC is the full information model with predictors: sex, age, income family and education level. As a result, we will apply this model to predict the percentage of popular vote for Liberal, Conservative and NDP parties.

First, We will exploit a logistic regression model to predict the proportion of votes for the Liberal party based on the rich demographic information that the census data provides.

Below is the mathematical representation of the logit regression

$$log(\frac{p_{Liberal}}{1-p_{Liberal}}) = \beta_0 + \beta_1 x_{sex} + \beta_2 x_{age\ group2} + \beta_3 x_{age\ group3} + \beta_4 x_{age\ group4} + \beta_5 x_{income\ level2} + \beta_6 x_{income\ level3} + \beta_7 x_{income\ level4} + \beta_8 x_{income\ level5} + \beta_9 x_{income\ level6} + \beta_{10} x_{education\ level2} + \beta_{11} x_{education\ level3} + \beta_{12} x_{education\ level4} + \beta_{13} x_{education\ level5} + \epsilon$$

Where $p_{Liberal}$ represents the probability of winning the election for Liberal. β_1 represents change in log odds for participants reporting their sex as male. β_0 represents the fixed baseline intercept $\{\beta_1, \beta_2, \beta_3, \beta_4\}$ represents the change in log odds for reporters in different age groups respectively. $\{\beta_5, \beta_6, \beta_7, \beta_8, \beta_9\}$ represents the change in log odds for reporters with different income levels respectively.

$\{\beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}\}$ represents the change in log odds for participants in different education levels.
The same logic applies for the Conservative and NDP:

$$log(\frac{p_{Conservative}}{1-p_{Conservative}}) = \beta_0 + \beta_1 x_{sex} + \beta_2 x_{age\ group2} + \beta_3 x_{age\ group3} + \beta_4 x_{age\ group4} + \beta_5 x_{income\ level2} + \beta_6 x_{income\ level3} + \beta_7 x_{income\ level4} + \beta_8 x_{income\ level5} + \beta_9 x_{income\ level6} + \beta_{10} x_{education\ level2} + \beta_{11} x_{education\ level3} + \beta_{12} x_{education\ level4} + \beta_{13} x_{education\ level5} + \epsilon$$

Where $p_{Conservative}$ represents the probability that people vote for Conservative

$$log(\frac{p_{NDP}}{1-p_{NDP}}) = \beta_0 + \beta_1 x_{sex} + \beta_2 x_{age\ group2} + \beta_3 x_{age\ group3} + \beta_4 x_{age\ group4} + \beta_5 x_{income\ level2} + \beta_6 x_{income\ level3} + \beta_7 x_{income\ level4} + \beta_8 x_{income\ level5} + \beta_9 x_{income\ level6} + \beta_{10} x_{education\ level2} + \beta_{11} x_{education\ level3} + \beta_{12} x_{education\ level4} + \beta_{13} x_{education\ level5} + \epsilon$$

Where p_{NDP} represents the probability that NDP wins the election

Note: `R version 4.0.2` was used to complete this section. Tables were generated using the kable() function from the knitr package. Figures were created using the ggplot() function from the Tidyverse package.

Post-Stratification

The idea of postratification is to make our sample data more representative of our target population. We should expect our estimation to be have less variance by adjusting the sampling weights. In our report, we partitioned the census data into different demographic cells to predict voter intent in each cell with the use of logistic regression, and then calculate the final estimation based on the strata-level estimates based on its proportion in the total population. More specifically, we weigh the phone survey results so that the over-represented responses have less effect and the under-represented responses have more strength in the result. For instance, if the sample has 99 percent of people supporting the Liberal party and one percent supporting other parties, the result would be biased as the true population will likely differ from that . The purpose of poststratification is to weigh the dataset so that voters within each cell can more accurately represent the population of interest with the specific demographic groups.

We choose to use a sub-data set of census data for the prediction. The cleaned census data includes sex, age, income and education level, as they are the basic demographic characteristics that we can extrapolate into how the entire population will vote. Initially, we have 20602 unique respondents, which is a large sample but might still be unrepresentative of the entire population. We need to further break down each demographic variable into different categories to partition the voters. Therefore, we choose to create cells based on combinations of:
(1)sex (2 categories)
(2)age group(4 categories)
(3)income level(6 categories)
(4)education level(5 categories)
Where in total we generate $2 \times 4 \times 6 \times 5 = 240$ cells. Hence we partition the census data into 240 cells for the poststratification analysis.

We choose to include sex while constructing the cells because there is a potential difference in political preferences between males and females, which might drive the results if we exclude them from the cells (Hatemi, McDemott, Bailey & Martin, 2012). The voter turnout by age in 2018 US election was (US News, 2022):

Table 7: 2018 US Election turnout rate

Age	turnout rate
age 18 to 24	30%
age 25 to 34	37%
age 35 to 44	44%
age 45 to 64	55%
age above 65	64%

The voter turnout suggests that senior people are more willing to vote compared to younger generations. Including the age variable when creating cells can separate the choices that seniors make as a large portion of voters are aged over 65, which could lead to an inaccurate result if we do not control for different age groups. Besides, the turnout rate by family income in the United of America found that people with higher incomes are more likely to vote than low-income families (Voting and Income, 2019). As a result, partitioning respondents into different income levels could better fit the model with less bias. The article "How, and For Whom, Does Higher Education Increase Voting?" reveals that individuals with a university or college education have a higher voting rate than those with less education (Ahearn, Brand & Zhou, 2022). Therefore, we need to consider the relationship between a voter's education level and the probability of voting since they might drive the result. Applying these four variables to generate cells helps us understand and extrapolate the effect within each cell. Then, we can aggregate each cell's weighted proportion of the electorate with the corresponding estimated outcome to approximate the result.

In order to estimate the winning likelihood of Liberal/Conservative/NDP at a population-level, we fit three logistic regressions for estimating candidate support in each cell.

Denoting p as the probability of winning for the relevant parties, the poststratification estimates of an overall sample proportion are defined by the following:

$$\hat{p}^{PS}_{Liberal} = \frac{\sum N_j \hat{p}_{j\,Liberal}}{\sum N_j}$$

$$\hat{p}^{PS}_{Conservative} = \frac{\sum N_j \hat{p}_{j\,Conservative}}{\sum N_j}$$

$$\hat{p}^{PS}_{NDP} = \frac{\sum N_j \hat{p}_{j\,NDP}}{\sum N_j}$$

Where $\hat{p}_j$ represents the estimation of $\hat{p}$ in strata j, and N_j refers to the size of the j-th strata in the census population.

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

Results

In this section, we present the results of our statistical analyses. Regarding the three performed logistic regressions, setting the significance level to 0.05 and employing z-statistic(since the variance of the population is known), for all 3 models, some coefficients were not statistically significant due to the p-value being larger than 0.05. Also, all 3 models differ in which coefficients are statistically significant or not, which affects the accuracy of our predictions(see Appendix 15, 16, and 17). However, our main goal is to predict the overall popular vote of the next 2025 Canadian federal election through re-weighting the post-stratified estimates based on the census data, so we still proceeded with our analyses.

Using our purposed models after post-stratification, we predict that in the 2025 Canadian Federal Election, the conservative party will receive the most popular votes, from 35.4% of the population followed by liberal party, where votes will come from 32.6% of the population and lastly, only 16% of the population will vote for the NDP, as shown in Table 5. This perfectly aligns with our initial hypothesis.

Table 8: The Predicted Percentage of Popular Votes in the 2025 Canadian Federal Election

Political Party	Percentage of Popular Votes
Liberal Party	0.3257243
Conservative Party	0.3542133
NDP	0.1599195

Conclusions

In this paper, we constructed forecasts of the 2025 Canadian Federal Election, and we hypothesized that the Conservative party would receive the greatest amount of popular votes followed by the Liberal party and then the NDP. Based on the 2019 Canadian Election Survey (CES) data, we fitted three multiple logistic models to obtain coefficients' estimates.

Then, we partitioned the census data into 239 demographic cells based on various categories of our demographic variables. For each cell, we applied our fitted models to predict the probability of voting for each of the three major parties. Lastly, reweigh each cell's estimate based on its relative proportion in the total population to make them more representative of the population. This allowed us to conclude that the Conservative party would likely win the Canadian Federal Election popular vote, followed by the Liberal party and then the NDP party.

More specifically, following our analysis and methods implementation we determined that the Conservative party would receive 35.4% of the popular vote, the Liberal party would receive 32.6% of the popular vote,

and the NDP party would receive 16% of the popular vote. Since our figures for these three parties comprise 84% of the popular vote, we are almost certain one of these parties will win the popular vote. This also implies we expect the Conservative party to win the popular vote in the 2025 Canadian Federal Election. However, our predictions for the popular vote received by the Liberal & Conservative party are relatively close so we cannot rule out the Liberal party claiming the popular vote.

While we have achieved our main goal of predicting the popular vote, limitations of our model and data exist. Multiple logistic regression was fitted at each cell level, resulting fairly simplistic models(i.e. Not multilevel modelling), so our models may not fully capture the true relationship between popular votes and demographics and the predictions may be biased. A limitation to our model selection is we only validated our models using AIC. If we considered other model selection criteria such as BIC, Corrected AIC, adjusted R^2, we may have found a model better suited to our task.

In terms of our data, the CES data (survey data) is from 2019 while the GSS data (census data) is from 2017. Hence, using the GSS data for poststratification is not ideal as the population data will have changed between 2017 and 2019. This could increase the bias of our estimates. In addition, the relationship between demographics and vote intention could change from 2019 to 2025 and the demographic composition in 2025 will almost surely differ from that in 2017. Both factors suggest potential bias in our estimates.Moreover, certain variables had a lower response rate than other variables in the census/survey data. For example, the census survey organizer noted that a high percentage of respondents did not answer the question about their household income. This could be due to many reasons such as the selection bias, where high income individuals do not want to publicize their income. Hence, we may be under-representing certain demographic groups in our model such as high income individuals. Thus, our model estimates could be biased and not representative of the population, especially with respect to income level.

For future analyses, we recommend the analysts to pick the most appropriate model using the one that has the smallest (or close to smallest) BIC, corrected AIC, AIC, and largest adjusted R^2. We would also recommend the future analyst to collect more data at both the census level and survey level (preferably from the same year). There were only ~20,000 respondents for the GSS census data and ~4,000 respondents for the CES survey data. More data could provide better estimates of model coefficients with smaller confidence intervals, less variance, and less bias. If the future analysts use the same methodology, they may also generate popular vote estimates for the next three largest parties: Bloc Québécois, Green Party, People's Party. Then, instead of being able to predict the exact voting preferences of 83% of the vote as our model currently has, we will be able to explain a larger share of the total vote.

Bibliography

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

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

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

Intro: Bélanger, Éric, and Jean-François Godbout. "Forecasting Canadian Federal Elections." PS: Political Science and Politics 43, no. 4 (2010): 691–99. http://www.jstor.org/stable/40927037.

Cohn, Nate. "Why the Surprise over 'Brexit'? Don't Blame the Polls." The New York Times. June 24, 2016. https://www.nytimes.com/2016/06/25/upshot/why-the-surprise-over-brexit-dont-blame-the-polls.html.

data:

4. Abendschön, Simone, and Stephanie Steinmetz. "The gender gap in voting revisited: Women's party preferences in a European context." Social Politics 21, no. 2 (2014): 315-344.

5. Arunachalam, Raj, and Sara Watson. "Height, Income and Voting." British Journal of Political Science 48, no. 4 (2018): 1027–51. doi:10.1017/S0007123416000211.

6. "General Social Survey - Family (GSS)." 2019. February 7, 2019. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4501.

7. Holland, Jenny L. Age gap? The influence of age on voting behavior and political preferences in the American electorate. Washington State University, 2013.

8. Kennedy, Lauren, Katharine Khanna, Daniel Simpson, and Andrew Gelman. "Using sex and gender in survey adjustment." arXiv preprint arXiv:2009.14401 (2020).

9. Marshall, John. "Education and voting Conservative: Evidence from a major schooling reform in Great Britain." The Journal of Politics 78, no. 2 (2016): 382-395.

10. 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

Hatemi, Peter K., Rose McDermott, J.Michael Bailey, and Nicholas G. Martin. 2012. "The Different Effects of Gender and Sex on Vote Choice." Political Research Quarterly 65 (1): 76–92. https://doi.org/10.1177/1065912910391475.

"Why Older Citizens Are More Likely to Vote - US News & World Report." Accessed November 29, 2022. https://money.usnews.com/money/retirement/aging/articles/why-older-citizens-are-more-likely-to-vote.

Akee, Randall, Randall Akee, Lynne Pepall and Dan Richards, Lynne Pepall, and Dan Richards. "Voting and Income." Econofact, February 7, 2019. https://econofact.org/voting-and-income.

Ahearn, C.E., Brand, J.E. & Zhou, X. How, and For Whom, Does Higher Education Increase Voting?. Res High Educ (2022). https://doi.org/10.1007/s11162-022-09717-4

Appendix

Plots of the remaining predictors (family income and sex)

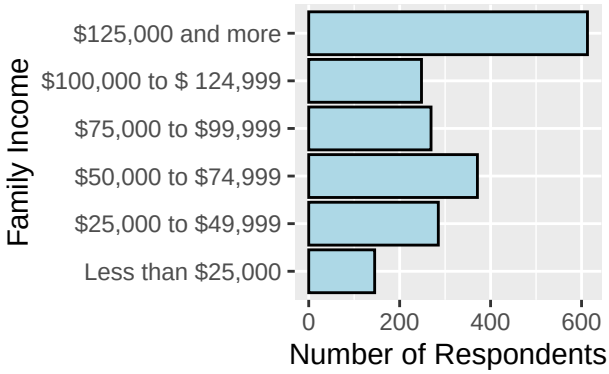


Figure 5: Survey Respondent Family Income Barplot

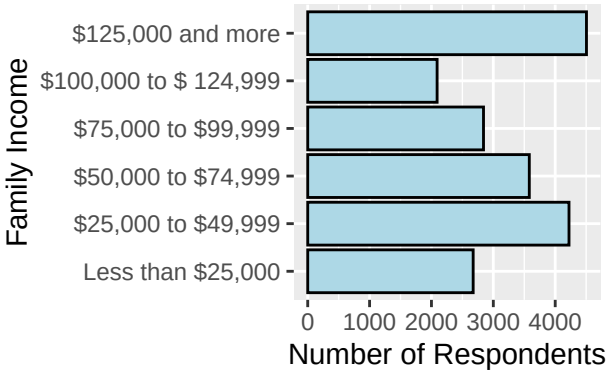


Figure 6: Census Respondent Family Income Barplot

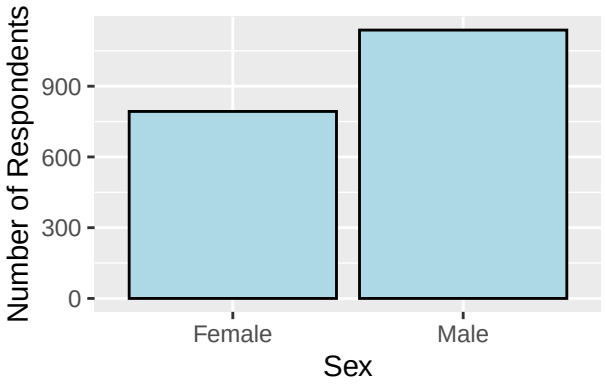


Figure 7: Survey Respondent Sex Barplot

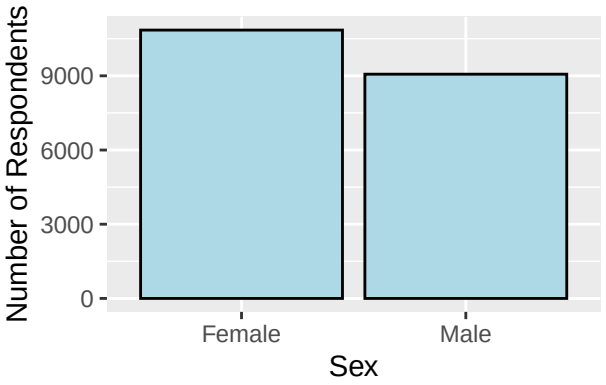


Figure 8: Census Respondent Sex Barplot

Numerical Summaries of the remaining predictors (education level & age group)

Table 9: The numbers of individuals in each age group in the survey data

age_group	Number	Proportion
ages 18 to 29	232	12%
ages 30 to 44	513	27%
ages 45 to 59	566	29%
ages 60+	620	32%

Table 10: The numbers of individuals in each age group in the census data

age_group	Number	Proportion
ages 18 to 29	2196	11%
ages 30 to 44	4732	24%
ages 45 to 59	4956	25%
ages 60+	8031	40%

Table 11: The numbers of individuals in each education level category in the survey data

education_level	Number	Proportion
Bachelor's degree	721	37%
College or Trade Diploma	523	27%
High school Diploma	332	17%
Less than High school Diploma	21	1%
Master degree or higher	334	17%

Table 12: The numbers of individuals in each education level category in the census data

education_level	Number	Proportion
Bachelor's degree	4484	23%
College or Trade Diploma	6049	30%
High school Diploma	4839	24%
Less than High school Diploma	2700	14%
Master degree or higher	1843	9%

AIC for different models

Table 13: AIC for different models

model	AIC_values
sex	2583.821
age	2579.728
income family	2589.094
education level	2555.169
sex + age	2577.592
sex + income family	2586.55
age + education level	2550.72
sex + age + income family	2577.818
sex + age + education level	2550.003
sex + income family + education level	2559.622
age + income family + education level	2555.696
sex + age + income family + education	2554.801

Table 14: The Predicted Percentage of Popular Votes in the 2025 Canadian Federal Election

Political Party	Percentage of Popular Votes
Liberal Party	0.3257243
Conservative Party	0.3542133
NDP	0.1599195

Summaries of the logistic regression model - Liberal Party

Table 15: Coefficient-Level Estimates for a Logistic Regression Model estimating the percentage of voters for the Liberal Party

Predictor	Coefficients	z-statistic	p-value
Intercept	-0.6053666	-2.8344380	0.0045906
Male	-0.1722909	-1.7115917	0.0869719
ages 30 to 44	0.0224044	0.1261134	0.8996422
ages 45 to 59	-0.0195269	-0.1109048	0.9116919
ages 60+	0.3490646	2.0689892	0.0385471
$125,000 and more	0.0896680	0.5435727	0.5867355
$25,000 to $49,999	0.1284306	0.6727312	0.5011183
$50,000 to $74,999	0.0062036	0.0344609	0.9725096
$75,000 to $99,999	0.2738504	1.4551734	0.1456213
Less than $25,000	-0.1876187	-0.7897816	0.4296553
College or Trade Diploma	-0.5199988	-4.0673316	0.0000476
High school Diploma	-0.4348771	-2.9311664	0.0033769
Less than High school Diploma	-0.1638857	-0.3447738	0.7302644
Master degree or higher	0.1985804	1.4432129	0.1489605

Summaries of the logistic regression model - Conservative Party

Table 16: Coefficient-Level Estimates for a Logistic Regression Model estimating the percentage of voters for the Conservative Party

Predictor	Coefficients	z-statistic	p-value
Intercept	-1.3936991	-6.3661773	0.0000000
Male	0.6151057	5.9200376	0.0000000
ages 30 to 44	0.1356664	0.7490017	0.4538562
ages 45 to 59	0.3482287	1.9766410	0.0480822
ages 60+	0.4632361	2.6854875	0.0072424
$125,000 and more	0.2666642	1.6465990	0.0996405
$25,000 to $49,999	-0.4018969	-2.0939821	0.0362616
$50,000 to $74,999	-0.1865360	-1.0518040	0.2928895
$75,000 to $99,999	-0.2452440	-1.2707265	0.2038260
Less than $25,000	-0.5070617	-2.1400375	0.0323517
College or Trade Diploma	0.5345719	4.3190384	0.0000157
High school Diploma	0.6875696	4.8145603	0.0000015
Less than High school Diploma	0.3366165	0.7160944	0.4739331
Master degree or higher	-0.5059366	-3.2166855	0.0012968

Summaries of the logistic regression model - NDP

Table 17: Coefficient-Level Estimates for a Logistic Regression Model estimating the percentage of voters for the NDP

Predictor	Coefficients	z-statistic	p-value
Intercept	-0.6281450	-2.4906492	0.0127510
Male	-0.5401759	-4.0700505	0.0000470
ages 30 to 44	-0.4627763	-2.3900088	0.0168480
ages 45 to 59	-0.8494876	-4.2381700	0.0000225

Predictor	Coefficients	z-statistic	p-value
ages 60+	-1.3900432	-6.7501215	0.0000000
$125,000 and more	-0.4845870	-2.1713333	0.0299060
$25,000 to $49,999	0.2346921	0.9758679	0.3291299
$50,000 to $74,999	0.0416389	0.1794671	0.8575710
$75,000 to $99,999	-0.0385812	-0.1564627	0.8756683
Less than $25,000	0.4295371	1.5539589	0.1201942
College or Trade Diploma	0.0133716	0.0823737	0.9343496
High school Diploma	-0.2118584	-1.0388037	0.2988960
Less than High school Diploma	0.0657553	0.1017598	0.9189473
Master degree or higher	0.2105714	1.1041355	0.2695344