Predicing overall popular vote of 2025 Canadian federal election by voters' sense of religious belonging and age

STA304 - Fall 2023 -Assignment 2

Wai Yu Amanda Ng (Completed individually)

16 November 2023

Part 1: Introduction

1.1 Background and terminology

The next Canadian federal election is tentatively happening in 2025. In this election, federal parties will compete for seats in the House of Commons, the lower house of the Parliament of Canada where members introduce, debate and pass bills. The winning party in the federal election is defined as the federal party which wins the most seats in the House of Commons. This is an important election since the leader of the winning party will be appointed as the prime minister, which is the most powerful position in Canadian politics.

The main federal parties include the Liberal Party of Canada, Conservative Party of Canada, Bloc Québécois, New Democratic Party and Green Party of Canada. Among all five of them, the Liberal Party of Canada which was founded in 1867, is the longest-serving and oldest active federal political party in the country [1]. The Liberal Party of Canada generally sits at the centre to centre-left of the Canadian political spectrum, meaning that it seeks to achieve social equality and egalitarianism through the expanding role of government [2]. It practices liberalism which aims to include individual freedom for present and future generations, responsibility, human dignity, a just society, political freedom, religious freedom, national unity, equality of opportunity, cultural diversity, bilingualism, and multilateralism [1]. Conservative Party of Canada is a right-wing party, which is Liberal Party of Canada's main rival,was founded in 2003 [2]. As opposed to achieving social equality, it views certain social orders and hierarchies as inevitable or desirable, individual rights and civil liberties are paramount and the role and power of the government is minimized [2]. New Democratic Party is a left-wing party founded in 1961 which practices social democracy that emphasizes on advocating economic and social interventions to promote social justice within the framework of a liberal-democratic polity and a capitalist-oriented mixed economy [3]. These three parties are the strongest candidates for the upcoming election since they are the parties with top three popular vote percentages in past elections [4].

As mentioned above, upholding religious freedom is one of the main objectives of the Liberal Party of Canada's ideology. On one hand, it may seem beneficial to attract voters from all religious backgrounds. On the other hand, previous studies show that stronger religious belonging (or strength of religiosity) is often associated with higher chances of voting for right-wing Conservative Party of Canada due to religious concerns toward controversial issues like legalizing same-sex marriage [6] [7]. Hence, religious belonging adds complexity to affecting voters' voting behaviour. Meanwhile, older generation voters tend to be more traditional and conservative, for instance, they may be less likely to accept new policies brought up by the Liberal Party of Canada that support the LGBTQ community [6]. The Liberal Party of Canada, which aims to achieve a wide range of social equality, may not be attractive to older voters. However, the party's objective of social equality also includes support for all citizens regardless of their classes which can be seen from its signature policies: universal health care and Canada pension plan [7], which may sound attractive to older voters. Hence, age should also be viewed as an important element in determining voters' voting behaviour.

Knowing that the Liberal Party of Canada has dominated federal politics of Canada for a long period of time [4], it would be interesting to predict if the Liberal Party of Canada will win in the upcoming election and continue to hold power in the federal politics of Canada with reference to the population's religion belonging and age composition.

1.2 Research question and data

In this study, we aim to use eligible Canadian voters' sense of religious belonging, and their age to predict which federal party will win in the next Canadian federal election by comparing the predicted voting percentages of the Liberal Party of Canada, Conservative Party of Canada and New Democratic Party. We specifically choose to compare results from these three federal parties as they are the strongest candidates for the upcoming election according to past election results [4].

The datasets we use are the General Social Survey (GSS) and the Canadian Election Study (CES). CES serves as the survey data and GSS surveys as the census data. We assume that the General Social Survey (GSS) dataset is a representative sample of the Canadian eligible voters population. For instance, the proportion of census respondents according to gender, age, and region categories should match with that of the true population.

1.3 Hypotheses

Given that total office time of all prime ministers who represent the Liberal Party of Canada is the highest among all federal political parties in Canada [4] and its inclusion of diversity in the Liberal Party of Canada's ideology [6] [7] which is beneficial to attracting support from a broad spectrum of voters, we hypothesize that the Liberal Party of Canada will win in the next Canadian federal election.

1.4 Problem

One of the main difficulties in this study is to account for the demographic differences in the survey data and the census data. Since the survey data was not collected using probability sampling, the demographic proportion of respondents in the survey data is significantly different from that in the true population (as reflected in the census data). In our study, we will utilize post-stratification to alleviate this problem by re-weighting predictions/estimates for different subgroups in the population (will be discussed further in section 3.3 Post-stratification).

Part 2: Data

2.1 Datasets

Our study consists of two datasets.

2.11 2021 Canadian Election Study [8]

The first dataset, which serves as the survey data, is the 2021 Canadian Election Study (CES). This is an online survey data which is a sample of Canadian citizens and permanent residents who were aged 18 or older, with quotas set for each region, gender and age. Hence, it adopted a non-probability sampling method, quota sampling.

The dataset is a two-wave panel with a modified rolling cross-section during the campaign period and a post-election recontact wave. The first wave of data (Campaign Period Survey) was collected between August 17, 2021 – September 19, 2021, with a sample size of 20,968. The second wave of data (Post-election survey) was collected between September 23, 2021 – October 4, 2021, with a sample size of 15,069. In this study, we will only use the first wave data (Campaign Period Survey) since the second wave data does not include religious belonging as a variable, which is one of our variables of interest.

2.12 2017 General Social Survey [9]

The second dataset, which serves as the census data, is the 2017 General Social Survey (GSS) with a sample size of 20968. Its target population was all non-institutionalized persons 15 years of age or older, living in the 10 provinces of Canada. The data was collected between February 1, 2017 and November 30, 2017, using a computer-assisted telephone interviewing method.

The census sample size is based on a stratified design employing probability sampling. Firstly, stratification was done at the census metropolitan area (CMA) level, resulting in 27 strata. For each province, minimum sample sizes were determined that would ensure estimates would have acceptable sampling variability at the stratum level. Once these stratum sample size targets had been met, the remaining samples were allocated to the strata in a way that balanced the need for precision of both national-level and stratum-level estimates.

Besides, it incorporates a two-stage sampling design. The sampling units are the groups of telephone numbers in which all numbers in each group are associated with the same address (hence the same household). The final stage units are individuals within the identified households. Therefore, only one eligible person per household was interviewed.

2.2 Cleaning process

Since our target population is eligible Canadian voters' and the minimum voting age in Canada is 18, we are only interested in responses from Canadian citizens aged 18 or above. Therefore, we have to clean the datasets to make sure observations are only from Canadian citizens aged 18 or above. In the survey data, observations were collected only from Canadian citizens and permanent residents who are aged 18 or older, so there is no further action needed in this case. However, in the census data, there are 424 respondents who are under-aged (i.e.aged below 18). It is acceptable to filter away these observations since they only account for around 2.058 % of the total number of observations. To ensure matching variable names between the two datasets, we created a new variable "age" in the survey data, which captures the responses in the variable "cps21_age" which essentially represents respondents' age. (Note that the census dataset has age saved under the variable "age".) Thereby, both datasets have "age" as the variable name for respondents' age.

In the survey data, "cps21_votechoice" is a variable which represents responses to the question "Which party do you think you will vote for?". The response is:

- 1 when the respondents think they will vote for the Liberal Party of Canada
- 2 when the respondents think they will vote for Conservative Party of Canaada
- 3 when the respondents think they will vote for New Democratic Party
- 4,5,6 or 7 when the respondents think they will vote for other federal parties.

We are interested in measuring the chance of eligible voters supporting the Liberal Party of Canada, Conservative Party of Canada and New Democratic Party respectively. So, we create a binary variable "vote liberal" as an indicator variable of the option "Liberal Party of Canada" by inserting a 1 when the response in "cps21_votechoice" is 1 and inserting a 0 when the response in "cps21_votechoice" is not 1. Similarly, we create a second binary variable "vote conservative" as an indicator variable of the option "Conservative Party of Canada" by inserting a 1 when the response in "cps21_votechoice" is 2 and inserting a 0 when the response in "cps21_votechoice" is not 2. Lastly, we create the third binary variable "vote NDP" as an indicator variable of the option "New Democratic Party" by inserting a 1 when the response in "cps21_votechoice" is 3 and inserting a 0 when the response in "cps21_votechoice" is not 3.

Although the identical question about religious belonging (i.e. "In your life, you would say religion is:_____") is asked in both datasets, there is a mismatch in terms of the response option codes. The census data has options: "Very Important", "Somewhat important", "Not very important", "Not at all important" and "Don't know" saved under its religious belonging variable "regilion_importance" (This is not a typo, see below) ; while the survey data has options: 1, 2, 3, 4, 5 saved under its religious belonging variable "cps21_rel_imp". Therefore, we created a new variable, named "religious belonging", in both datasets to define religion belonging using the options provided in the census data. Numerical values in the survey data are transformed to match with the census data by referring to the category each number represents (1 = "Very Important", 2 = "Somewhat important", 3 = "Not very important", 4 = "Not at all important" and 5 = "Don't know"). As we want to avoid incorrectly interpreting respondents who reported "Don't know" as their religious belonging level, we will filter away observations with "Don't know" as the response to this question. Note that these "Don't know" responses only account for a very small proportion in both datasets (Census data: 1.131% ; Survey data: 1.645 %), hence it is acceptable to filter them away without creating major impacts to the representation of the datasets.

After conducting the above cleaning procedures, we select our variable of interests, "age" and "religion belonging", in both datasets and additionally binary variables "vote liberal", "vote conservative" and "vote NDP" in the survey data to conduct regression later.

Lastly, we remove observations with missing data (i.e. not available data) on any one of our variable of interests from the datasets. The census data after cleaning has size 1.9699×10^4, which means 4.3830696% of observations are deleted after cleaning. Note that there is no missing data in the census data, all observations that are removed are due to invalid responses (as explained above). The survey data after cleaning has the size 9926. (Will discuss in greater detail about missing data in section 3.2.)

Note to TA grader: "regilion_importance" is not a typo, it was really coded as regilion_importance in the provided data set but we should understand it as "religion_importance".

2.3 Important variables

1. age

It is a numerical variable indicating the age of the respondents. In both the census and the survey datasets, the minimum value is 18. This variable has a maximum value of 80 in the census dataset while that in the survey dataset is 97. Note that the variable age is corrected to the nearest 1 decimal place in the census data set while it is an integer variable in the survey dataset. In both cases, we will assume it is a continuous variable.

2. religious belonging

It is a categorical variable, with possible options of "Very Important", "Somewhat important", "Not very important", "Not at all important" and "Don't know", indicating respondents' subjective importance level of religion in their lives. However, all observations with "Don't know" as the response have been removed from both datasets during the cleaning process.

3. vote liberal

It is a binary indicator variable, indicating whether the respondents will vote for the Liberal Party of Canada. It is "1" if the respondents think they will vote for the Liberal Party of Canada and "0" otherwise.

4. vote conservative

It is a binary indicator variable, indicating whether the respondents will vote for Conservative Party of Canada. It is "1" if the respondents think they will vote for Conservative Party of Canada and "0" otherwise.

5. vote NDP

It is a binary indicator variable, indicating whether the respondents will vote for New Democratic Party. It is "1" if the respondents think they will vote for New Democratic Party and "0" otherwise.

2.4 Numerical summary

2.41 Statistics about census data by religious belonging level

Table 1: Statistics about census data

In your life, you would say religion is:	Group size	Mean age	Standard deviation of age
Not at all important	3864	46.767	16.746
Not very important	3035	47.752	16.953
Somewhat important	5914	53.788	16.535
Very important	6886	57.772	16.281

There are around 0.196 % of census respondents who think religion is not at all important to them, and around 0.154 % of census respondents who thinks religion is not very important to them. The mean ages in these two groups are similar, which are 46.767 and 47.752 respectively.

There are around 0.3 % of census respondents who think religion is somewhat important to them and 0.35 % of census respondents who thinks religion is very important to them. The mean age in the "Somewhat important" group is 53.788. Meanwhile, the mean age in the "Very important" group is 57.772, which is slightly higher than that of the "Somewhat important" group and is the highest among all four groups.

Overall, around 64.978% of census respondents hold a higher level of sense of religious belonging (i.e. "Very important" and "Somewhat important") in the census data set and they on average older in age than those of lower level ("Not very important" and "Not at all important"). Nonetheless, all groups have similar standard deviations in age, which fall between 16.5 and 17.0.

2.42 Statistics about survey data by religious belonging level

Table 2: Statistics about survey data by religion importance level

In your life, you would say religion is:	Group size	Mean age	Standard deviation of age
Not at all important	1946	51.690	15.934
Not very important	2750	54.047	16.259
Somewhat important	3244	53.579	16.973
Very important	1986	52.417	17.522

There are around 0.196 % of survey respondents who think religion is not at all important to them and 0.277 % of survey respondents who thinks religion is not very important to them. The mean age in "Not at all important" group is 51.69 with standard deviation 15.934. The mean age in the "Not very important" group is 54.047 , which is the highest among all four groups. Its age standard deviation is 16.259.

There are around 0.327 % of survey respondents who think religion is somewhat important to them, which is the highest among all four groups. Meanwhile, there are around 0.2 % of survey respondents who thinks religion is very important to them. The mean ages in these two groups are similar, which are 53.579 and 52.417 respectively. The age standard deviation in "Somewhat important" group is 16.973 while that in "Very important" is 17.522, which is the highest among all four groups.

Overall, around 52.690% of survey respondents hold a higher level of sense of religious belonging (i.e. "Very important" and "Somewhat important") in the survey data set but the average ages for each religious belonging group are similar. Note that proportion of observations in each religious belonging level group and mean age of some groups are quite different between the census data and the survey data.

2.43 Statistics about survey data by voting preferences

Table 3: Statistics about survey data by whether respondents will vote for the Liberal Party of Canada

Will you vote for the Liberal Party of Canada?	Group size	Mean age	Standard deviation of age
No	7148	52.589	16.848
Yes	2778	54.434	16.289

As shown in Table 3, around 27.987% of survey data respondents revealed that they would vote for the Liberal Party of Canada. Besides, people who will vote for the Liberal Party of Canada are on average older than those who will not.

Table 4: Statistics about survey data by whether respondents will vote for Conservative Party of Canada

Will you vote for the Conservative Party of Canada?	Group size	Mean age	Standard deviation of age
No	7144	51.913	16.811
Yes	2782	56.167	16.060

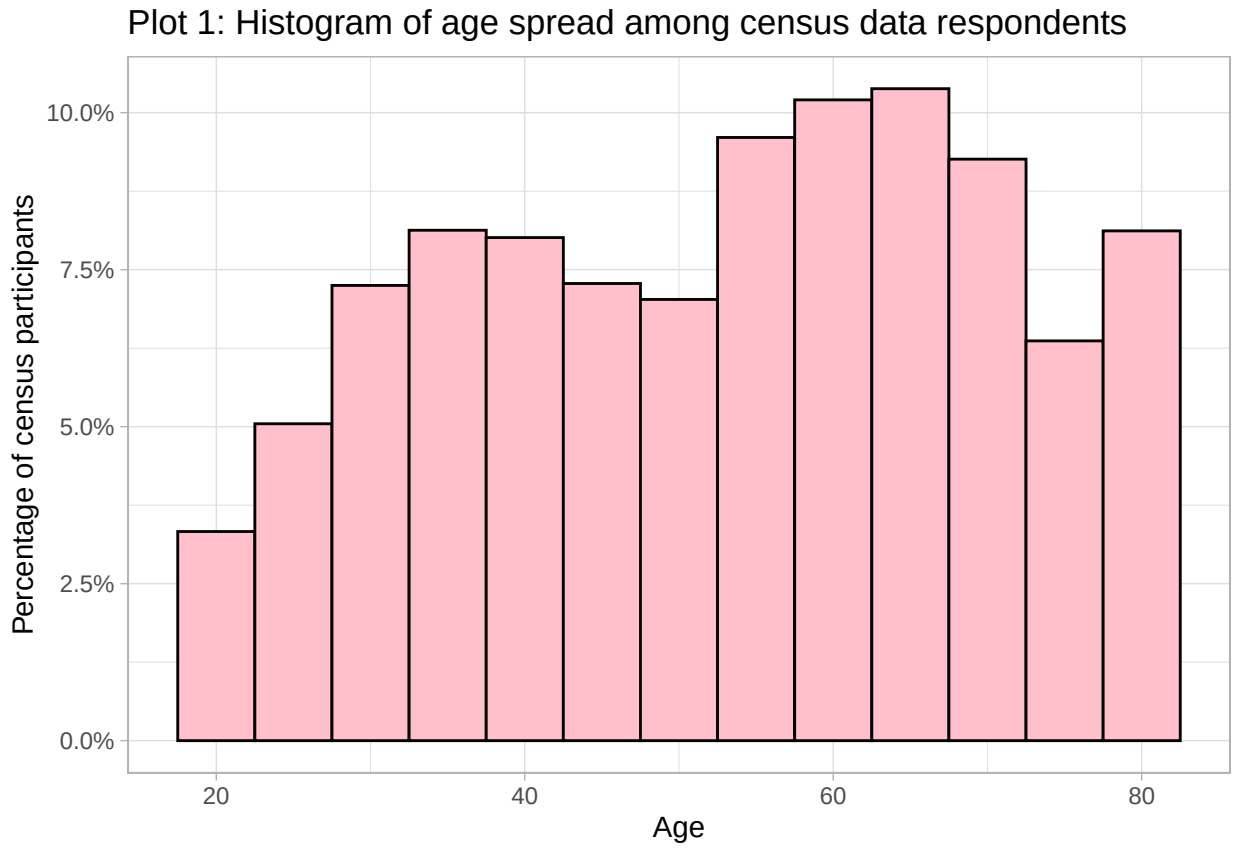
As shown in Table 4, around 28.027% of survey data respondents revealed that they would vote for Conservative Party of Canada. Besides, people who will vote for Conservative Party of Canada are on average older than those who will not.

Table 5: Statistics about survey data by whether respondents will vote for New Democratic Party

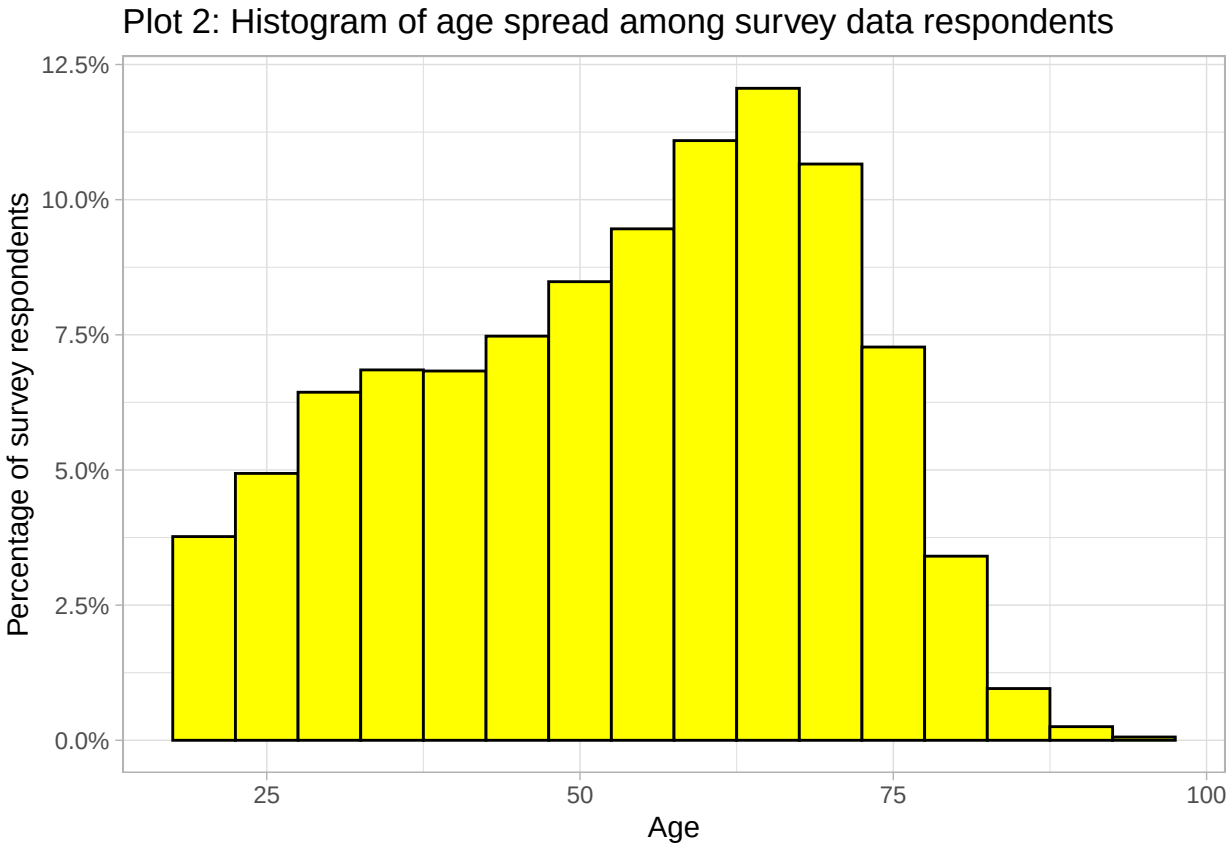
Will you vote for New Democratic Party?	Group size	Mean age	Standard deviation of age
No	8334	54.626	16.101
Yes	1592	45.147	17.584

As shown in Table 4, around 16.039% of survey data respondents revealed that they would vote for New Democratic Party. Besides, people who will vote for Conservative Party of Canada are on average younger than those who will not.

2.5 Plots



This histogram shows the distribution of age among the census data respondents. As shown in the plot, the distribution is approximately bimodal with peaks at around 35-year-old and 65-year-old. The overall mean age of the census data respondents is 52.874 and the overall age standard deviation is 17.146. Additionally, note that 29.149% of census respondents are aged 65 or above.



This histogram shows the distribution of age among the survey data respondents. As shown in the plot, the distribution is fairly unimodal with a peak at around 65-year-old. The distribution is also slightly right-skewed since we can observe a tail on the right of the plot. Most of the survey data respondents are middle-aged. Compared to the census data, the proportion of respondents who are aged 65 or above (15%) is significantly lower. The overall mean age of the survey data respondents is 53.105 and the overall age standard deviation is 16.713.

All analysis for this report was programmed using `R version 4.0.2` [10], the above tables and plots are generated using `R knitr package` and `R Tidyverse package`[11].

Part 3: Methods

In this study, our parameters of interest are the overall proportions of eligible voters voting for the Liberal Party of Canada, Conservative Party of Canada and New Democratic Party respectively in the next Canadian federal election. We are interested in predicting these proportions based on eligible Canadian voters' level of sense of religious belonging and age.

3.1 Logistic regression [12]

Logistic regression is a statistical method that uses a linear combination of one or more independent variables (they can be continuous or discrete) to predict the logit of the probability of whether an event takes place. Note that if P is the corresponding probability, we call $ln(\frac{P}{1-P})$ as the logit. In other words, given predictors (they can be continuous or discrete), we can use a logistic regression model to predict the value of a response variable which consists of only two classes. Suppose we label the response variable values with 0 and 1. Then, we can use the logistic regression model to predict the logit and convert it back to the corresponding probability of value labelled 1. Note that this probability can vary between 0 (meaning certainly the value 0) and 1 (meaning certainly the value 1).

3.11 Model Specifics

In this study, we will adopt logistic regression as the prediction model. Our model equations are shown below:

Model 1:

$$ln(\frac{P_{liberal}}{1-P_{liberal}}) = \beta_0 + \beta_1 x_{age} + \beta_2 I_{Not\,very\,important} + \beta_3 I_{Somewhat\,important} + \beta_4 I_{Very\,important}$$

Model 2:

$$ln(\frac{P_{conservative}}{1-P_{conservative}}) = \beta_0 + \beta_1 x_{age} + \beta_2 I_{Not\,very\,important} + \beta_3 I_{Somewhat\,important} + \beta_4 I_{Very\,important}$$

Model 3:

$$ln(\frac{P_{NDP}}{1-P_{NDP}}) = \beta_0 + \beta_1 x_{age} + \beta_2 I_{Not\,very\,important} + \beta_3 I_{Somewhat\,important} + \beta_4 I_{Very\,important}$$

where

- $P_{liberal}$ represents the probability of the individual voting for the Liberal Party of Canada in the next Canadian federal election.
- $P_{conservative}$ represents the probability of the individual voting for Conservative Party of Canada in the next Canadian federal election.
- P_{NDP} represents the probability of the individual voting for New Democratic Party in the next Canadian federal election.
- x_{age} represents the individual's age.
- $I_{Not\,very\,important}$ is 1 when the individual rates their religious belonging level as "Not very important" and 0 otherwise.
- $I_{Somewhat\,important}$ is 1 when the individual rates their religious belonging level as "Somewhat important" and 0 otherwise.
- $I_{Very\,important}$ is "1" when the individual rates their religious belonging level as "Very important" and "0" otherwise.
- β_0 is the logit $ln(\frac{P}{1-P})$ of an individual whose religious belonging level is "Not at all important" and age is 0.
- β_1 is the change in logit $ln(\frac{P}{1-P})$ for a unit increase in age.

- β_i (for i = 2, 3, or 4) coefficients represent the change in logit $ln(\frac{P}{1-P})$ of an individual whose religious belonging level is what the attached indicator variable represents compared to the logit of an individual with "Not at all important" religious belonging level, given that age is fixed.

These three model allow us to predict the logit $ln(\frac{P}{1-P})$ and hence the probability P of an individual voting for each of the three parties in the next Canadian federal election based on their age and religious importance level. (Model 1 aids predicting the probability of an individual voting for the Liberal Party of Canada; Model 2 aids predicting the probability of an individual voting for Conservative Party of Canada; Model 3 aids predicting the probability of an individual voting for New Democratic Party.)

3.12 Justification of model

There are four assumptions associated with logistic regression models:

1. The outcome is binary

All three of our response variables "vote liberal", "vote conservative", "vote NDP" are binary variables since they are indicators of whether the respondents will vote for each party (Liberal Party of Canada, Conservative Party of Canada and New democratic Party), they can only take on two values: 1 for yes and 0 for no. So, this assumption is satisfied.

2. Linearity in the logit for continuous variables

The only continuous variable in our model is age (although it can only take on integer values in the survey data, we will treat it as a continuous variable here). This assumption states that the continuous variable "age" should be linear to the logit of our response variable "vote liberal". As shown in Plot 3-5 (refer to section 7.2 Supplementary Materials), age is quite linearly associated with the outcome variables "vote liberal", "vote conservative" and "vote NDP" in the logit scale respective since we can generally fit a straight line to describe their relationships. So, this assumption is satisfied.

3. Absence of multicollinearity

Multicollinearity arises when a predictor is related/correlated to another predictor in the regression model. Since the only predictors we have in our models are "age" and "religion belonging", they are two very distinct concepts and one most probably cannot be used to explain the other variable. Hence, it is reasonable to assume that they are not correlated to each other. So, this assumption is satisfied.

4. Lack of strongly influential outliers

By computing standard residuals and Cook's distance [13] (refer to section 7.2 Supplementary Materials) for each of our three proposed model, we found no outliers or influential observations in our survey dataset. So, this assumption is satisfied.

Since all 4 assumptions are satisfied, the logistic regression models are appropriate models for this study.

3.2 Missing data

In the cleaning process, we removed observations with missing data (i.e. no data value is stored as an observation in a variable) from the survey data. There are 4660 observations with missing responses in the voting party preference variable "cps21_votechoice", age variable "age", or in the religious belonging variable "cps21_rel_imp". It is important to note that these missing data account for around 22.224% of the observations in the original survey data. However, we still opt to delete these observations since we do not want to impose any inaccurate information on the observations. Imputation often involves borrowing information from other observations (that are not missed data observations). Yet, we want to treat each respondent as an independent observation and would not risk borrowing information from other observations and assume the missed data observations behave similarly or follows a general pattern as other observations. Therefore, we chose to remove these missed data from our data analysis to avoid including questionable assumptions.

3.3 Post-stratification

3.31 What is post-stratification? [14]

In order to estimate the proportions of eligible voters in the population voting for the Liberal Party of Canada, Conservative Party of Canada and New Democratic Party respectively in the next Canadian federal election, we will conduct post-stratification. This is a correction method imposed on non-probability-based samples to adjust the weights of estimates for different subgroups in the population with reference to the subgroup proportions in the population. So, the overall estimates are more representative values for the actual target population. In the following data analysis, we will assume our (cleaned) census data is representative of the true population and all demographic distributions observed in this census data are consistent with the true population.

3.32 Why is post-stratification appropriate?

This is an appropriate method since correction is needed on our survey sample which is obtained through non-probability sampling. As discussed earlier, the survey data used in this study is a based on non-probability sampling (quota sampling). Therefore, the survey data may be non-representative of the true population. In particular, we can see from section 2.4 Numeric Summary and section 2.5 Plots that the proportions of respondents distributions by religious belonging groups and age are significantly differences between the survey data and the census data (which is a representative sample of the true population).

Additionally, there is a serious non-response rate occurred in the survey sample where we had to remove almost 20% of the observation due to missing data. Post-stratification allows us to 're-weight' in a way that includes the uncertainty in these missing data so that the final population-level estimate is not evaluated with restriction by the potentially biased small survey sample.

3.33 Post-stratification process

The implementation of post-stratification involves 3 steps:

1. Partitioning the data into demographic cells.

In our study, we will use both "age" and "religious belonging" as the variables to partition the census data in cells because

- they are both predictors in our regression model built earlier;
- they are believed to be likely to influence voter outcome according to past studies [4] [5]; and
- we observed a significant difference in the proportion of observations in each religious belonging level group and age distribution between the survey data and the true population census data.

The partitioning of data can be done by grouping the census data according to the values stored under "age" and "religion belonging". For example, respondents with "age" 20 and "religion belonging" "Very important" will be grouped to one cell, respondents with "age" 20 and "religion belonging" "Somewhat important" will be grouped to another cell, respondents with age 30 and religion belonging "Somewhat important" will be grouped to the third cell, etc. We then calculate the relative proportion of each demographic cell in the population by summarizing the number of observations in each cell divided by the total number of observations in the census data (i.e. 1.9699×10^4). This set of relative proportion values is saved under a new variable "prop" in the census data.

Here, we assume that, within a demographic cell, the sample is drawn at random from the larger population that includes all eligible voters who share the same demographic characteristic (which is used to define the demographic cell), this yields an unbiased estimate.

2. Estimating response variable for each cell using regression models.

In section 3.1 Logistic Regression, we built three separate regression models which relates individual-level survey "vote liberal", "vote conservative", "vote NDP" responses to two demographic characteristics, "age" and "religion belonging", based on the survey data. Using the models, we can compute the probabilities

$P_{liberal}$, $P_{conservative}$, P_{NDP} which indicate the probabilities of an individual voting for the Liberal Party of Canada, Conservative Party of Canada and New Democratic Party respectively based on their age and religious belonging level.

We apply each model to the census data to predict the logits of "vote liberal" responses, logits of "vote conservative" responses and logits of "vote NDP" responses. Hence, we obtain the probabilities $P_{liberal}$, $P_{conservative}$, P_{NDP} for all observations. Next, we save the results of these predictions under new variables created in census data "estimate liberal", "estimate conservative" and "estimate NDP" respectively.

To calculate the mean response for each cell, we group observations in the census data by "age" and "religion belonging" (i.e. by the cells defined in Step 1), then sum up all estimates that belong to the same cell and divide the sum by total number of observations in that particular cell. Since we have three estimates (i.e. "estimate liberal", "estimate conservative" and "estimate NDP"), we will do this procedure once for each of the estimates. The sets of cell-level mean values is saved under new variables "mean cell estimate liberal", "estimate conservative" and "estimate NDP" created in the census data.

3. Aggregating the cell-level estimates up to a population-level estimate by weighting each cell by its relative proportion in the population.

This can be done by multiplying each cell-level mean estimate (i.e. values stored under the variables "mean cell estimate liberal", "estimate conservative" and "estimate NDP") by the relative proportion of that cell in the population (i.e. value stored under the variable "prop"). Lastly, we sum up all weighted cell means to obtain the total population-level estimates.

3.34 Post-stratification equation and parameters

Below is the post-stratification equation:

$$\hat{y}^{PS} = \frac{\sum_i N_i \hat{y}_i}{\sum_i N_i} = \sum_i \frac{N_i}{19699} \hat{y}_i$$

where

- $\hat{y}^{PS}$ is the post-stratified mean population-level estimate: the proportion of eligible voters in the population voting for the Liberal Party of Canada/ Conservative Party of Canada/ New Democratic Party in the next Canadian federal election.
- $\hat{y}_i$ is the mean estimate of the i_{th} demographic cell: the proportion of eligible voters belonging to i_{th} demographic cell voting for the Liberal Party of Canada/ Conservative Party of Canada/ New Democratic Party in the next Canadian federal election.
- N_i is the population size of the i_{th} demographic cell.
- $\sum_{i=1}^{2347} N_i$ is the total population size which is 19699 in our census data.
- $\frac{N_i}{19699}$ is the relative proportion of the i_{th} demographic cell in the population, it is stored under the variable "prop".

Note that demographic cells are defined as groups of observations that share the same age and same religious belonging level.

All analysis for this report was programmed using `R version 4.0.2` [10].

Part 4: Results

4.1 Regression model

4.11 Fitted model equations

The logistic regression models obtained after fitting the survey data are:

Model 1:

$$ln(\frac{P_{liberal}}{1-P_{liberal}}) = -1.449+0.006x_{age}+0.108I_{Not\,very\,important}+0.279I_{Somewhat\,important}+0.166I_{Very\,important}$$

Model 2:

$$ln(\frac{P_{conservative}}{1-P_{conservative}}) = -2.072+0.112x_{age}+0.337I_{Not\,very\,important}+0.622I_{Somewhat\,important}+0.157I_{Very\,important}$$

Model 3:

$$ln(\frac{P_{NDP}}{1-P_{NDP}}) = 0.238-0.137x_{age}-0.188I_{Not\,very\,important}-0.435I_{Somewhat\,important}+-0.034I_{Very\,important}$$

where

- $P_{liberal}$ represents the probability of the individual voting for the Liberal Party of Canada in the next Canadian federal election".
- $P_{conservative}$ represents the probability of the individual voting for Conservative Party of Canada in the next Canadian federal election".
- P_{NDP} represents the probability of the individual voting for New Democratic Party in the next Canadian federal election" .
- x_{age} represents the individual's age.
- $I_{Not\,very\,important}$ is 1 when the individual rates their religious belonging level as "Not very important" and 0 otherwise.
- $I_{Somewhat\,important}$ is 1 when the individual rates their religious importance level as "Somewhat important" and 0 otherwise.
- $I_{Very\,important}$ is 1 when the individual rates their religious belonging level as "Very important" and 0 otherwise.

4.12 Coefficient interpretation

The intercepts (-1.499 in Model 1, -2.072 in Model 2, 0.238 in Model 3) represents the logit of the probability (i.e. $log(\frac{P}{1-P})$) when the individual has age 0 and rates their religious belonging level as "Not at all important". However, since eligible voters must be at least 18 years old, the above explanation is just a statistical interpretation.

- β_1 (0.006 in Model 1, 0.112 in Model 2, - 0.137 in Model 3) is the average change in the logit of the probability (i.e. $log(\frac{P}{1-P})$) for every one unit increase in age.
- β_2 (0.108 in Model 1, 0.337 in Model 2, -0.188 in Model 3) is the average change in the logit of the probability (i.e. $log(\frac{P}{1-P})$) for an individual with religion belonging level as "Not very important" compared to an individual with religion belonging level as "Not at all important", given that age is fixed. 0.622

- β_3 (0.279 in Model 1, 0.622 in Model 2, - 0.435in Model 3) is the average change in the logit of the probability (i.e. $log(\frac{P}{1-P})$) for an individual with religion belonging level as "Somewhat important" compared to an individual with religion belonging level as "Not at all important", given that age is fixed.
- β_4 (0.166 in Model 1, 0.157 in Model 2, -0.034 in Model 3) is the average change in the logit of the probability (i.e. $log(\frac{P}{1-P})$) for an individual with religion belonging level as "Very important" compared to an individual with religion belonging level as "Not at all important", given that age is fixed.

Overall, eligible voters who are older on average have a higher (logit and hence) probability to vote for the Liberal Party of Canada and Conservative Party of Canada in the next Canadian federal election and have a smaller probability to vote for New Democratic Party.

Besides, the average (logit and hence) probability is the highest among eligible voters who rate their religious belonging level as "Somewhat important", followed by those with religious belonging level as "Very important" in Model 1 and Model 2. In general, eligible voters with higher religious belonging levels (i.e. "Very important" and "Somewhat important") yield higher chances to vote for the Liberal Party of Canada and Conservative Party of Canada in the next Canadian federal election compared to those with lower religion belonging levels (i.e. "Not very important" and "Not at all important").

Conversely, the average (logit and hence) probability is the lowest among eligible voters who rate their religious belonging level as "Somewhat important", followed by "Not very important" as the second lowest, "Very important" as the third lowest and "Not at all important" as the highest in Model 3. Therefore, eligible voters with extreme religious belonging level (i.e. "Not at all important" and "Very important") yield higher chances to vote for New Democratic Party in the next Canadian federal election compared to those with less extreme religious belonging level (i.e. "Somewhat important" and "Not very important").

4.13 Sensibility of models

1. Model 1:

Model 1 suggests that voters with higher sense of religious belonging have higher chances to vote for the Liberal Party of Canada. It is a sensible result since the Liberal Party of Canada's aim to achieve social equality including religious freedom. Voters with a stronger sense of religious belonging may be interested in supporting parties that aim to guarantee a safe and respectful environment for them to practice religious activities and therefore have a higher chance to vote for the Liberal Party of Canada. Model 1 shows that there is a positive association between age and the probability of voting for the Liberal Party of Canada, which is still a sensible result. Since the Liberal Party of Canada's objective of achieving social equality includes providing resources and opportunities to all citizens regardless of their classes, this may sound attractive to older voters. The elderly usually do not participate actively in the labour market and some of them may experience financial difficulties or face discrimination since they no longer contribute to the country's economy. The objective of social equality in the Liberal Party of Canada's ideology may direct the policies to alleviate these social disparity problems. So, older voters may be interested in voting for the Liberal Party of Canada.

2. Model 2:

Model 2 suggests that voters with higher level of religious belonging and older voters have higher chances to vote for Conservative Party of Canada which is in line with past studies conclusions [6] [7].

3. Model 3:

Model 3 suggests that voters with less extreme level of religious belonging and younger voters have higher chances to vote for New Democratic Party. This is a reasonable result since New Democratic Party focuses on ensuring affordable cost of living and economy of the country which are interests of the general public instead of specifically targeting at religious citizens. On the other hand, news has reported that New Democratic Party has been working on attracting young voters using digital platforms recently [15], so it is sensible that younger voters have higher chances to vote for New Democratic Party.

4.2 Post-stratification results

The partitioning of data is done by grouping the census data according to the value stored under "religion belonging" and "age" and we obtain a total of 2437 cells, meaning that there are 2437 different combinations of "religion belonging" and "age" values. By post-stratification, the overall popular vote in the next Canadian federal election of

- Liberal Party of Canada is 28.055%,
- Conservative Party of Canada is 29.532%,
- New Democratic Party is 15.883%.

Among these three major federal parties, Conservative Party of Canada has the highest overall popular vote in the next Canadian federal election, hence we predict that Conservative Party of Canada will win the upcoming Canadian federal election.

This is a reasonable result. Observe that Liberal Party of Canada and Conservative Party of Canada have very similar percentage of overall popular vote while New Democratic Party has a significantly lower percentage. This can be explained by the fact that Liberal Party of Canada and Conservative Party of Canada have been the biggest two parties and it is usually these two rival parties who win most of the votes. As mentioned in past research, voters with stronger religious belonging tend to vote for the Conservative Party of Canada [6] [7]. According to the demographic distribution of the census data presented in the Data section, more than half of the population (around 65%) hold a higher level of religious belonging. Therefore, it is expected that the proportion of popular vote for Conservative Party of Canada would be the highest due to the effect of religion belonging on a relatively high proportion of high religion belonging level sub-population.

Part 5: Conclusions

5.1 Brief recap

In this study, we used eligible Canadian voters' sense of religious belonging, and their age to predict which federal party will win in the next Canadian federal election by comparing the predicted voting percentages of the three main federal parties (Liberal Party of Canada, Conservative Party of Canada and New Democratic Party). We hypothesized that Liberal Party of Canada would win the next Canadian federal election based on its past winning history.

We first built (three) logistic regression models by fitting the survey data. Each model allows us to predict the probability of an individual voting for the Liberal Party of Canada/ Conservative Party of Canada/ New Democratic Party in the next Canadian federal election respectively based on their age and religious belonging level.

From the results of the model for Liberal Party of Canada and the model for Conservative Party of Canada, we found that both age and religious belonging level are positively associated with (the logit and hence) the probability of voting for these two parties in the next Canadian federal election. In other words, an increase in age and religious belonging level yields higher chances to vote for the Liberal Party of Canada/ Conservative Party of Canada in the next Canadian federal election. This can be explained by the social-equality-based ideology [1] of the Liberal Party of Canada which may sound appealing to religious activity practitioners, who usually have a higher sense of religious belonging, and older voters, who may be vulnerable and require government assistance. Meanwhile, Conservative Party of Canada's focuses on stability and emphasis on traditions [2] may attract higher sense of religious belonging and older voters as well. On the contrary, results of the model for New Democratic Party suggests that younger and less extreme sense of religious belonging voters are positively associated with (the logit and hence) the probability of voting for New Democratic Party in the next Canadian federal election. This is once again be explained by the party's target population [15] and economy-focused objectives [3].

We then used these models to make predictions on the census data observations and obtain three sets of estimated probabilities (one set for each party). Post-stratification is conducted to re-weight the estimates according to their corresponding age and religion belonging levels. So, the relative proportions of age and religious belonging levels subgroups in the population are taken into consideration while computing the overall population-level estimate. The overall population-level estimates of Liberal Party of Canada, Conservative Party of Canada and New Democratic Party after post-stratification are 28.055%, 29.532% and 15.883% respectively. Since Conservative Party of Canada has the highest popular vote percentage, we conclude that Conservative Party of Canada, instead of Liberal Party of Canada, will win the next Canadian federal election.

5.2 Limitations

We assumed that the (cleaned) census data is a representative sample of the Canadian eligible voters population. However, this assumption may not be solid since the census data only has size 1.9699×10^4 while the Canadian population is over 40 million in the year 2023 [16]. The relatively small sample size of the census data suggests that a large amount of responses from the rest of the Canadian population is being left out and hence any data analysis conducted on this census data may not fully reflect the situation in the true population. Also, all respondents who participated in the census were volunteers, making the sample vulnerable to volunteer bias.

Moreover, our single-level frequentist approach of logistic regression modelling post-stratification ignores uncertainty and only uses central cell estimates (i.e. mean estimates) to generate the population-level estimates. It may be insufficient to only report central estimates.

5.3 Next step

We adopted a frequentist logistic regression approach in this study and obtained a single value as the population-level estimate for each party's vote percentage in the next Canadian federal election. In future studies, we may consider using a Bayesian approach (e.g. building a Bayesian logistic regression model) which can produce plausible intervals for the population-level vote percentage estimates.

Additionally, we may consider including more predictors in the regression model. For example, we can include religion categories as a variable and observe the differences in voting behaviour by religion and post-stratify according to religion categories to obtain population-level estimate of each party's vote percentage in the next Canadian federal election.

Part 6: Bibliography

1. Rayside, D. (2023, November 8). Liberal Party of Canada. Encyclopedia Britannica. https://www.britannica.com/topic/Liberal-Party-of-Canada
2. Conservative Party of Canada. (2023). Retrieved November 14, 2023, from https://www.conservative.ca/
3. Canada's NDP. (2023). Canada's NDP. Retrieved November 14, 2023, from https://www.ndp.ca/ndp.ca

4.. Elections Canada (2023, October 10). Elections. https://www.elections.ca/content.aspx?section=ele&document=index&lang=e Forward. For Everyone. (n.d.).

5. Wilkins-Laflamme, S. (2016). The Changing Religious Cleavage in Canadians' Voting Behaviour. Canadian Journal of Political Science / Revue Canadienne de Science Politique, 49(3), 499–518.
6. Wilkins-Laflamme, S., & Reimer, S. (2019). Religion and Grassroots Social Conservatism in Canada. Canadian Journal of Political Science, 52(4), 865–881. https://doi.org/10.1017/S0008423919000544
7. Our Progress | Liberal Party of Canada. (2023). Retrieved November 11, 2023, from https://liberal.ca/our-progress/
8. Stephenson, Laura B., Allison Harell, Daniel Rubenson and Peter John Loewen. (2022). Canadian Election Study 2021 Survey Codebook.
9. Government of Canada, S. C. (2019, February 7). General Social Survey—Family (GSS). https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=335816
10. RStudio Team. (2020). *RStudio: Integrated Development for R*. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.
11. Allaire, J.J., et. el. *References: Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: September 22, 2023)
12. Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An Introduction to Logistic Regression Analysis and Reporting. The Journal of Educational Research, 96(1), 3–14. https://doi.org/10.1080/00220670209598786
13. Daignault, K. (2023). STA302 Module 8: Problem with Problematic Observations
14. Hanretty, C. (2020). An Introduction to Multilevel Regression and Post-Stratification for Estimating Constituency Opinion. Political Studies Review, 18(4), 630–645. https://doi.org/10.1177/1478929919864773
15. Jagmeet Singh wants to attract more young voters, and he's using TikTok to do it. (2021, July 26). The Globe and Mail. https://www.theglobeandmail.com/politics/article-jagmeet-singh-wants-to-attract-more-young-voters-and-hes-using-tiktok/
16. Canada's population reaches 40 million. (n.d.). Retrieved November 11, 2023, from https://www.statcan.gc.ca/en/subjects-start/population_and_demography/40-million

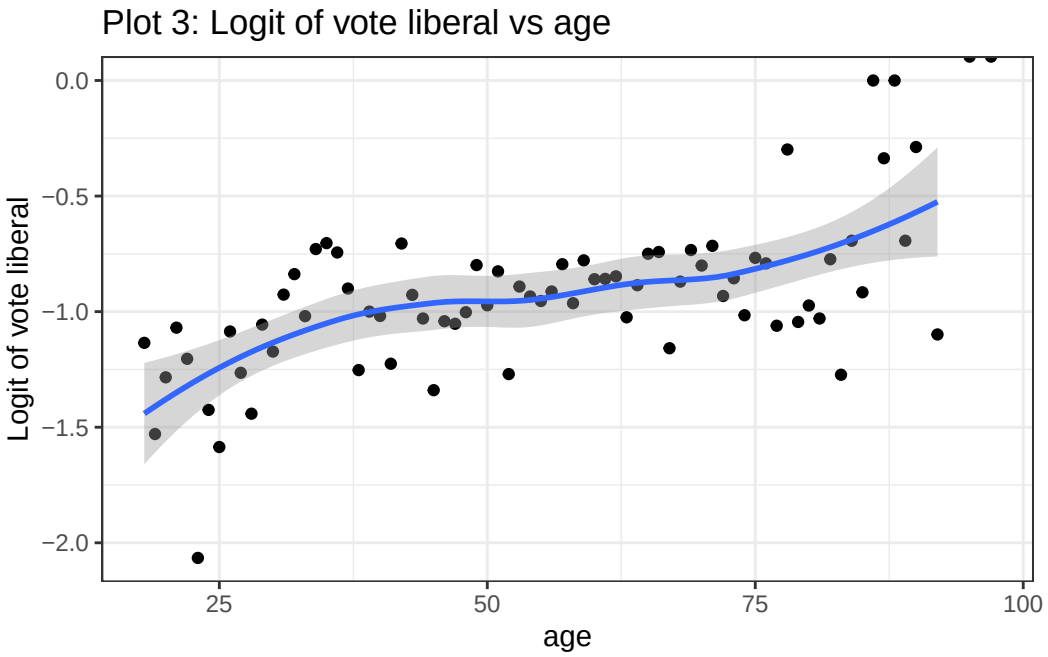
Part 7: Appendix

7.1 Generative AI Statement

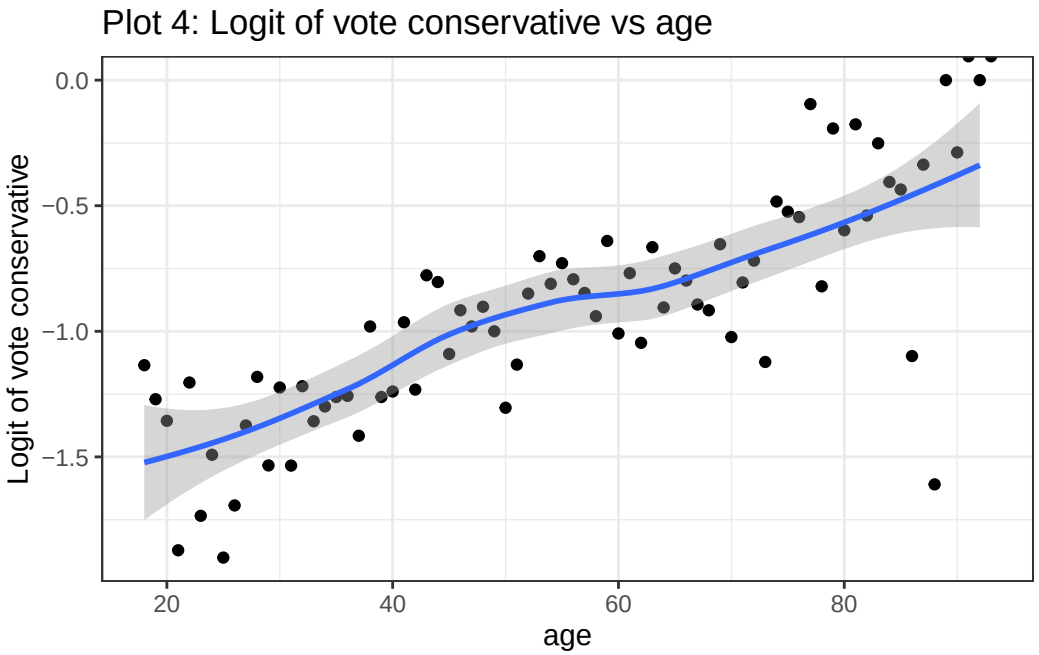
Generative artificial intelligence (AI) tools are not used in this study. These tools do not promise quality or accuracy in the information they provide. There are concerns about implausible content provided and violation of intellectual property (i.e. plagiarism) in these tools. Instead of merely copying information output from AI tools, I prefer incorporating knowledge taught in lectures and other credible sources such as textbooks and peer-reviewed academic journals and further provoking thoughts on the materials to develop critical thinking skills to finish assignments.

7.2 Supplementary Materials (model assumptions plots and results)

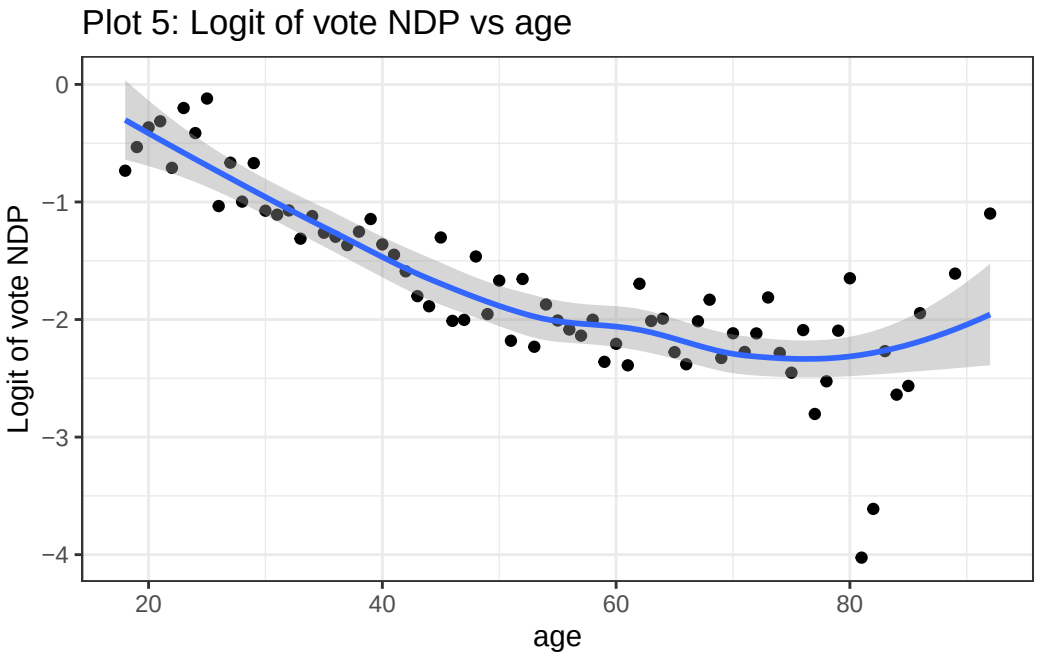
Plot 3: Logit of "vote liberal" vs "age"



Plot 4: Logit of "vote conservative" vs "age"



Plot 5: Logit of "vote conservative" vs "age"



Note that there may exist a slightly quadratic relationship between age and logit of "vote NDP" but we will assume they are linearly related for this assignment purpose.

Number of outliers [13]

We can use the `rstandard()` function in `R broom package`[11] to calculate the standardized residual values of the data points for each model and the cut off point of the absolute value of the standardized residual values is 4 [13]. We found that there is

- 0 outlier in Model 1
- 0 outlier in Model 2
- 0 outlier in Model 3

Number of influential points [13]

We can use the `cooks.distance()` function in `R broom package`[11] to calculate the Cook's distance of the data points in each model. The cut off point is the 50% quantile of F-distribution with degrees of freedom 2 (number of predictors) and 9924 (number of observations minus number of predictors) which is 0.6931956[13]. We found that there is

- 0 influential point in Model 1
- 0 influential point in Model 2
- 0 influential point in Model 3