An analysis: Prediction of proportion for voting Canadian Liberal party

STA304 - Assignment 3

GROUP 99: Qianlin Gao, Helen Han, Ruizhe Huang, Xiaoke Zeng

November 5, 2021

Introduction

Elections Canada is an independent, non-partisan institution of Parliament that the Chief Electoral Officer of Canada leads. Its major responsibility is to be ready to administer an election at all times. (Elections Canada, 2021)
The Liberal Party of Canada is the country's oldest and longest-serving federal political party. It dominated federal politics for all of Canada's history, holding power for about 70 years in the 20th century. (Liberal party in Canada, 2021)
The Liberal Party of Canada is committed to the view that the dignity of each person is the cardinal principle of a democratic society and the primary purpose of all political organization and activity in such a society. (The constitution of the Liberal Party of Canada, 2021) We decided to predict the proportion of voters who will vote for the Liberal party in 2019. In order to research this topic, we have the 2019 Canadian election phone survey data and the census data.
As we found from Statistics Canada, age has been found to matter a great deal regarding voting participation. Another potentially important factor is the family status of prospective voters. (Factors associated with voting 2015) Moreover, there are significant differences when it comes to women and men voting in the federal election. (Belmonte, 2019) Furthermore, research from the University of Waterloo suggests that the religious effect on voting behavior is one of Canada's strongest sociodemographic effects on vote choice. (How religious beliefs affect voting behavior in Canada 2019) According to these research papers, we found out that age, gender, religion has affiliation, and household size would impact the election voting results.
Therefore, Our research question is **By using age, gender, religion has affiliation, and household size to predict the proportion of voters who will vote for the Liberal party in 2019.**
Moreover, the 2015 Canadian federal election results show that the Liberal Party obtained around 40% of the votes. (Results of the 2015 Canadian federal Election,2021)
Therefore, we hypothesize that the Liberal Party will have around 40% of the votes in 2019 as well.
In the data section, there will be a data description in detail. Moreover, I have numerical summaries and graphical summaries of the variable of interest: age group, gender, religion has affiliation and household size.
In the methods section, we will introduce the logistic regression model, and the post-stratification explained in detail.
In the results section, by using logistic regression model and the post-stratification, as a result, predict the weighted proportion of respondents vote for the liberal party.
In the end, there will be a conclusion to summarize the main points of this report.

Terminology

citizenship: The citizenship in this report is specifically Canadian citizenship. It is divided into two cases: by birth and by naturalization. A person becomes a Canadian citizen by birth in two ways, either by birth

within Canada or by birth outside Canada but with at least one first-generation Canadian parent. Naturalized citizenship means that the person first immigrates to Canada and then applies to become a Canadian citizen a few years later. (Al Parsai, 2021)
logistic regression model: Logistic regression is the appropriate regression analysis to conduct when the dependent variable is binary. Logistic regression is a predictive analysis. Logistic regression does not make many of the key assumptions of linear regression models based on ordinary least squares algorithms. Particularly regarding linearity, normality, homoscedasticity, and measurement level. (Assumptions of logistic regression, 2021)
post-stratification: The process of adjusting the estimates, essentially a weighted average of estimates from all possible combinations of attributes such as age and gender. Each combination is sometimes called a "cell." (Multilevel regression with poststratification 2021)

Data

Data collection

The dataset we will be using is General Social Survey(Census data) and the Canadian Election Study(Survey data).
General Social Survey(Census data) is a short-form Census of Population questionnaire sent to (75%) of Canadian households, collecting basic demographic information, such as age, marital status, and language; a sample of 25% of Canadian households receive a long-form questionnaire. In the end, all respondents' responses were organized into the raw dataset; the census dataset we will be using is a cleaned version for better understanding with only a few variables detail stated in the data cleaning section. (Government of Canada, 2021)
The other dataset, "Survey data" is collected based on the Canadian Election Study, an annual survey of voting and other preferences, and further measurements that are likely to pertain to the political behaviors of Canadian voters. During the election campaign, telephone interviews were completed with 4,021 Canadian citizens. (The 2019 Canadian Election Study -Phone Survey, 2019) The datasets that will use are from the year 2019.

Data Cleaning Process

Since the dataset was downloaded from the website and was unorganized with lots of missing values and some variables that are of no use to my analysis, I will perform data cleaning at first.
For the **census dataset**, the original dataset contains 20602 observations and 81 variables with lots of missing values.
I removed the missing values for the following variables *citizenship status*, *age*, *sex*, *household size*, and *religion has affiliation*.
Since the age in the original dataset was with decimals, thus I round them into the integers as it is easier for analyzing. Moreover, we need to match the variable names in the census dataset and the survey dataset. Therefore, I renamed the *sex* in the census dataset to *gender* and renamed the *hh size* to *household size*.
Since only a Canadian citizen and 18 years old or older on election day can vote in a Canadian federal election (A guide to voting in the Canadian federal election, 2021). Thus, I only choose those who are 18 years old or older, and their citizenship status is Canadian by birth and naturalization.
Besides, I remove the people who do not know their religious affiliation in the census dataset.
In the end, no use variables were removed from the dataset, and only five critical variables were kept, including *citizenship status*, *age*, *sex*, *household size*, and *religion has affiliation*. Since these variables are the main factors for the election voting, I will use these variables to predict the probability of voting for the Liberal party in this report.
The **cleaned census dataset** contains 18773 observations and five variables without any missing values.
For the **survey dataset**, the original dataset contains 4021 observations and 278 variables with lots of

missing values.
I removed the missing values for the following variables *age*, *gender*, *religion has affiliation*, *household size*, *citizenship status*, and *vote liberal*.
Moreover, I removed those who do not know or refused to answer whether they have a religious affiliation and those who do not know or refused to answer whether they will vote for the Liberal party.
Some new variables were created in the survey dataset. The new variable *age* represents the age of the respondents until 2019, in years. It was calculated by the difference between 2019 and their birth year. The new variable *gender* is classified by the value in the survey question 3 when the value equal to 1 will be "Male" when the value equal to 2 will be "Female." Otherwise, it will be the "Other" gender. The new variable *religion has affiliation* classified by the value in the survey question 62, when the value equal to 21 will be "No religious affiliation," otherwise it will be "Has religious affiliation." The new variable *household size* is just the answer to survey question 71, representing the number of family members for each family. The new variable *citizenship status* is just the answer to survey question 1, representing whether the respondents are Canadian. The last new variable, *vote_liberal*, is classified by the answer to survey question 11. If the respondents choose to vote Liberal party, it will equal 1; otherwise, it will equal 0.
Since there is only one respondent belongs to the "Other" gender. Therefore, I removed this respondent as it would not have a large negative impact on my analysis. Moreover, I removed the household size is less than 0, that is, who do not know or refused to answer their household size. Furthermore, I only choose those 18 years old or older, and their citizenship status is Canadian. Since only a Canadian citizen and 18 years old or older on election day can vote in a Canadian federal election (A guide to voting in the Canadian federal election, 2021).
Eventually, I only kept 6 variables in the cleaned survey dataset, including *age*, *gender*, *religion has affiliation*, *household size*, *citizenship status* and *vote liberal*. Since these variables are the main factors for the election voting, I will use these variables to predict the probability of voting for the Liberal party in this report.
The **cleaned survey dataset** contains 2675 observations and 6 variables without any missing values.

Variable Description

Table 1: Description of variables that exist in the both data

Variable	Description
Age	The age of each respondents in 2019
Gender	Gender of the respondent
Citizenship_status	Canadian citizen by birth/naturalization
Religion has affiliation	Whether the respondent has Religious affiliation
Household size	Household size of respondent
Vote Liberal	Whether the respondent vote for Liberal

Table 1 describes selected essential variables that exist in both Census and Survey data after cleaning. Only a few variables were kept to give readers a picture of the critical factors for each respondent's response. This report will be mainly investigate the following variables: Age, Gender, Religion has affiliation and Household size.
Variable "Age" in both of the datasets indicates each respondent's age up until the year of the election, which is 2019. It was a newly created variable, using 2019 minus the year of born. Variable "Gender" indicates the gender of each respondent, "Male" or "Female"; a response with gender "other" has been removed during the data cleaning process. Variable "Citizenship_status" represents whether the respondent was born in Canada or the citizenship was given by naturalization. "Vote_Liberal" represents whether the respondent did vote for the Liberal party or planning to vote for the Liberal party, response "0" indicates not voting or does not plan to vote for the Liberal party, "1" indicates the respondent voted for Liberal or planning to vote for Liberal. "Household_size" represents the number of family members the respondent has. "Religion_has_affiliation" indicates whether the respondent has religion or not, such as Muslim, Baptist, and Pentecostal, etc.

Numerical summaries

Table 2: Numerical Summaries of age and household size variables

Variables	Mean	Median	Min	Max	Standard deviation	IQR
age	50.9	51	18	100	17.1	27
household_size	2.66	2	1	15	1.43	2

Table 2 is the numerical summaries of the number of age and household size variables. From Table 2, we know that among 2675 observations, the sample mean and sample median of age is 50.9 and 51, respectively. So, It is almost symmetric. The sample standard deviation of age, which is 17.1, expected a moderate fluctuation of the data. The sample range of age is from 18 to 100, which is a really extensive range. The sample IQR of age is 27, which means that the range between the 1st and the 3rd quantile is 27, a relatively large spread.
For the household size, the sample mean is 2.66, which is higher than its sample median of 2. It implies that the household size is right-skewed. The sample standard deviation of the household size is 1.43. It would expect a relatively small fluctuation of the data. The sample range is between 1 and 15, which is a fairly large range for household size. The sample IQR is 2, which means that the range between the 1st and the 3rd quantile is 2, a relatively small spread.
In summary, age has a relatively more considerable fluctuation and spread. That means practically the age of respondents varies a lot.

Table 3: Proportion of important variables in the phone survey

term	Not vote Liberal	Vote Liberal	Has religious affiliation	No religious affiliation	Female	Male
proportion	67%	33%	64%	36%	42%	58%
count	1801	874	1704	971	1122	1553

Table 4: Proportion of important variables in the census

term	Has religious affiliation	No religious affiliation	Female	Male
proportion	80%	20%	55%	45%
count	15109	3664	10277	8496

From Table 3 the proportion of important variables in the phone survey, we know the number of votes not voting Liberal Party is 1801, and the number of voting liberal party is 874 among the total 2675 respondents from the survey. Respondents that do not plan to vote or did not vote for Liberal Party accounted for the most significant part, and it is nearly two times that of the respondents who plan to vote for the Liberal Party. The proportion of those who vote for liberal party is around 33 percent, and the proportion of those who did not vote for liberal party is around 67 percent, which means they plan to vote or already voted for other parties, including Conservatives, NDP, Bloc Québécois, Green Part, and People's Party. Thus, combining the intention of voting, the Liberal party accounts for a significant part among all parties.
From Table 3 and Table 4, in the survey the number of respondents who have religious affiliation is 1704, and respondents without religious affiliation are 971. In the census, the number of respondents with a religious affiliation is 15109, and respondents without religious affiliation are 3664. In both datasets, more voters have a religious affiliation than non-religious affiliation ones. The proportion of voters who have religious affiliation in the phone survey is 64 percent, while it is 80 percent in the census. It indicates a significant difference between the two datasets, with more people having religious affiliation in the census. In contrast,

respondents without religious affiliation accounted for 36 percent in the phone survey and 20 percent in the census.
The proportion of males and females varies considerably in the phone survey and the census. In the phone survey data, 1122 respondents are female, about 42 percent. Moreover, 1553 are male, about 58 percent. While in the census, only 8496 respondents are male, about 45 percent. Moreover, 10,277 respondents are female, about 55 percent. Although the difference between the male and female percentages in both dataset is not significant, the results in the phone survey and the census are opposite, with more male voters than females in the phone survey and fewer than females in the census.
In summary, there are around 33 percent of respondents support the liberal party in the survey. In both datasets, more voters have a religious affiliation than non-religious affiliation ones. The proportion of voters who have religious affiliation in the phone survey is 64 percent, while it is 80 percent in the census. The difference between the male and female percentages in both dataset is not significant,the results in the phone survey and the census are opposite, in the phone survey data, about 42 percent of respondents are female and about 58 percent of respondents are male. While in the census, about 45 percent of respondents are male and about 55 percent of respondents are female.
There will be some graphical summaries analyzing these variables later as well.

Graphical summaries

Figure 1: Histograms of respondents age in the phone survey and census

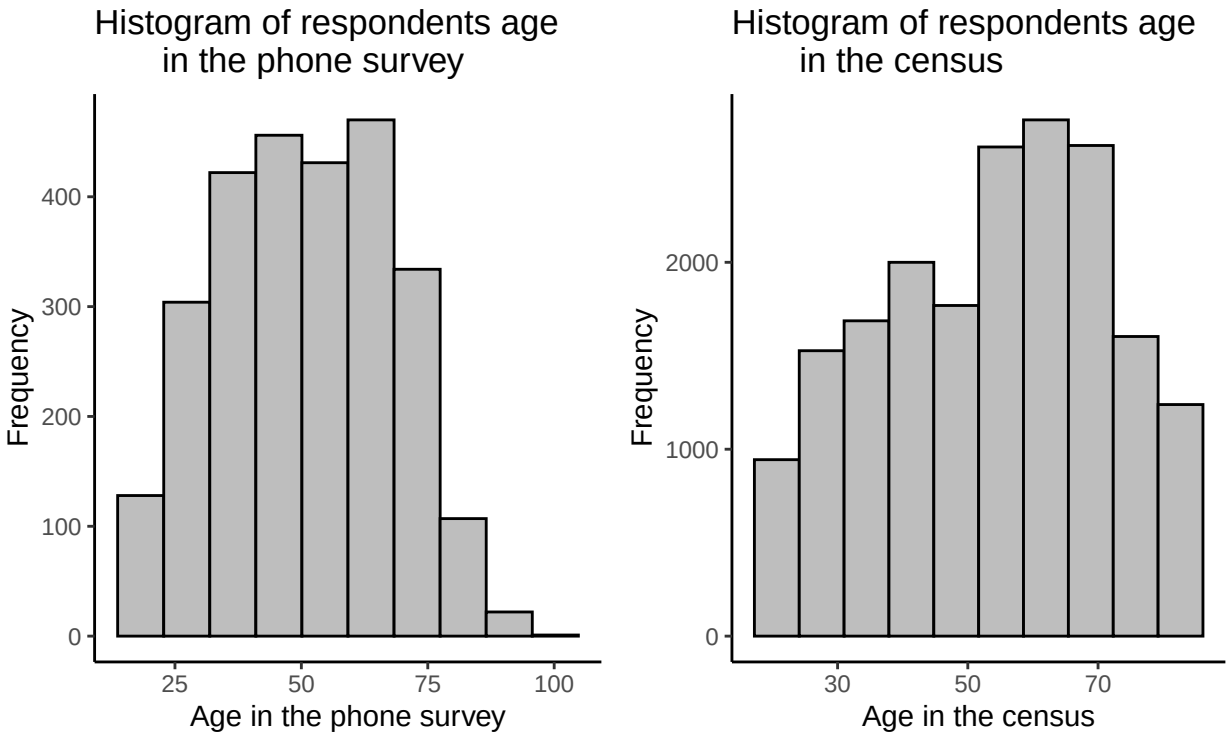

Figure 1 shows the histograms of respondents' age in both phone survey and the census in 2019.
The histogram on the left demonstrates that respondents' age in the phone survey is roughly symmetric, and the other histogram in the census seems a little bit left-skewed. The center of the plot for the phone survey is around 50 years old, which means that mid-age people mostly filled the survey at around 40~60. The histogram on the right demonstrates the age of respondents that were in the census. It centered around the age of 65 years old, which means that mid-age people mostly filled the census at around 50~70.

Combining the two histograms, there are no significant modes. Besides, it is clear that the overall age of respondents in the census is greater than the respondents who took the survey; since respondents at the age of 40~60 account for a significant part of the survey, while census respondents count more at the age of 50~70. The spread of the age for both the survey and the census is from 18 to 100 since the legal age of voting is 18.

Figure 2: Histograms of household size of respondents in the phone survey and census

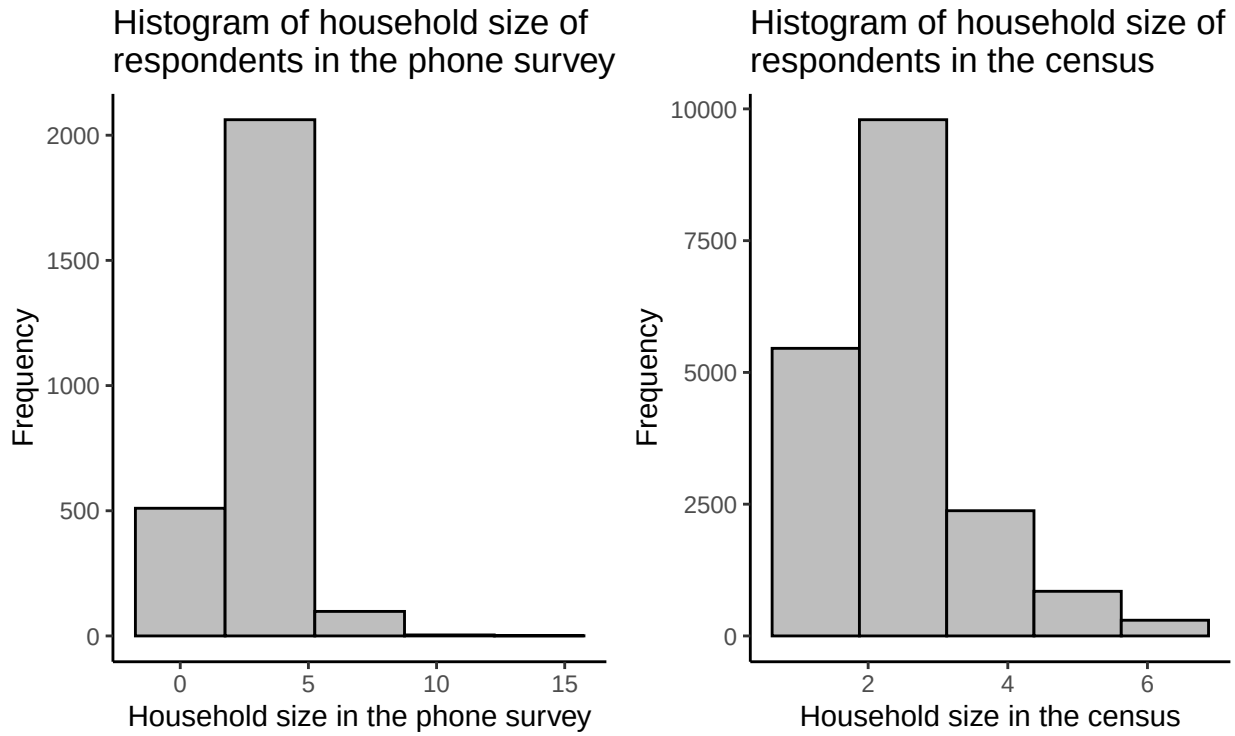

Figure 2 illustrates the histograms of the household size of respondents in the phone survey and the census, and both are right-skewed distributions. The center of the histograms is 4 and 3 respectively in the phone survey and census, which means a household with four and three people showed the most participation in voting in the phone survey and the census data, respectively. From those plots, we found that respondents in the phone survey are in the household size is between 1 and 15, it coincides with the range in the numerical summaries table, where the range is between 1 and 15. Moreover, the range of household size of respondents in the census is from 1 to 6.
In summary, most household size among respondents is between 1 and 5.

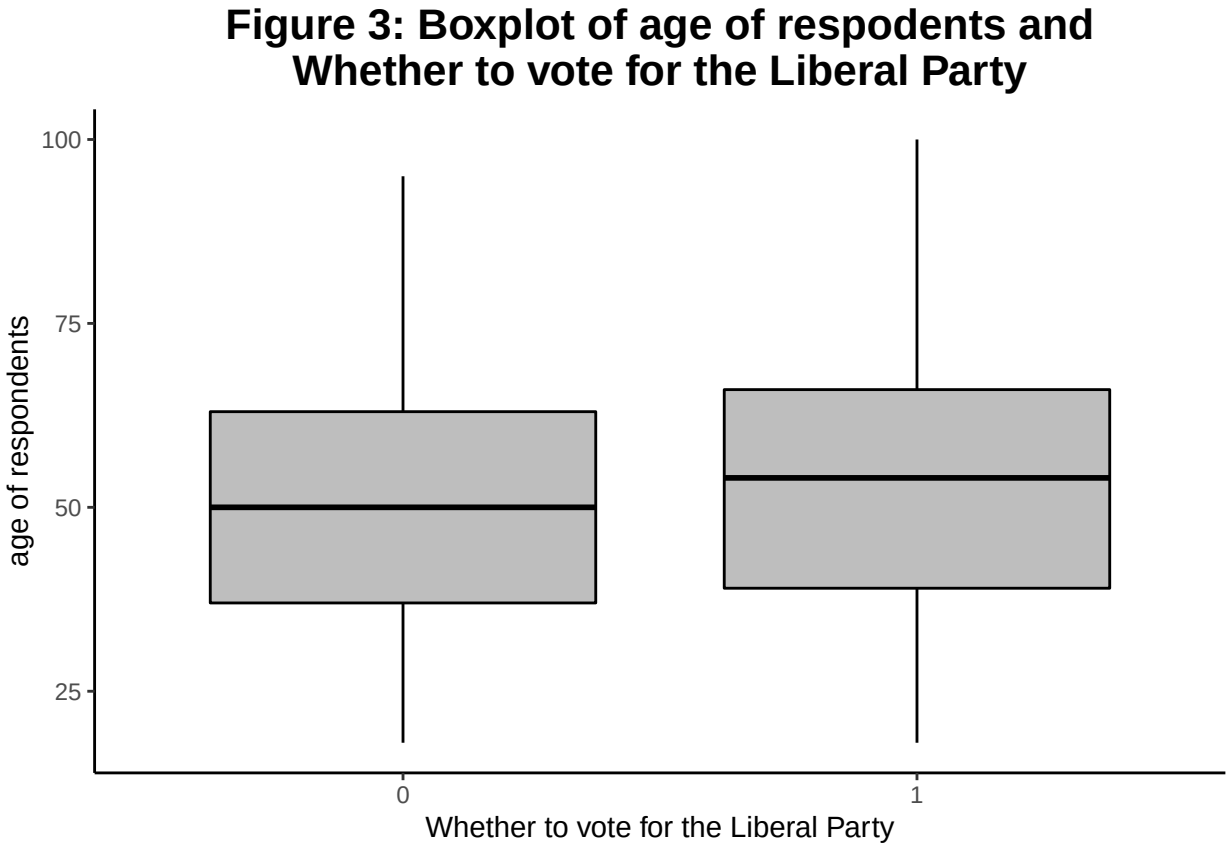

Figure 3 is the boxplot of the age of respondents and whether to vote for the Liberal Party. The two values in the x-axis are represented by 0 for those who did not vote for the Liberal Party and 1 for those who did. Figure 3 shows a side-by-side boxplot of the age of respondents who voted for the Liberal Party and who did not in the phone survey data in 2019. The shape of the plots is slightly right-skewed for both boxplots, and they do not have outliers. The center of the left boxplot is 48, which means the median age of respondents who did not vote for the Liberal Party is around 48 years old. The first quantile is nearly 37 years old, and the third quantile is nearly 63 years old. So the IQR of the age of respondents who did not vote for the Liberal Party is approximately 26 years old. It implies that the spread of the age of respondents who did not vote for the Liberal Party is moderate.

Furthermore, the median age of respondents who vote for the Liberal Party is around 52 years old, slightly higher than the respondents who do not vote for the Liberal Party. Its first quantile is around 38 years old, and the third quantile is around 65 years old. The IQR of the respondents' age in the vote for the Liberal group is approximately 27 years old, which is similar to the other group. It implies that the spread of the age of respondents who vote for the Liberal Party is almost the same as those who did not.

In conclusion, both have the almost same spread and fluctuation, but the right boxplot of the age of respondents who vote for the Liberal Party is a little higher overall with a higher range and center. Whether they voted for the Liberal Party or not, the age of the respondents is around 50 years old, which means that older people are more willing to vote. Nevertheless, the respondents who vote for the Liberal Party have a higher age.

Wrapping up for data section

We want to emphasize several critical results in the data section. From Table 2, age has a relatively more considerable fluctuation and spread. That means practically the age of respondents varies a lot. Around 33 percent of respondents support the liberal party in the survey, shown in Table 3. In Table 3 and 4, more voters have a religious affiliation than a non-religious affiliation in both datasets. The proportion of

voters who have religious affiliation in the phone survey is 64 percent, while 80 percent in the census. The difference between the male and female percentages in both datasets is not significant. The results in the phone survey and the census are opposite, in the phone survey data, about 42 percent of respondents are female, and about 58 percent of respondents are male. While in the census, about 45 percent of respondents are male, and about 55 percent of respondents are female.Based on Figure 1, it is clear that the overall age of respondents in the census is greater than the respondents who took the survey since respondents at the age of 40~60 account for a significant part of the survey, while census respondents count more at the age of 50~70.. Figure 2 displays the most household size among respondents is between 1 and 5. From Figure 3, both boxplots have almost the same spread and fluctuation, but the boxplot of the age of respondents who vote for the Liberal Party is a little higher overall with a higher range and center. Whether they voted for the Liberal Party or not, the age of the respondents is around 50 years old, which means that older people are more willing to vote. Nevertheless, the respondents who vote for the Liberal Party have a higher age.

Methods

To predict the proportion of who will vote for the Liberal party, we will use age(numerical), gender(categorical), religion has affiliation(categorical), and household size(numerical) as the predictors. Since *vote liberal* is a binary response variable. Therefore we can use the logistic regression model. Assume the data in both datasets are independent and accurate.
The method section will use the post-stratification and logistic regression model to predict the estimated probability for voting for the Liberal party. I will describe the model specifics and how to do the post-stratification as follows.

Model Specifics

We will be using a logistic regression model to get the proportion of voters who will vote for the Liberal party. Since the predicted result is a binary response variable, the model result is the log of odds for voting for the Liberal party. The logistic regression model we will be using is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{\text{age}} + \beta_2 x_{\text{gender}} + \beta_3 x_{\text{religion has affiliation}} + \beta_4 x_{\text{household size}}$$

Where:
x_{age} and $x_{\text{household size}}$ are both numerical variables, as we have described in the data section. x_{gender}: when *gender* equal to 1, it will be male. Otherwise, when *gender* is equal to 0, it will be female. $x_{\text{religious has affiliation}}$: when *religion has affiliation* equal to 1, it will be No religious affiliation. Otherwise, when *religion has affiliation* equal to 0, it will be Has religious affiliation.
p represents the probability of voting for the Liberal party.
β_0 represents the intercept of the model and is the log of odds of voting for the Liberal party when *age*, *household size* is 0, and when *gender* equal to 0, that is female, *religion has affiliation* equal to 0, that has a religious affiliation. However, in this case, the intercept makes nonsense as age cannot be equal to 0.
β_1 represents the slope of *age* in the model. We expect a β_1 increase log odds of voting for the liberal party when holding other predictors constant for every one unit increase in age.
β_2 represents the average difference in log odds of voting for the liberal party between males and females when holding other predictors constant.
β_3 represents the average difference in log odds of voting for the liberal party between has religious affiliation and no religious affiliation when holding other predictors constant.
β_4 represents the slope of *household size* in the model. We expect a β_4 increase log odds of voting for the liberal party when holding other predictors constant for every unit increase in household size.

Model selection process

For the model selection process, since the research question is: By using age, gender, religion has affiliation, and household size to predict the proportion of voters who will vote for the Liberal party.
Therefore, I will use the *vote liberal* as the response variable. Then, I will use the *age*, *gender*, *religion has affiliation*, and *household size* as the predictors by the practical rationale stated in the introduction.

Post-Stratification

Assume the data in both datasets are independent and accurate. Moreover, the data in each group is disjoint. In the data section, we have found the inconsistency of the characteristics of the respondents in the census and the survey (age distribution, gender distribution, etc.). We will introduce the post-stratification method, which is the process of adjusting the estimates, essentially a weighted average of estimates from all possible combinations of attributes such as age and gender. Each combination is sometimes called a "cell." (Multilevel regression with poststratification 2021)

In order to estimate the proportion of voters who will vote for the liberal party, I will perform a post-stratification analysis. We created cells based on different age, gender, religion has affiliation, and household size. Using the logistic regression model mentioned above, we will estimate the proportion of respondents who vote for the liberal party for each group (bin). And then weighted each proportion estimate (within each bin) to the respective population size of that bin, added the values together, and divided by the entire population size.
Here is the mathematical expression for the post-stratification method:

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

Where:
N_j: population size of each group. $\widehat{y}_j$: the proportion of respondents vote for liberal party.

Results

I selected the variables in table 5 to construct the logistic regression model. *vote liberal* is the response variable. *age*, *gender*, *religion has affiliation* and *household size* are the predictors. I removed the missing values in each selected variable as well.

Table 5: Description of logistic regression model variables

Variable name	Description	Type
Age	The age of each respondents in 2019	Numerical
Gender	Gender of the respondent	Categorical
Religion has affiliation	Whether the respondent has Religious affiliation	Categorical
Household size	Household size of respondent	Numerical
Vote Liberal	Whether the respondent vote for Liberal	Categorical

Table 5 is the variables description table for the variables used in the logistic regression model. A more detailed description can be found in the data variable description section.

logistic regression model

Table 6: estimated coefficients of logistic regression

term	estimate	standard error	statistic	p-value
Intercept	-0.971415	0.2022702	-4.8025614	1.5664876×10^{-6}
age	0.0084053	0.002673	3.1445407	0.0016635
gender Male	-0.2276623	0.0836903	-2.7202944	0.0065224
religion has affiliation No religious affiliation	-0.0488454	0.0896104	-0.5450864	0.5856941
household size	-0.0137135	0.0313017	-0.4381072	0.6613086

Table 6 is a summary table for the coefficient of the logistic regression model. From table 6, eventually the estimated logistic regression model is:

$$\log(\frac{\hat{p}}{1-\hat{p}}) = -0.971 + 0.008x_{\text{age}} - 0.228x_{\text{gender}} - 0.049x_{\text{religion has affiliation}} - 0.014x_{\text{household size}}$$

where p represents the estimated probability of voting for the Liberal party.
For $\hat{\beta}_1$=0.0084053, when age increase by one year, the average log odds of voting for the liberal party will increase by 0.0084053 when holding other predictors constant.
For $\hat{\beta}_4$=-0.0137135, when household size increase by 1 unit, the average log odds of voting for the liberal party will increase by -0.0137135 when holding other predictors constant.
As $\hat{\beta}_0$=-0.971415, that is the intercept of the model. It means when *age*, *household size* is 0, and when *gender* equal to 0, that is female, *religion has affiliation* equal to 0, that has religious affiliation. The expected log odds of voting for the liberal party is -0.971415. However, for example, we will not have the age is 0. Thus, the intercept of the model seems does not seem to make sense in real life.
For $\hat{\beta}_2$=-0.2276623 represents the average difference in log odds of voting for the liberal party between Male and Female when holding other predictors constant is -0.2276623.
For $\hat{\beta}_3$=-0.0488454 represents the average difference in log odds of voting for the liberal party between has religious affiliation and no religious affiliation when holding other predictors constant is -0.0488454.
The p-value for $H_0 : \beta_i = 0$, $H_a : \beta_i \neq 0$. Where β_i is coefficients of the logistic regression model. As the common significant level is 0.05. It means that when the p-value is lower than 0.05, we have strong evidence against that $H_0 : \beta_i = 0$. From Table 6, we found that the p-value of age is 0.0016635, which is lower than 0.05. We have strong evidence to support that $H_a : \beta_1 \neq 0$. The p-value of gender is 0.0065224, which is lower than 0.05. We have strong evidence to support that $H_a : \beta_2 \neq 0$. From Table 6, we found that the p-value of religion has affiliation is 0.5856941 and the p-value of household size is 0.6613086. Both of them the p-value are larger than 0.05, which are not very significant predictors. Since the research question of the report is to predict the voting for the liberal party, I do not need to remove them from the model.
Altogether, we know that the estimated model when gender is male and religion has affiliation is No religious affiliation would be:

$$\hat{Y}_i = -1.248 + 0.008x_{age} - 0.014x_{\text{household size}}$$

where $\hat{Y}_i$ is the expected log odds of voting for the liberal party.
Then, we know that the estimated model when gender is female, and religion has affiliation is Has religious affiliation would be:

$$\hat{Y}_i = -0.971 + 0.008x_{age} - 0.014x_{\text{household size}}$$

where $\hat{Y}_i$ is the expected log odds of voting for the liberal party.
Practically, when we find out that when age increase by one year, the average log odds of voting for the liberal party will increase by 0.0084053 for both gender and religion has affiliation when holding household size constant.
We find out that when household size increase by 1 unit, the average log odds of voting for the liberal party will increase by -0.0137135 for both gender and religion has affiliation when holding age constant.

Post-Stratification

Table 7: Group90-94 voters from census

age	gender	religion_has_affiliation	household_size	N	N_prop
22	Female	Has religious affiliation	1	8	0.0004261
22	Female	Has religious affiliation	2	9	0.0004794
22	Female	Has religious affiliation	3	14	0.0007458
22	Female	Has religious affiliation	4	14	0.0007458
22	Female	Has religious affiliation	5	7	0.0003729

Table 8: Group54-58 voters from survey

age	gender	religion_has_affiliation	household_size	n	n_prop
22	Female	Has religious affiliation	1	1	0.0003738
22	Female	Has religious affiliation	2	1	0.0003738
22	Female	Has religious affiliation	3	1	0.0003738
22	Female	Has religious affiliation	4	1	0.0003738
22	Female	Has religious affiliation	5	1	0.0003738

Table 7 and Table 8 both have five groups of voters from the census and survey. As we can see, these groups have the same age, gender, religion has affiliation, and household size. However, their proportion in the census and survey is different.
For example, the respondents who are 22 years old female has religious affiliation with one household size. There are 8 of them in the census, and their proportion in the census is 0.04261%. By contrast, we only have one respondent with the same attributes in the survey, and their proportion in the survey is 0.03738%. Clearly, the proportion of the respondents with the same attributes are different in census and survey. Therefore, the post-stratification is useful to adjusting the estimates, essentially a weighted average of estimates from all possible combinations of attributes.

Table 9: First 15 groups of voters from census

age	gender	religion_has_affiliation	household_size	n	estimate
18	Female	Has religious affiliation	1	2	0.3028351
18	Female	Has religious affiliation	2	3	0.2999477
18	Female	Has religious affiliation	3	7	0.2970761
18	Female	Has religious affiliation	4	14	0.2942204
18	Female	Has religious affiliation	5	8	0.2913808
18	Female	Has religious affiliation	6	4	0.2885574
18	Female	No religious affiliation	1	1	0.2926229
18	Female	No religious affiliation	3	5	0.2869782
18	Female	No religious affiliation	4	11	0.2841803
18	Female	No religious affiliation	5	3	0.2813990
18	Female	No religious affiliation	6	3	0.2786342
18	Male	Has religious affiliation	2	2	0.2544137
18	Male	Has religious affiliation	3	9	0.2518212
18	Male	Has religious affiliation	4	17	0.2492463
18	Male	Has religious affiliation	5	13	0.2466890

Table 9 shows the first 15 groups of voters from the census. Where n is N_j, that is the population size of each group. Moreover, the estimate is $\widehat{y}_j$ is the proportion of respondents who vote for the liberal party. The estimate is calculated by the logistic regression model mentioned previously.
For example, we found that the respondent, 18 years old female, has a religious affiliation with one household size. Two of them have this attribute in the population, and their proportion of vote for the liberal party is around 30.28%.
Eventually, we will estimate the proportion of respondents who vote for the liberal party for each group by the logistic regression model mentioned previously. And then weighted each proportion estimate within each group to the respective population size of that group, added the values together, and divided by the entire population size. After that, we will get the weighted proportion of respondents who vote for the liberal party by population size.

Table 10: Total Groups

n
1234

Table 10 shows the total groups in the census. It suggests that we have a total of 1234 groups in the census. These groups are the cells based on different age, gender, religion has affiliation and household size.

Table 11: Prediction of weighted proportion of respondents vote for liberal party by population size

Predict
0.3409564

Table 11 is the result of predicting the weighted proportion of respondents who vote for the liberal party by population size. It is calculated by using the logistic regression model mentioned previously. We estimated the proportion of respondents who vote for the liberal party for each group (bin). And then weighted each proportion estimate (within each bin) to the respective population size of that bin, added the values together, and divided by the entire population size.
The result suggests that there will be around 34.1% of respondents will vote for the liberal party. It is smaller than 40%, which we hypothesized in the introduction part. Thus, the result does not capture our hypothesis, though it is close. However, it is still a large proportion of votes.

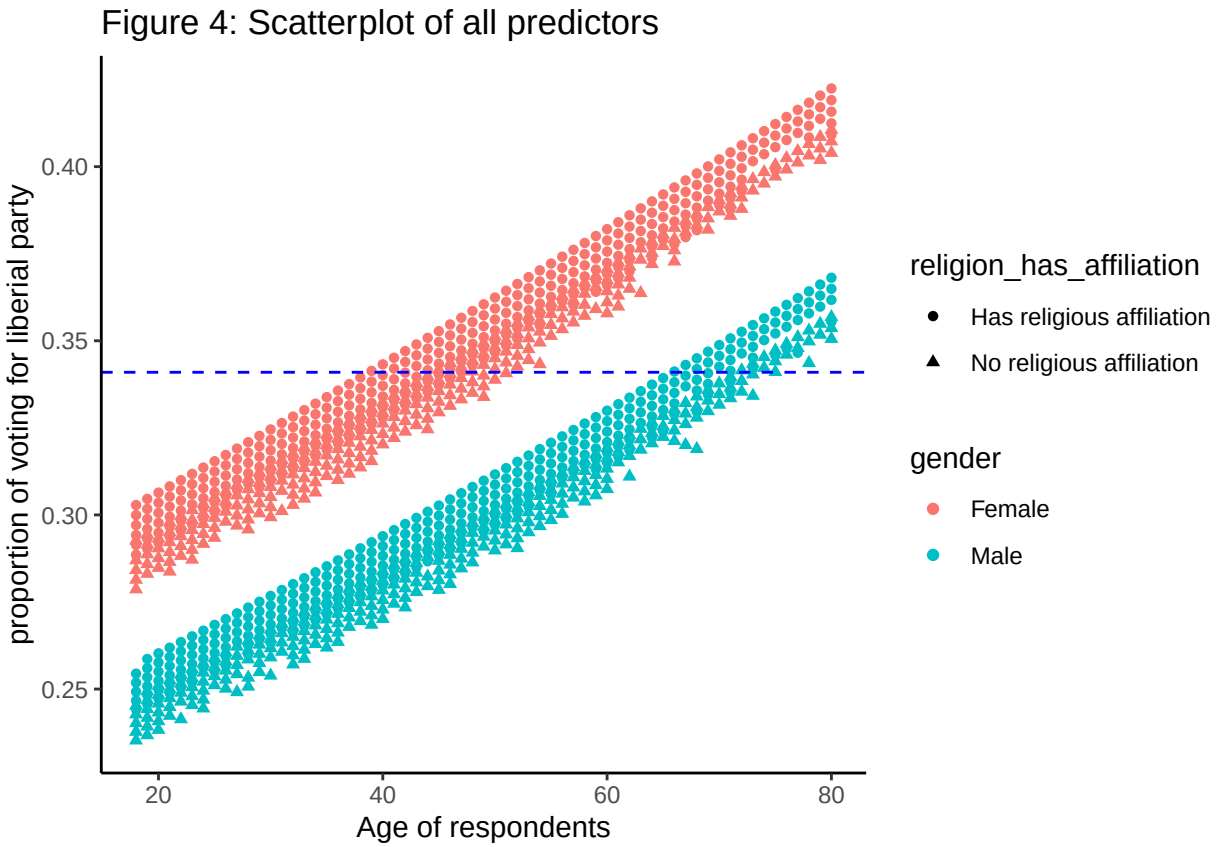

Figure 4 is the scatterplot of all predictors. The blue dashed line is the result of the prediction of the

weighted proportion of respondents voting for the liberal party by population size.
Figure 8 suggests a strong positive linear relationship between the age of respondents and the proportion of respondents who vote for the liberal party. As the age of respondents increases, the proportion of voting for the liberal party increases. Moreover, females are more likely to vote for the liberal party than males. Furthermore, the respondents who have religious affiliation are more likely to vote for the liberal party.

Wrapping up for result section

In the result section, I used the practical rationale to determine the estimated logistic regression model for this report. That is $\log(\frac{\hat{p}}{1-\hat{p}}) = -0.971 + 0.008x_{\text{age}} - 0.228x_{\text{gender}} - 0.049x_{\text{religion has affiliation}} - 0.014x_{\text{household size}}$, where $\log(\frac{\hat{p}}{1-\hat{p}})$ is the expected log odds of voting for liberal party. Practically, when we find out that when age increase by one year, the average log odds of voting for the liberal party will increase by 0.0084053 for both gender and religion has affiliation when holding household size constant.
We find out that when household size increase by 1 unit, the average log odds of voting for the liberal party will increase by -0.0137135 for both gender and religion has affiliation when holding age constant.
Moreover, there will be around 34.1% of respondents will vote for the liberal party. It is smaller than 40%, which we hypothesized in the introduction part. Thus, the result does not capture our hypothesis, though it is close. However, it is still a large proportion of votes.
In addition, there is a strong positive linear relationship between the age of respondents and the proportion of respondents who vote for the liberal party. As the age of respondents increases, the proportion of voting for the liberal party increases. Moreover, females are more likely to vote for the liberal party than males. Furthermore, the respondents who have religious affiliations are more likely to vote for the liberal party.

Conclusions

As specified in the introduction, my research question is: By using age, gender, religion has affiliation, and household size to predict the proportion of voters who will vote for the Liberal party in 2019. We hypothesize that the Liberal Party will have around 40% of the votes in 2019.
We have assumed that the data in both datasets are independent and accurate. Moreover, the data in each group is disjoint. We have a logistic regression model between the log of odds of voting Liberal Party and four predictors: age, gender, religion has affiliation, and household size. Taken together, we used the practical rationale to obtain the estimated logistic regression model for the report. That is $\log(\frac{\hat{p}}{1-\hat{p}}) = -0.971 + 0.008x_{\text{age}} - 0.228x_{\text{gender}} - 0.049x_{\text{religion has affiliation}} - 0.014x_{\text{household size}}$ Practically, when we find out that when age increase by 1 year, the average log odds of voting for Liberal Party will increase by 0.0084053 for both gender and religion has affiliation when holding household size constant. And when household size increase by 1 unit, the average log odds of voting for Liberal Party will increase by -0.0137135 for both gender and religion has affiliation when holding age constant.
We predicted the weighted proportion of respondents who vote for the liberal party by population size. It is calculated by using the logistic regression model mentioned previously. We estimated the proportion of respondents who vote for the liberal party for each group (bin). And then weighted each proportion estimate (within each bin) to the respective population size of that bin, added the values together, and divided by the entire population size.
As a result, there will be around 34.1% of respondents will vote for the liberal party. It is smaller than 40%, which we hypothesized in the introduction part. Thus, the result does not capture our hypothesis, though it is close. However, it is still a large proportion of votes.
In addition, there is a strong positive linear relationship between the age of respondents and the proportion of respondents who vote for the liberal party. As the age of respondents increases, the proportion of voting for the liberal party increases. Moreover, females are more likely to vote for the liberal party than males. Furthermore, the respondents who have religious affiliations are more likely to vote for the liberal party.
Besides, I have a detailed data collection and cleaning process with variable descriptions in the data section. Also, I want to emphasize several critical results in the data section. From Table 2, age has a relatively more

considerable fluctuation and spread. That means practically the age of respondents varies a lot. Around 33 percent of respondents support the liberal party in the survey, which is shown in Table 3. In Table 3 and 4, more voters have a religious affiliation than non-religious affiliation ones in both datasets. The proportion of voters who have religious affiliation in the phone survey is 64 percent, while 80 percent in the census. The difference between the male and female percentages in both datasets is not significant. The results in the phone survey and the census are opposite, in the phone survey data, about 42 percent of respondents are female, and about 58 percent of respondents are male. While in the census, about 45 percent of respondents are male, and about 55 percent of respondents are female.
Based on Figure 1, it is clear that the overall age of respondents in the census is greater than the respondents who took the survey since respondents at the age of 40~60 account for a significant part of the survey, while census respondents count more at the age of 50~70. The spread of the age for both the survey and the census is from 18 to 100 since the legal age of voting is 18. Figure 2 displays the most household size among respondents is between 1 and 5. From Figure 3, both boxplots have almost the same spread and fluctuation, but the right boxplot of the age of respondents who vote for the Liberal Party is a little higher overall with a higher range and center. Whether they voted for the Liberal Party or not, the age of the respondents is around 50 years old, which means that older people are more willing to vote. Nevertheless, the respondents who vote for the Liberal Party have a higher age.

Weaknesses

The p-value of the household size and religion has affiliation are larger than 0.05, which is not significant. We do not reject the null hypothesis of coefficients of the two variables. However, it is not removed because we are only looking for prediction the proportion of voting Liberal Party, not for the relationship between the response and predictors. So it looks like it is not particularly perfect for this model, which is the weakness of this model.

Next Steps

It could be interesting to research the confounding variables that are not chosen in this report from the original dataset. Because we only use the four variables that we think have the most influence on the prediction, other variables in the dataset influence this prediction of the voting Liberal Party. However, we cannot discuss all variables here since we do not have enough data and literature about other variables.
Also, if future researchers have data about future elections, they can use this model to predict the results of voter voting for the Liberal Party in advance.

Discussion

Throughout the report, my research question is by using age, gender, religion has affiliation, and household size to predict the proportion of voters who will vote for the Liberal party in 2019. Firstly, we investigated the numerical and graphical summaries for age, household size, gender, religion has affiliation, and vote liberal in the data section. Then, we came up with the method I used to predict the proportion of voting Liberal Party, and we found that 34.1% of respondents will vote for Liberal Party by using post-stratification. The methodologies we used were introduced in the method section, and all corresponding results are in the result section.

Bibliography

All analysis for this report was programmed using `R version 4.1.2`.
1. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)
2. *A guide to voting in the Canadian federal election.* Settlement.Org | Information Newcomers Can Trust. (2021, September 13). Retrieved November 4, 2021, from https://settlement.org/ontario/immigration-citizenship/canadian-government/voting/a-guide-to-voting-in-the-canadian-federal-election/.
3. Al Parsai. (2021, August 22). *Naturalized citizen versus born citizen in Canada.* Parsai Immigration Services. Retrieved November 5, 2021, from https://www.settler.ca/english/naturalized-citizen-versus-born-citizen/.
4. Assumptions of logistic regression. Statistics Solutions. (2021, August 11). Retrieved November 6, 2021, from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/.
5. Belmonte, L. (2019, September 28). *Voting in Canada differs a lot between men & women.* Narcity. Retrieved November 5, 2021, from https://www.narcity.com/voting-in-canada-differs-a-lot-between-men-and-women.
6. Canada, E. (n.d.). Home. – *Elections Canada.* Retrieved November 4, 2021, from https://www.elections.ca/content.aspx?section=abo&dir=role&document=index&lang=e.
7. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.
8. *Factors associated with voting.* Statistics Canada: Canada's national statistical agency / Statistique Canada : Organisme statistique national du Canada. (2015, November 27). Retrieved November 5, 2021, from https://www150.statcan.gc.ca/n1/pub/75-001-x/2012001/article/11629-eng.htm.
9. Government of Canada, S. C. (2021, October 28). *Frequently asked questions-general information.* Government of Canada, Statistics Canada. Retrieved November 5, 2021, from https://census.gc.ca/faq/general-eng.htm.
10. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)
11. *How religious beliefs affect voting behavior in Canada.* Arts. (2019, October 15). Retrieved November 5, 2021, from https://uwaterloo.ca/arts/news/how-religious-beliefs-affect-voting-behavior-canada.
12. Stephenson, Laura B.,Allison Harell,Daniel Rubenson and Peter John Loewen. *The 2019 Canadian Election Study -Phone Survey.* [dataset].
13. *The constitution of the Liberal Party of Canada.* (2021, April 11). Retrieved November 5, 2021, from https://liberal.ca/wp-content/uploads/sites/292/2021/04/The-Constitution-of-the-Liberal-Party-of-Canada.pdf.
14. What is logistic regression? Statistics Solutions. (2021, August 11). Retrieved November 6, 2021, from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-logistic-regression/.
15. Wikimedia Foundation. (2021, September 24). *Multilevel regression with poststratification.* Wikipedia. Retrieved November 4, 2021, from https://en.wikipedia.org/wiki/Multilevel_regression_with_poststratification.
16. Wikimedia Foundation. (2021, November 2). *2019 Canadian federal election.* Wikipedia. Retrieved November 4, 2021, from https://en.wikipedia.org/wiki/2019_Canadian_federal_election#Liberal.
17. Wikimedia Foundation. (2021, October 27). *Liberal Party of Canada.* Wikipedia. Retrieved November 5, 2021, from https://en.wikipedia.org/wiki/Liberal_Party_of_Canada.
18. Wikimedia Foundation. (2021, September 22). *Results of the 2015 Canadian federal election.* Wikipedia. Retrieved November 5, 2021, from https://en.wikipedia.org/wiki/Results_of_the_2015_Canadian_federal_election.

Appendix

The 2019 Canadian Election Study -Phone Survey questions we used in this report:
Question 1: Are you a Canadian Citizen?
Answers:
1-Yes
2-No
-8-Refused
-9-Don't know

Question 2: In what year were you born?
Answers: _____
-8-Refused
-9-Don't know

Question 3: With which gender do you identify?
Answers:
1-Male
2-Female
3-Other
-8-Refused
-9-Don't know

Question 11: (if Certain or Likely to Vote) Which party do you think you will vote for?
(if Already Voted) Which party did you vote for?
(if Unlikely to vote) If you decide to vote, which party do you think you will vote for?
Answers:
1-Liberal (Grits)
2-Conservatives (Tory, PCs, Conservative Party of Canada)
3-NDP (New Democratic Party, New Democrats, NDPers)
4-Bloc Québécois (BQ, PQ, Bloc, Parti Québécois) (Show if Quebec)
5-Green Party (Greens)
6-People's Party
7-Other (specify): _____
8-None of these
9-Will not vote
10-Will spoil ballet
-8-Refused
-9-Don't know/undecided

Question 62: Please tell me what is your religion, if you have one?
Answers:
1-Anglican / Church of England
2-Muslim / Islam
3-Baptist
4-Pentecostal / Fundamentalist / Born Again / Evangelical
5-Buddhist / Buddhism
6-Catholic / Roman Catholic / RC
7-Presbyterian
8-Greek Orthodox / Ukrainian Orthodox / Russian Orthodox / Eastern Orthodox
9-Protestant (only after probe)
10-Sikh / Sikhism

11-Hindu
12-United Church of Canada
13-Jehovah's Witness
14-Christian (only after probe)
15-Jewish / Judaism / Jewish Orthodox
16-Christian Reformed
17-Lutheran
18-Salvation Army
19-Mormon / Church of Jesus Christ of Latter Day Saints
20-Mennonite
21-None, don't have one / Atheist
22-Other (specify): _____
-8-Refused
-9-Don't know / Agnostic

Question 71: How many people live in your household?
Answers: _____
-8-Refused
-9-Don't know

Source: Stephenson, Laura B.,Allison Harell,Daniel Rubenson and Peter John Loewen. *The 2019 Canadian Election Study -Phone Survey.* [dataset].