Predictions for People Voting for the Conservatives in the Next Election Report

STA304 - Assignment 3

GROUP 58: Yukun Zhang, Zehao Li, Suheng Yao, Weiyu Zhao

November 5, 2021

Introduction

The federal election is essential to a democratic country. It is a group decision-making procedure with an occurrence at regular intervals. Fair and free elections will bring about a dynamic, progressive, and mature society. Through the election, the public opinion will be discussed, and individuals can choose their representative with the same ideas as themselves to be the governor in the next few years, which effectively improves the existing policies and systems. Every adult citizen of the country has the right to vote and becomes a stakeholder in the country's social and economic development. In this study, we will predict the overall popular vote of the Conservative Party in the next Canadian federal election. They declare its principle is Canada's democratic process, favouring lower taxes, small government, strong national defence, law and order, equal treatment for all Canadians, and upholding individual rights and freedom. Our prediction will help them estimate their support rate and improve and change their strategies favouring more voters. We will use data sets collected by the Canadian government to predict: the census data of General Social Survey 2019, an extensive data set with a population of 25000 composed by using Random Digit Dialling and computer-assisted telephone interviewing. (Caregiving, families, time use, social identity, volunteering and victimization are included) and the survey data of Canadian Election Survey 2019 with complete responses of 4021 observations collected by an online survey tool. The other studies about federal elections have proven the relations between church attendance with vote preference and the political opinions with their religious beliefs. (Plotkin, 2014) And the gap between people of more and less education is widening in terms of party identification. (Suls, 2016) Gender also has an impact on political choice. More men voted for Conservative among all ages and social classes because Men tend to earn more, and the low tax policies of the Conservative party benefit them more. Similarly, people's concerns are different in various ages, and older people are more likely to vote for traditional policies of Conservative. (BBC Bitesize) Based on the research above, to predict the election result, our hypothesis is that the different ages, genders, education levels, and religions of people will have significant impacts on whether they will choose the Conservative party or not. The Conservative Party is one of the two most popular parties in Canada, so we estimate that about 40% of people in the population would vote for it. We will use logistic regression with those impacts and following analysis to predict the proportion of voters who will vote for the Conservative party and to prove our estimation.

Data

In this assignment, I used two data sets. One is the census data, which is called the General Social Survey. This large data set is used as evidence behind government programs to improve the well-being of Canadians, informed research about social life [1].It is consisted of 5 themes, including care giving, families, time use, social identity, volunteering and victimization [1]. The population of the data set is about 25000, and the whole survey process will last 6 to 12 months [1]. It was collected using Random Digit Dialing (RDD), which

is a program randomly generates phone numbers based on in-use area codes [1]. Aside from this method, computer assisted telephone interviewing is used to lower the labor cost [1].
Another data set that I used is the survey data, which is called 2019 Canadian Election Survey. It contains two surveys: one is the Campaign Period Survey, and another one is the Post-Election Survey. In this assignment, I mainly used the Campaign Period Survey because it can more accurately reflects people's choices during the election. It is a online survey generated by Qualtrics, which is a online survey tool [2]. The initial sample space of the survey is 37822 members of the Canadian population, however, after removing incomplete responses, only 4021 observations remaining [2]. In conclusion, these are the two data sets that I use to predict the election result.

The first step is to do some additional cleaning on the census data. Firstly, I imported the initially cleaned census data using the read_csv() function. Since I don't need to use all the variables, I used the select() function to only keep the variables of my interests. These variables are: citizenship_status, age, sex, education and religion importance. I chose variables citizenship_status and age because they could help me filter the data in the next cleaning step. For variables sex and education, sex or gender can largely affect one's decision on voting. According to research, women tend to be more ideologically left-leaning and progressive than men, so women outside of Quebec are more likely than men to support left-wing parties like the New Democratic Party (NDP) and less likely to vote for right-wing parties like the Conservative Party [3]. When it comes to education, more formal education results in a higher probability of voting. People who have higher education are regarded as having received more information about politics than those, who did not receive higher education [4]. For the variable regarding religion, I chose it just because of my own experience. Sometimes, religion can change a person's values and how they see things around him/her. Maybe the party leader believe in the same religion as you, and he/she keeps pushing the development of the religion forward, so you may want to vote for him/her. These are the reasons why I chose these variables in census data. After selecting these variables, I noticed that the age column of the census data is in character type, and it will be easier to analyze if I change it into numerical type. So I used the as.numeric() function to change the variable into numeric type. Also, in this data set, we only want to analyze people who can vote only. So I use the filter() function to keep the observations with age greater than or equal to 18 and citizenship_status equal to "By birth" or "By naturalization". The final step is to remove the NA values because observations containing NA values are incomplete, and they are useless in the analysis. Thus, I used the function na.omit() to remove all these NA values. Above all, these are the cleaning process for census data.

Next, I will introduce how I clean the survey data. First, I used the cesR package to import the survey data into R using the get_ces() function. After the import, since the data would be easier to analyze using table format, so I used the as_tibble() function to make the data set become a table. Then, I selected the columns corresponding to questions relating to the important variables in the census data, which were q1, q2, age, q3, q10, q11, q61, q63 and q64. However, after I kept these variables, it was not very clear what each column represents because they were all question numbers. So I use the rename() function to rename "q1" column to citizenship, "q2" column to year_born, "q3" column to gender, "q10" column to whether_to_vote, "q11" column to party_vote, "q61" column to education, "q63" column to religion_value and "q64" column to country_born. After changing the column names, I also noticed that most of the values in the table cell were numbers, which makes the values hard to understand. Therefore, I used the case_when() function to decode these numbers to show what each number really means. Here, I will only introduce the important variables. Aside the NA value in each column, Citizenship variable contains two unique values: yes and no; Gender variable contains three unique values: male, female and other; whether_to_vote contains 5 values: Certain, Likely, Unlikely, Certain not to vote, Already voted in advanced poll, which describes whether a person really goes to vote; party_vote variable contains 6 values, which are all parties in Canada; education variable contains 11 values, which are all different degrees a person can get during his/her student life, from elementary school to post-secondary education; religion_value contains 5 values: Very Important, Somewhat important, Not very important, Not important at all and Don't know, which are all the importance the religion played in one's daily life. Also, in the survey data, the age column is in character type, so I changed it into numerical type. Same to the census data, NA values have no use to me, so I used the filter() function to remove the NA values. Finally, I only needed to analyze the people who are eligible to vote,i.e. this person has to be a Canadian citizen and over 18 years old, besides, this person has to be certain that he/she will go to vote. So I used the filter() function again to filter these observations out. Above all, these are the cleaning steps for the survey data.

In this part, in order to make it easier to apply post-stratification method, the important variables in the census data and survey data have to be consistent. For the first step, I created the response variable in the survey data, which was the variable vote_Conservative. The function I used is mutate() and ifelse() statement. All the models created in the assignment later will be used to predict this variable. The important variables are age, gender, education_level and religion_importance. Detailed introduction will be written later. For age variable, since both data sets have it, there is no actions needed. For the gender variable, since in the survey data, there is an additional value "Other", because there is a small amount of observations contain this value($<$ 1% of the total data), I just simply eliminate these observations in the survey data to match the census data because these data will not make large differences to the model predictions. For the census data part, I just simply created a new column called gender, using the data from "sex" column. For the variable education_level, although two data sets all have "education" variable, but their values do not match. Census data has 7 unique values, but survey data has 10 unique values. To make it consistent, I again used the mutate() function and case_when statement to map the "Completed secondary / high school" in survey data to "High school diploma or a high school equivalency certificate"; the "Bachelor's degree" in survey data to "Bachelor's degree (e.g. B.A., B.Sc., LL.B.)" in census data; the "Some technical, community college, CEGEP, College Classique" and "Completed technical, community college, CEGEP, College Classique" in survey data to "College, CEGEP or other non-university certificate or di…" in census data; the "Some secondary / high school", "Some elementary school" and "Complete elementary school" in survey data to "Less than high school diploma or its equivalent" in census data; the "Some university" in survey data to "University certificate or diploma below the bachelor's level" in census data; the "Professional degree or doctorate" and "Master's degree" in survey data to "University certificate, diploma or degree above the bach…" in census data. After making these matches, I still found that there was not any values in survey data corresponding to the "Trade certificate or diploma" in census data. Thus, to solve this problem, I first use the "education" column to create "education_level" column in census data, then I used the which() statement to change all the "Trade certificate or diploma" value to "Less than high school diploma or its equivalent". For the variable religion_importance, the only difference is the value "Not important at all" and "Very Important" in the survey data, so I changed them to "Not at all important" and "Very important" using mutate() function and case_when() statement. For census data, I just simply used the regilion_importance column to create religion_importance column. In conclusion, above are the steps to make the survey data be consistent with census data.

Table 1: Important Variable Descriptions

Variables	Variable type	Description
vote_Conservatives	numeric	Whether this person vote for the Conservatives
age	numeric	The person's age
gender	character	The person's gender
education_level	character	The highest education the person receive
religion_importance	character	whether religion plays an important role in this person's life

Table 2: Some numerical summaries of variable age

min	Q1	median	Q3	max	IQR	mean	SD	Lower_Outliers	Higher_Outliers
18	38	51	64	100	26	50.89033	16.83581	0	0

Some graphs introducing the important variables

Table 3: How many people vote for the Conservatives

Whether the person vote for the Conservatives	Number of People
No	852
Yes	664

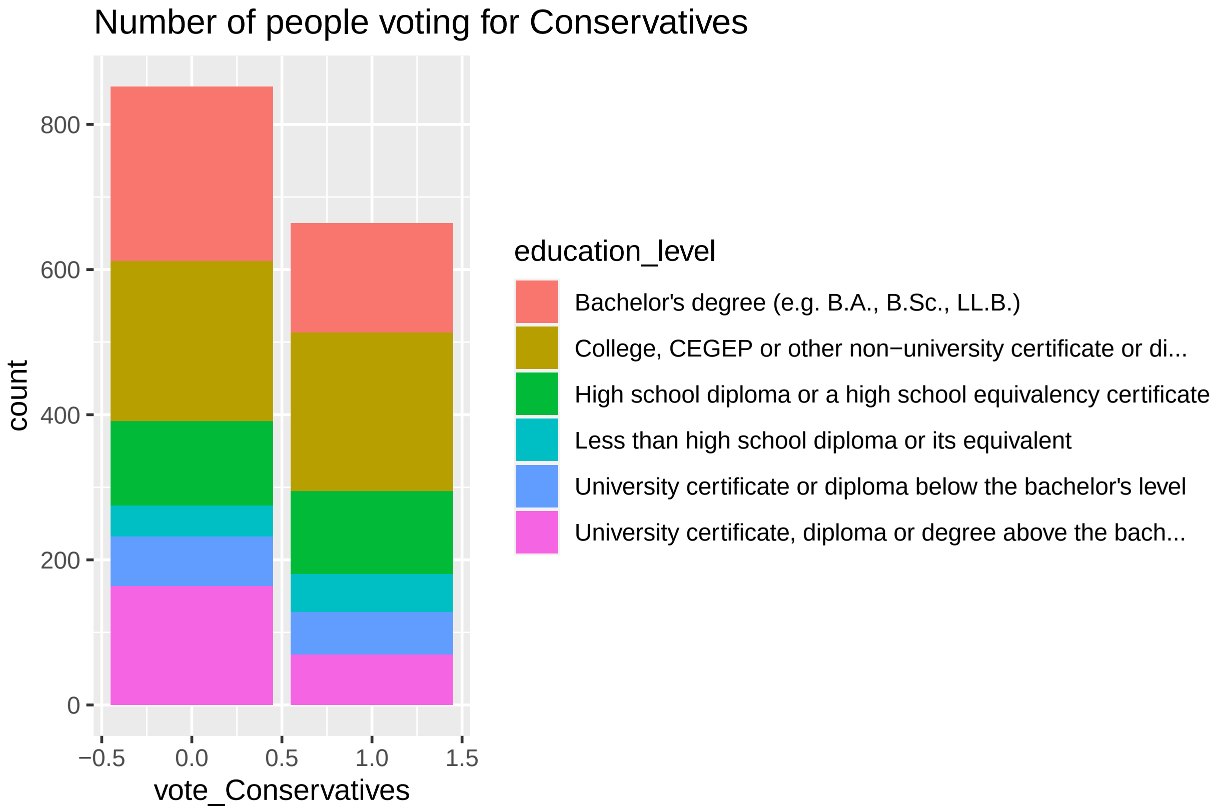

Figure 1

This is a barplot description of the variable vote_Conservatives in relation to the variable education_level in the survey data. It is clear in this graph that there were more people not voting for the Conservatives than people voting for the Conservatives in 2019. In the people who did not vote for the Conservatives, people who get bachelor's degree is the most, and number of people who goes to college is slight lower. However, in people who vote for the conservatives, the number of people who goes to college is clearly higher than that of people who get bachelor's degree. Other education levels have similar behaviors between two groups except people with degree high than bachelor. The number of people who have a degree higher than bachelor but did not vote for the Conservatives is clearly higher than the number of people who have a degree higher than bachelor but vote for the Conservatives.

Table 4: Religion Importance in survey

Religion Importance	Number of People
Don't know	4
Not at all important	141
Not very important	282
Somewhat important	580
Very important	509

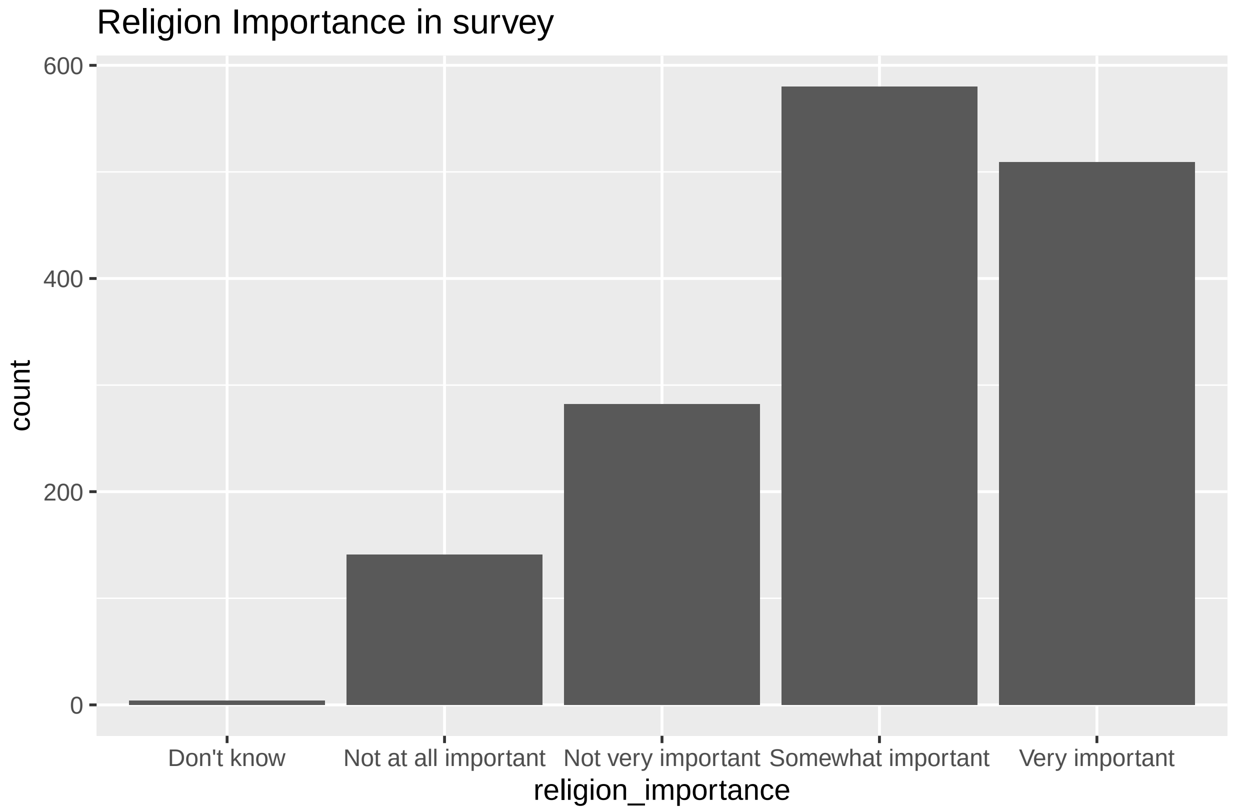

Figure 2

Table 5: Religion Importance in census

Religion Importance	Number of People
Don’t know	201
Not at all important	3632
Not very important	2866
Somewhat important	5630
Very important	6285

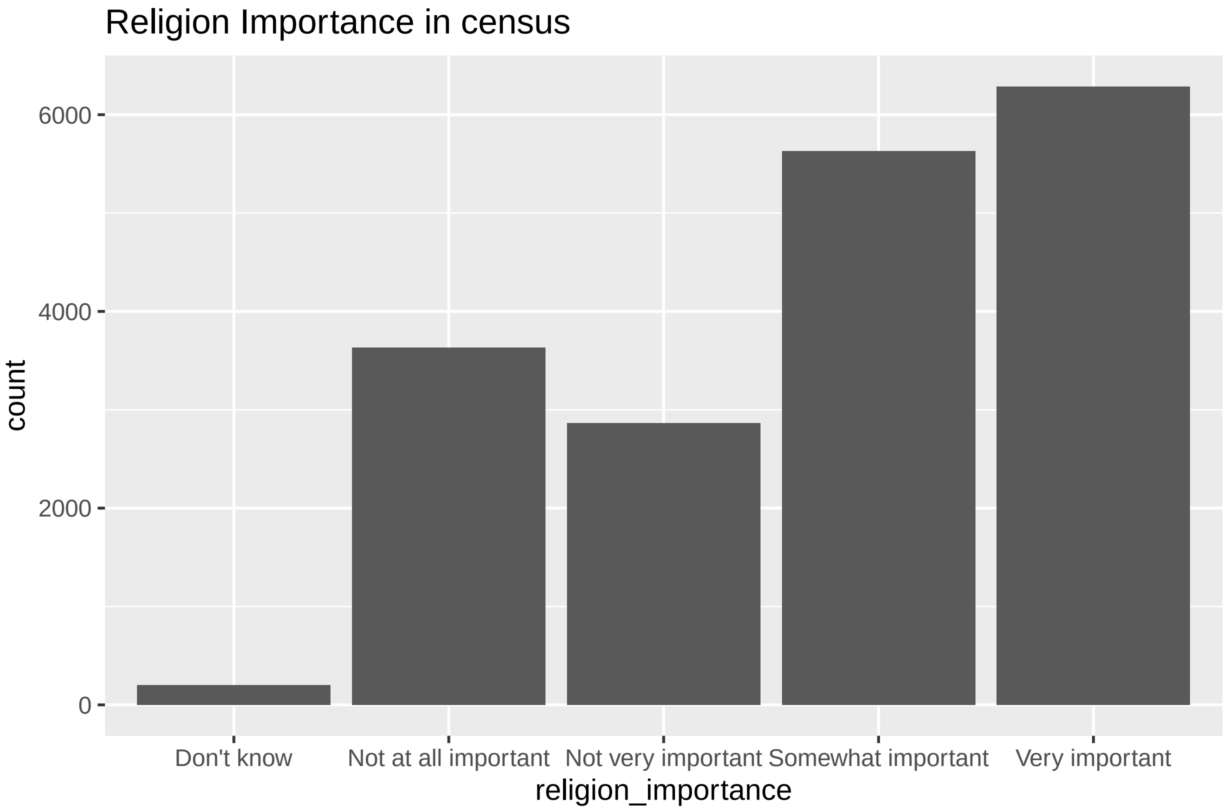

Figure 3

These are the two barplots for the variable religion_importance both in census and survey data. The variable's behavior is similar in these two data sets. They both have higher number of people who think religion is somewhat important or very important in their lives. In survey, the number of "somewhat important" people is higher than that of "very important" people, but in census, the situation is completely the opposite, the number of "very important" people is higher. Both data sets have very few people in "Don't know" section. Both of these two barplots are left-skewed, but this pattern is less clear in the census plot because the number of "Not very important" people is lower than that of "Not at all important" people. However, in survey data, the number of "Not very important" people is higher compared to the bar of "Not at all important". Overall, the variable behaves similar in these two data sets, so it should not be used in the post-stratification later.

All analysis for this report was programmed using `R version 4.1.1`.

Methods

Logistic Regression

Logistic regression is an effective and powerful method to analyze the influence of a group of predictor variables on binary results by quantifying the unique contribution of each predictor variable. It models the probability with logit-transformed as a linear relationship with the predictor variables. More formally, let Y be the outcome variable indicating failure or success with 0 and 1 and p be the probability of Y being the positive outcome p(Y=1). And let X_1,X_2,…,X_n be a a group of independent variables. Then the logistic regression of Y on X_1,X_2,…,X_n estimates parameter values for β_0, β_1,…,β_n via the maximum likelihood method of this equation:

$$logit(p) = log\frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + +\beta_n X_n$$

Assumptions in Logistic Regression

In simple linear regression we know that if we want to fit this model we must satisfy these four assumptions which is linearity, homoscedasticity, independence and normality. And same logic in logistic regression. It doesn't work for all conditions but only works under certain conditions. Therefore, based on Statology Study (2020) we made four assumptions below:

Assumption 1: The Response Variables are Binary

We assume that the response variables can only have two possible outcomes.
For instance: Male or Female, Success or Failure, Yes or No, Good or Bad, Malignant or Benign

Assumption 2: Each Observation is Independent

We assume that the observations in the dataset are independent, which means that each observation should come from non-repeated measurements of the different individual or unrelated to each other.

Assumption 3: Exists a Linear Relationship Between Explanatory Variables and Logit of Response Variable

We assume that there exists a linear relationship between each explanatory variable and the logit of the response variable, which defined as:

logit(p) = log(p/(1-p)) where p is the probability of positive outcome.

Assumption 4: Sample Size is Sufficiently Large

We assume that the sample size of the dataset is large enough to get valid conclusions from the fitted logistic regression model.

Model Specifics

In this project, we are going to use logistic regression model to build the proportion of voters who will vote for the Conservative Party since our predicted variables have numeric variable, binary variable and categorical variables.

The model outcomes are the log of odds for age (numeric variable), gender (binary variable with Male or Female), education level (categorical variable) and religion importance (categorical variable).

As we mentioned before, we assume that these response variables are binary, each observation is independent, there exists a linear relationship between explanatory variables and logit of response variable and sample size is sufficiently large enough.

Then, we fit this logistic regression model on survey data, which is

$$logit(p) = log\frac{p}{1-p} = \beta_0 + \beta_1 x_{age} + \beta_2 x_{gender} + \beta_3 x_{educationlevel} + \beta_4 x_{regilionimportance}$$

where p represents is the estimated probability of the positive event occurrence, which is the probability that people vote yes for the Conservative Party.

β_0 represents the intercept of the logistic model, which means that the log of odds of voting for the Conservative Party when the observations are at a certain age, gender, education level and religion importance.

β_1 represents the slope of age, which means that when age increases for one unit, the log of odds of voting for the Conservatives Party increase or decrease by a factor of controlling other variables constant.

β_2 represents the average difference in log odds of voting for the Conservative Party between gender (Male and Female) for a certain age, education level and religion importance.

β_3 represents the average difference in log odds of voting for the Conservative Party between education level for a certain age, gender and religion importance.

β_4 represents the average difference in log odds of voting for the Conservative Party between religion importance for a certain age, gender and education level.

Post-Stratification

Post-stratification is a statistical technique that used for correcting model estimates for obvious differences between the sample population (census data) and the target population (survey data). To be more specific, it is a process of adjusting the estimates, essentially a weighted average of estimates from all possible combinations of attributes. Each combination we called is a "cell" based on different category of variables. And then, using the logistic regression model we just fit to do estimation in each cell with too little data by using overall or nearby averages.

In this project, we will create cells based on different ages, gender, education level and religion importance by using the logistic regression model we just fit on survey data to estimate the proportion of voters in each group (bin). Then, we will weight each proportion estimate (within each bin) by the respective population size of that bin and sum these values and divide that by the total population size.

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

where N_j is the bin number and $\hat{y_j}$ is the expected value of each bin.

All analysis for this report was programmed using `R version 4.1.1`.

Results

$$log\frac{\hat{p}}{1-\hat{p}} = -1.359435 - 0.003030X_{age} + 0.712453X_{male} + 0.424613Xcollege + 0.386933X_{highschool}$$
$$+0.487533X_{university} - 0.427686X_{abovebachelor} + 0.303352X_{notimportantatall} + 0.201569X_{notveryimportant}$$
$$+0.645027X_{omeimportant} + 1.116041X_{veryimportant}$$

Model Performance Results

Table 6: Model Performance

term	estimate	std.error	statistic	p.value
(Intercept)	- 1.3594352	1.1795768	- 1.1524770	0.2491251
age	- 0.0030304	0.0032981	- 0.9188413	0.3581786
genderMale	0.7124535	0.1102908	6.4597698	0.0000000
education_levelCollege, CEGEP or other non-university certificate or di...	0.4246131	0.1446527	2.9353979	0.0033312

term	estimate	std.error	statistic	p.value
education_levelHigh school diploma or a high school equivalency certificate	0.3869331	0.1719299	2.2505283	0.0244154
education_levelLess than high school diploma or its equivalent	0.4875330	0.2379379	2.0489931	0.0404628
education_levelUniversity certificate or diploma below the bachelor's level	0.3505559	0.2132511	1.6438644	0.1002042
education_levelUniversity certificate, diploma or degree above the bach…	-0.4276858	0.1821514	-2.3479689	0.0188761
religion_importanceNot at all important	0.3033524	1.1728668	0.2586419	0.7959116
religion_importanceNot very important	0.2015692	1.1660248	0.1728687	0.8627547
religion_importanceSomewhat important	0.6450272	1.1621095	0.5550486	0.5788614
religion_importanceVery important	1.1160412	1.1625466	0.9599969	0.3370568

Table 7: post-stratification

predict
0.4391372

Table 8: First five groups of post-stratification for each group

age	gender	education_level	religion_importance	N	estimate
18	Female	High school diploma or a high school equivalency certificate	Somewhat important	2	0.4056391
18	Female	Less than high school diploma or its equivalent	Not at all important	1	0.3490772
18	Female	Less than high school diploma or its equivalent	Somewhat important	1	0.4301049
18	Female	Less than high school diploma or its equivalent	Very important	1	0.5472563
18	Male	High school diploma or a high school equivalency certificate	Not very important	1	0.4717732

Table 9: Total Groups

n
11340

Finally, we decided to use GLM, which includes 4 predictors, namely age, gender, education level and religion importance. We used these 4 variables to predict that people would vote for the Conservative Party in this model with a post stratification of 0.4391372, which means that about 43.91% of people in the population would vote for the Conservative Party. On the other hand, the sample proportion of votes for Conservative Party is around 43.7%, so it is safe to say that our estimation is pretty accurate.

Table 2 shows that we have divided into four groups: age, gender, education level and religious importance. We divided 11,340 groups and calculated what percentage of people in each group would vote for the Conservative Party. Finally, we calculated the average of these groups, which is 43.91%.Worth to know that even some variables have really high p-value, which means they are not significant. In these predictors, gender, education level at college, education level at high school,education level at less than high school and education level above bachelor are significant. Other predictors are insignificant such as age and religion importance. Age and education level at above bachelor have negative relationship with the log odd ratio, other predictors have positive relationship with log odd ratio. However, the goal of this model is predict how many people are going to vote for Conservatives party, not for canalize which predictor effect how many people going to vote. So it is fine to keep insignificant predictors.

For post stratification of 0.4391372, this figure is reasonable because even if the Conservatives lose the election in 2019, they still have many supporters in the Vancouver and Toronto areas. Also the Conservatives have said they will spend $60 billion on health care if their campaign is successful. This move has allowed him to gain many older supporters. Also they have plans for better control of immigration, jobs, etc. In the post stratification, we have 11,340 groups, and each group is divided into 4 categories: age, gender, education level and religion importance.In the table we can see the situation of each group and the predicted value of each group.

Conclusions

Our hypothesis is to assume the ages, genders, education levels, and religions of people have a significant influence on the support rate of the Conservative party. Therefore, we use logistic regression to model the proportion of voters who will vote for the Conservative party by those variables we think will impact. Then we use the post-stratification calculation to overcome that our survey data of the target population does not reflect the distribution of variables in the census data of the population. After dividing our data into 11,340 groups based on our hypothesis and using the logistic model to estimate the proportion of voters in each group, the post-stratification of 0.4391372 comes up, which means we estimate that about 43.91% of people in the population would vote for the Conservative Party. The census data of 2019 this report reached is not up-to-date, which may affect the study results. Also, only four possible impact factors have been discussed. Some of them with high p-values are not significant, so we can include more factors that really influence the choices of voters and newer census data in the future study. Moreover, our model for estimation may not be the most appropriate. The multilevel regression model will be better for the data measured for different levels, such as individuals and groups.

Bibliography

1. Government of Canada, S. C. (2017, February 27). The General Social Survey: An overview. Government of Canada, Statistics Canada. Retrieved November 5, 2021, from https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2013001-eng.htm.

2. 2019 Canadian Election Study - Online Survey Technical Report and Codebook.pdf - Harvard Dataverse. (2021). Retrieved 5 November 2021, from https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/DUS88V/HRZ21G&version=1.0)

3. Voting Behaviour in Canada | The Canadian Encyclopedia. (2021). Retrieved 5 November 2021, from https://www.thecanadianencyclopedia.ca/en/article/electoral-behaviour

4. Feess, S. (2021). Does education influence voter turnout?. Retrieved 5 November 2021, from https://www.grin.com/document/101356

5. Statology Study. (2020, Octber 13). *The 6 Assumptions of Logistic Regression.* Retrieved from: https://www.statology.org/assumptions-of-logistic-regression/.

6. UCLA Institute for Digital Research & Education Statistical Consulting. (2020). *How do I Interpret Odds Ratios in logistic regression model?.* Retrieved from: https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/.

7. Wikipedia contributors. *Multilevel Regression with Post-Stratification.* In Wikipedia, The Free Encyclopedia. Retrieved from: https://en.wikipedia.org/wiki/Multilevel_regression_with_poststratification. (Last Edited: September 24, 2021).

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

9. Olsen, H. (2021, September 20). Opinion | conservatives in Canada are hitting a political sweet spot. The Washington Post. Retrieved November 5, 2021, from https://www.washingtonpost.com/opinions/2021/09/20/conservatives-canada-are-hitting-political-sweet-spot/.

10. Wikimedia Foundation. (2021, November 4). Election. Wikipedia. Retrieved November 5, 2021, from https://en.wikipedia.org/wiki/Election.

11. Religious views as a predictor of ... - Chapman University. (n.d.). Retrieved November 5, 2021, from https://digitalcommons.chapman.edu/cgi/viewcontent.cgi?article=1034&context=e-Research

12. Wikimedia Foundation. (2021, October 31). Conservative Party of Canada. Wikipedia. Retrieved November 5, 2021, from https://en.wikipedia.org/wiki/Conservative_Party_of_Canada#Principles_and_policies.

13. BBC. (n.d.). Long-term factors - factors influencing voting behaviour - higher modern studies revision - BBC Bitesize. BBC News. Retrieved November 5, 2021, from https://www.bbc.co.uk/bitesize/guides/zd9bd6f/revision/9.

14. Suls, R. (2020, August 28). Educational divide in vote preferences on track to be wider than in recent elections. Pew Research Center. Retrieved November 4, 2021, from https://www.pewresearch.org/fact-tank/2016/09/15/educational-divide-in-vote-preferences-on-track-to-be-wider-than-in-recent-elections/