The analysis of popular vote for the Liberal and the Conservative party in 2025 Canadian Federal Election

STA304 - Assignment 2

Yiliu Cao

November 24, 2022

Introduction

The Canadian federal election is held every four years, and all Canadian aged or above 18 are eligible to vote [10]. The percentage of votes cast for a candidate or his party by voters is called the overall popular vote of this party, and it plays an important role in deciding who will be the new prime minister and which party will govern the country. In this analysis, I will predict the overall popular vote for the Liberal and the Conservative party and the winner of the 2025 Canadian federal election.

In this analysis, I will only predict the overall popular vote for the Liberal and the Conservative party. This is because the popular vote of these two parties is much higher than the rest parties in the last federal election (around 30%) [4], and the highest popular vote of the rest is just 17%. Therefore, the winner of the next federal election should still be between these two parties, and thus I will only make predictions for these two parties.

During the last federal election (2021) [4], the Liberal party beat the Conservative party, and Justin Trudeau won his third term as the prime minister of Canada. However, the Liberal party only won the minority government instead of the majority government, and one reason is that the Liberal had less popular vote than the Conservative party. The majority government formed by a party means this party has more than half of the seats in the House of Commons, unlike the minority government, which is formed when no political party has a majority of seats in the House of Commons [5]. In other words, the majority government can have more power than the minority ones. However, Justin Trudeau hopes to win the majority government in the next federal election (2025) to have more power. Therefore the Liberal party needs a higher popular vote than in the last federal election. Meanwhile, his main competitor, the Conservative party, won't allow the Liberal party to win the majority government as they will be less powerful. Hence, the Conservative party also needs a higher popular vote to beat the liberal party in the next federal election.

Therefore, it is necessary and important to predict the overall popular vote for the Liberal and the Conservative party for the next federal election (2025). The prediction can help these two parties to take appropriate actions in advance to increase their popular vote if the results do not match their expectations and take specific strategies for the next election to beat their competitors. For the voters and residents, predicting the winner of the next federal election can tell them which party will govern the country, the possible changes in the country's development, and how the changes may impact their lives.

The data I will use to predict the popular vote for the two parties are the **General Social Survey Data 2017 (GSS 2017)** [6] and **Canada Election Survey Data 2019 (CES 2019)** [7]. The General Social Survey Data 2017 is the census data of Canada in 2017, and it is made by Statistics Canada. Its target population is all non-institutionalized persons living in Canada with 15 years old and older. It aims to find the changes in living conditions and well-being and Canadians over time. GSS 2017 is the census data in our analysis, which can provide information about different Canadians. The Canada Election Survey Data 2019 records the phone survey conducted by the CES team to gather Canadians' attitudes and opinions of political behavior during and after the 2019 Canadian federal election. It is the sample data of this analysis that can help us to build the model to predict the popular vote for the two parties.

Moreover, from the results of the last federal election (2021) on Wikipedia [4], we can see that the popular vote for the Liberal party is less than the Conservative party (32.62% vs. 33.74%). Therefore, **my hypothesis will be: Based on the outcome of the 2021 federal election [4], I predict that the Conservative party will again yield a higher popular vote than the Liberal party in the 2025 Canadian federal election.** Besides, it is wired that even though the Liberal party has a lower popular vote than the Conservative, the Liberal party won and governed the country. The reasons are complicated, and it is not the topic of this assignment. I will also assume that **the party with a higher overall popular vote will be the winner in the federal election.**

Combining the hypotheses and my assumption, my research question will be:
Based on the data of GSS 2017 and CES 2019, which party of the Liberal party and the Conservative party is more likely to be the winner of the 2025 Canadian federal election? I will conclude the answers to my research question in a different section, including Data, Methods, Results, and Conclusion.

Data

In the introduction, I introduced the data I will use in this analysis: the General Social Survey Data 2017 (GSS 2017) and Canada Election Survey Data 2019 (CES 2019). However, the collection process of these two data sets is different, and I will introduce it separately.

For General Social Survey Data 2017 (GSS 2017):
It was conducted from February 2nd to November 30th, 2017, and was made by Statistics Canada [7]. It is collected by phone, and all interviewing took place using centralized telephone facilities in five of Statistics Canada's regional offices. Each record in the survey frame was assigned to a stratum within its province, and simple random sample without replacement of records was next performed in each stratum. To grab the General Social Survey Data 2017, you can access the CHASS website of the U of T library [8].

For the Canada Election Survey Data 2019:
It will be conducted starting on October 21st, 2019, and collected via phone and email. The people on the Campaign-Period Survey (CPS) were called or emailed after the election according to their stated preference and asked to complete the Post-Election Survey (PES). The collection process is finished 31 days after the election, with the final interviews completed on November 21st, 2019 [12].
You can find CES data on their official website [6]. Alternatively, you can also use the R package "cesR" to grab the data [9].

Cleaning process

The original forms of the two data sets are too big and contain many variables we may not need for the analysis; therefore, I will clean both data sets.

For **GSS 2017**:

1. I remove the participants (observations) whose age is below 18 as only the people with age is or higher than 18 are eligible for voting [10]. I also remove the observations if any values of age, sex, and education level are missing.

2. I transfer the age of each participant to a rounded number, meaning that the values of the age of each participant are now integers instead of decimal numbers. For example, the participant's age will be 55 instead of 55.1.

3. I create a new variable indicating the University completion of each participant. If the participants' education level is or higher than the university (e.g., Bachelor's degree), the value of this variable will be "Completed." Otherwise, it is "Uncompleted," meaning that their highest education level is lower than the university (e.g., the high school diploma). Besides, those participants whose education level can not indicate whether they completed the university will have a value of "Unknown."

4. I only chose the variables indicating the participants' age, sex, and university completion (the new variable I created), and removed all other variables.

For **CES 2019**:
Comparing to the GSS 2017, the CES 2019 can give us extra information about each participant's voting preference, and the cleaning process will be different.

1. I create two new variables indicating whether the participants will vote for the Liberal and the Conservative party. For each variable, "1" means they will vote for the party, and "0" means they will not vote for this party. This refers to Question eleven (q11) in the CES 2019. Question eleven is "Which party will you likely vote for" and "1" means voting for the Liberal party, and "2" means voting for the Conservative party.
   Besides, if both variables have a value of "0," it means that they will not vote for any of these two parties (they may vote for other parties or refuse/skip the question when they were surveyed).

2. I create a new variable indicating each participant's age. The age of each participant is calculated based on their born years from the survey data, and it equals to survey year (2019) minus their born year. The born year refers to Question 2 (q2) on the original data, which is "In what year were you born?".

3. I create a new variable indicating the sex of each participant. The value of this new variable will be "Male" for male participants and "Female" for female participants. The sex of each participant refers to Question 3 (q3) in the CES 2019 ("1" means "Male" and "2" for "Female"). For any other values, I will set them to "Other."

4. Similar to the previous part, I create a new variable indicating the whether the voter has completed their university. If the participants have an education level is or higher than the university (e.g., a Master's degree), they will be marked as "Completed" for this variable. Otherwise, it is "Uncompleted," such as Completed secondary/high school) Besides, the value of this new variable will be "Unknown" for those participants whose education level is unknown (This could be such cases that they refuse or skip the question, etc.). This refers to Question 61 (q61) in CES 2019, which is "What is the highest level of education that you have completed".

5. I only select the five new variables I create and remove all the rest variables.

6. I only keep the participants whose age is higher than 18 as only the people aged higher than 18 are eligible for voting. I also remove all voters whose any values of the five variables is a missing value.

Descriptions and summary of variables

After introducing the data sets and the cleaning process, I want to introduce all the important variables in the two cleaned data sets. Note that only CES 2019 contains the variable indicating the party they will vote for.

The common variables (appear in both data sets) are:
Age:
This is a numerical variable indicating each participant's age, and it is an integer. The coded name is **"age"**. The voter's age is essential as the order and young people may vote for different parties. For example, older people may vote for the party that cares more about people's health, whereas young people may not.
The below table shows the summary measures for age in two data sets.

Table 1: Summry measures for age

Data Name	Variable name	Minimum Age	Median Age	Maximum Age	Mean Age	Variance	IQR
GSS 2017	age	18	55	80	52.93408	294.0764	29
CES 2019	age	18	51	100	50.89992	283.5215	26

From Table 1, we can see that the minimum age is 18 in both data sets, and the median and mean ages for the two data sets are all around 50. Besides, the age spread in both data is similar, 29 for GSS 2019 and 26 for CES 2019. However, the maximum age in CES 2019 is higher than GSS 2017 (100 vs. 80).

Sex:
This is a categorical variable and indicates the sex of each participant. It has three possible outcomes: Male, Female and Other. The coded name is **"sex"**.
People of a different sex may also have different voting preferences. For example, females will vote for the party supporting women's rights more and protecting women's status in the workplace, whereas Males will not.
The below table shows the summary measures of sex in the two data sets.

Table 2: Summary measures for sex

Data Name	Sex	Total Number	Proportion
GSS 2017	Male	9032	0.4552190
	Female	10809	0.5447810
CES 2019	Male	2218	0.5633731
	Female	1718	0.4363729
	Other	1	0.0002540

From Table 2, we can see that the proportion of Males and Females in each data set is approximately fifty to fifty, but the total number in GSS 2017 is larger than in CES 2019. Besides, only the CES 2019 has an observation whose gender is "Other."

Completion of University:
This is a categorical variable and indicates whether each participant's highest education level is higher than the university. "Completed" means they at least completed their university, and "Uncompleted" means their highest education level is lower than the university, such as high school. The coded name is **"complete_university"**.
As with the first two variables, people with different highest levels of education may also have different voting preferences. People with less education may be less informed about politics and may not think long-term when voting. However, highly educated people may better understand how society and politics work, and they may vote based on which party's policies are really good for the country in the future. For example, people with higher education may consider economic impacts (globalization, trade, etc.) for different parties governing the country, which may change their voting preference.
The below table shows the summary measures of education level in the two data sets.

Table 3: Summary measures for completion of University

Data Name	Status	Total	Proportion
GSS 2017	Completed	6327	0.3188851
	Uncompleted	13514	0.6811149
	Unknown	0	0.0000000
CES 2019	Completed	2049	0.5204470
	Uncompleted	1877	0.4767590
	Unknown	11	0.0027940

From Table 3, we can see that there are relatively more people who have a value of "Completed" in CES 2019 than in GSS 2017 (0.52 vs. 0.31), meaning that there are more people with higher education in CES 2019. However, the number of "Unknown" completion of University is 0 in GSS 2017 but 11 in CES 2019. This means we do not know the 11 participants' status of university completion in GSS 2017.

The following two variables are unique in CES 2019:
Voting for the Liberal party:
This is a binary variable indicating whether the participant will vote for the Liberal party. The binary outcomes are "1" and "0". "1" means that the participant will vote for the Liberal party, and "0" means not. The coded name is **"vote_liberal"**.

Table 4: Summary measures for voting for the Liberal party

Variable Name	Sum of voting	Total number of people	Proportion
Voting for the Liberal party	909	3937	0.2308865

Voting for the Conservative party:
Similar to vote_liberal, this is also a binary variable indicating whether the participant will vote for the Conservative party. The binary outcomes are also "1" and "0". "1" means that the participant will vote for the Conservative party, and "0" means not. The coded name is **"vote_conservative"**.

Table 5: Summary measures for voting for the Conservative party

Variable Name	Sum of voting	Total number of people	Proportion
Voting for the Conservative party	980	3937	0.2489205

Tables 4 and 5 show the number and proportion of participants voting for the Liberal and the Conservative party in CES 2019, respectively. We can see that the proportion of people voting for the Conservative Party is about 0.01 higher than the Liberal party, which is reasonable as it matches the results of the last (2021) federal election [4].

Furthermore, in order to take a closer look at whether voters with different education levels will influence their voting preference. I will draw two figures.

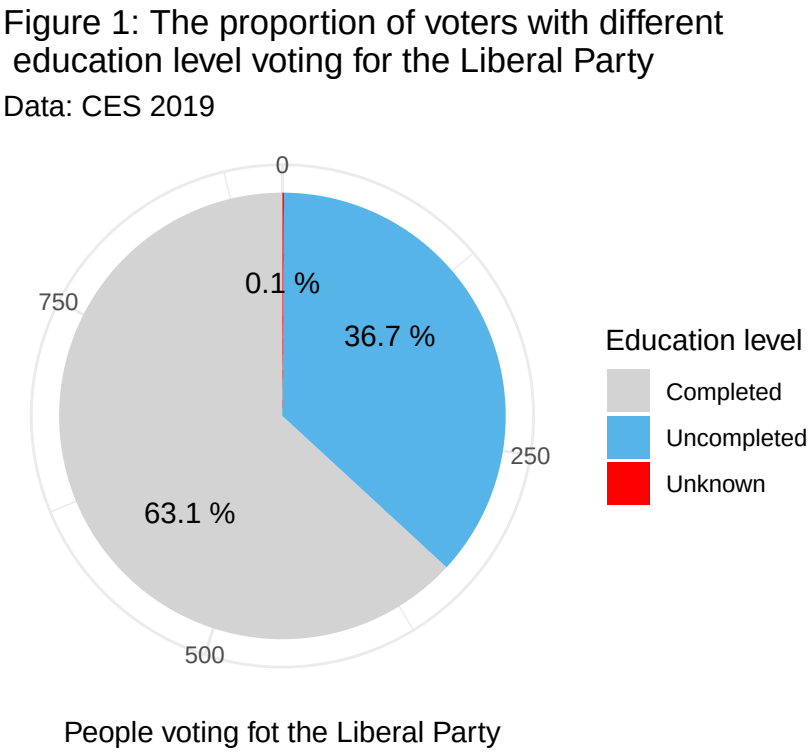
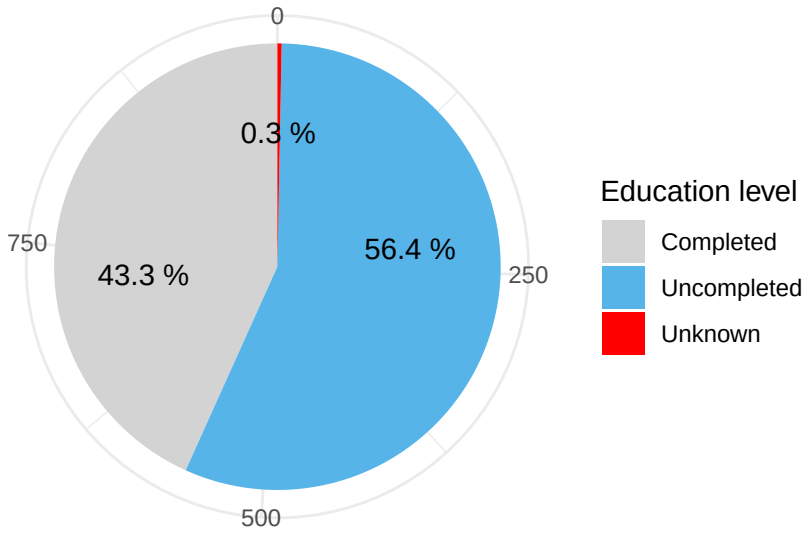


Figure 1 shows the proportion of voters with different education level voting for the Liberal party. We can see that 63.1% of people voting for the Liberal party have at least completed their university, which is much higher than the proportion of people who did not complete their university (36.7%). It seems that the people with a higher education level have a flavor of the Liberal Party.

Figure 2: The proportion of voters with different education level voting for the Conservative Party

Data: CES 2019



People voting for the Conservative party

Figure 2 shows the proportion of voters with different education level voting for the Conservative party. We can find that 56.4% of voters who vote for the Conservative party have not completed their university, which is about 13% higher than the proportion of people who have completed their university.
Comparing Figure 1 and Figure 2, it is interesting that the voters who at least completed their university dominate the voters voting for the Liberal party (63.1% vs. 36.7%). In contrast, more voters voting for the Conservative party did not complete their university (43.3% vs. 56.4%). It seems that the people with a higher education level seem more likely to vote for the Liberal party, and those people without a university education seem more inclined to the Conservative party. I will go through more details in the next part.

Methods

As I introduced in the introduction, this analysis aims to predict the overall popular vote for different parties and the winner of the 2025 Canadian federal election. And the previous part shows you the data I will use. In this part, I will show you the methods to predict the overall popular vote for each party.

Model Specifics

We know that there will be binary outcomes for each voter when they consider voting for a party: vote or not vote. Statistically, the probability of a voter voting for the Liberal/Conservative party is 1 or 0. If we want to predict the likelihood of votes voting for the Liberal or the Conservative party, we need to use the models which can predict the probability of each outcome given the voters' information. The logistic models are the most appropriate ones.
The logistic regression model is the statistical model that uses a logistic function to model a binary response variable and find the probability of each outcome. Thus we can use the logistic models to predict the probability of voters voting for a party. Moreover, since we want to predict the popular vote for both the Liberal and the Conservative party, we need to build a logistic model for each party.
Besides, I will also use poststratification methods to divide the population into different cells and calculate the popular vote for each cell. This required that each variable in the two data sets have the same categories.

However, we know from the Data section that only the GSS 2017 has participants with a value of "Other" in sex and "Unknown" in completion of university. Thus we have to remove this observation to ensure the two data sets match each other.

Hence, the logistic model for each party is:

For the Liberal party:
To predict the overall popular vote for the Liberal party, I will use a multiple logistic regression model to model the probability of a voter who will vote for the Liberal party. I will use age, sex, and status of university completion to model the probability of voting for the Liberal party. The multiple logistic regression model I am using is:

Equation 1:

$$log(\frac{\hat{p}_l}{1-\hat{p}_l}) = \hat{\beta}_0 + \hat{\beta}_1 x_{age} + \hat{\beta}_2 x_{Male} + \hat{\beta}_3 x_{Uncompleted}$$

Equation 1 is the logistic model for predicting the popular vote for the Liberal party, where:
- p_l represents the probability of voters voting for the Liberal party.
- β_0 represents the intercept of the logistic model. The value of voting for the Liberal party will be - β_0 if all other variables are 0.
- β_1 represents the change in log odds for every one unit increase in age.
- β_2 represents the difference in log odds for different sex. There will be a β_2 increase in log odds of males voting for the Liberal compared to females.
- β_3 represents the difference in log odds of voting for the Liberal people for people who did not complete the university compared to those who finished their university.

For the Conservative party:
To predict the overall popular vote for the Conservative party, the model is similar to the previous one. I will also use a multiple logistic regression model to model the probability of a voter who will vote for the Conservative party. I will use age, gender, and completion of the university to model the probability of voting for the liberal party. The logistic model will be:

Equation 2:

$$log(\frac{\hat{p}_c}{1-\hat{p}_c}) = \hat{\beta}_0 + \hat{\beta}_1 x_{age} + \hat{\beta}_2 x_{Male} + \hat{\beta}_3 x_{Uncompleted}$$

Equation 2 is the logistic model for predicting the popular vote for the Conservative party, where:
- p_c represents the probability of voters voting for the Conservative party.
- β_0 represents the intercept of the logistic model. The value of voting for the Conservative party will be β_0 if all other variables are 0.
- β_1 represents the change in log odds for every one unit increase in age.
- β_2 represents the difference in log odds for different sex. There will be a β_2 increase in log odds of males voting for the conservative compared to females.
- β_3 represents the difference in log odds of voting for the Liberal people for people who did not complete the university compared to those who finished their university.

Check assumptions of the two logistic models

After we build the logistic models for each party, we need to check their assumptions. We may need to use some tools from R package "car" [11] to check the assumptions [2]:

1. Outcome is binary: p_l and p_c can only equal 1 or 0, meaning that the voter will or will not vote for the Liberal party and the Conservative party. Therefore the assumption of binary outcome for both models is satisfied.

2. Linearity in the logit for continuous variable: We will use Box-Tidwell Test to test the assumption of linearity. The Box-Tidwell test shows a strong linearity between the log odds and "age" for the Conservative party model. But there is not a very strong linearity between the log odds and "age." for the Liberal party model. However, it does not necessarily mean there is NO linearity between them. I will discuss this limitation in the "Conclusion" part.

3. Absence of multicollinearity: This assumption is to check whether the data contain highly correlated predictor variables. We can check it by variance inflation factors (vif). Variance inflation factors measure the amount of multicollinearity in regression analysis. After we calculate the vif values for the two models, it suggests that there is no multicollinearity in both models, meaning that the three variables in both models are independent of each other.

4. Lack of strongly influential outliers: We will use Cook's distance by drawing the plot to check any influential points. The plots show no data points in both models with an absolute standardized residual above 3. Therefore the assumption of a lack of strongly influential outliers is held.

After I introduce the logistic models I will use, I want to introduce the method of post-stratification.

Post-Stratification

From the Data section, we know that the collection of the CES 2019 is via phone and email. However, we also know that only those who completed the Campaign-Period Survey (CPS) were called or emailed after the election. They did not randomly choose the people to fill out this survey (i.e., no random selection). This means that the samples in CES 2019 may be non-probability sampling, and the sample may not be representative of all residents in Canada. Therefore, in order to predict the popular vote in a more precise way, we need to use the census data (GSS 2017) and use the method of poststratification to predict the overall popular vote of each party.
Poststratification is the process of dividing the total population into different cells and calculating the estimates we are interested in each cell as well as their weights to the total population. Then summarize the weight of each cell and the estimate, we can have an adjusted estimate $\hat{y}^{PS}$. The poststratification method can help us "extrapolate" how the entire population will behave (e.g., how the entire population will vote). Since we divide the population into different cells and find each cell's weights, it can decrease the bias and increase the precision of our model even though the sample in CES 2019 may be non-representative.
Therefore, we can use poststratification to predict the overall popular vote for the two parties. From the logistic models we constructed previously, I already stated that voters' age, sex, and education level could influence their voting preferences (Figure 1 and Figure 2). Thus the cell partitions will be age, sex, and status of university completion, meaning that people within a cell will have the same age, sex, and university completion status. For example, one cell can be all 30 years old males who have completed their university or 40 years old females who did not complete their university.

If we express the poststratification with our model in terms of mathematical form, it will be look like:

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

where:
- $\hat{y}_j$ is the popular vote for the Liberal party or the Conservative party in each cell.
- N_j is the population size in the jth cell, and thus $\sum N_j$ is the total population.
- $\hat{y}^{PS}$ is hence the overall popular vote for each party.

After we build the logistic model for each party, we can fit the logistic models and estimate the coefficients. Then combine with the poststratification, we can have the results of the overall vote for each party. The results are shown in the next part.

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

Results

Table 6: The results of the logistic model for the Liberal party

Coefficients	Corresponding preidctor	Estimate	Standard Error	Statistics	P-value
$\hat{\beta}_0$	NA	-1.3271757	0.1350909	-9.8243148	0.000000
$\hat{\beta}_1$	x_{age}	0.0081799	0.0022816	3.5851460	0.000337
$\hat{\beta}_2$	x_{Male}	-0.0698075	0.0768922	-0.9078621	0.363951
$\hat{\beta}_3$	$x_{Uncompleted}$	-0.5840980	0.0781342	-7.4755751	0.000000

The table 6 shows the results of the logistic model of voting for the Liberal party. From table 6, we can identity our regression equation, which is:

Equation 3:

$$log(\frac{\hat{p}_l}{1-\hat{p}_l}) = -1.327 + 0.008x_{age} - 0.070x_{Male} - 0.584x_{Uncompleted}$$

The equation 3 is the true logistic model of voting for the Liberal party. Therefore, we can interpret the model:
• The intercept of our model is -1.327, meaning that the log odds of a female voter with a age of 0 and completed her university will be -1.327. However, the age can not be 0.
• Every one unit increase in a participant's age, there will be 0.008 increase in log odds of voting for the Liberal party.
• The difference in log odds of voting for the Liberal party between different sex (Male and Female) is -0.070, meaning that the log odds of Female is 0.070 lower than the Male under a certain age and education level.
• The difference in log odds of voting for the Liberal party between different education level (completed and uncompleted university) is -0.584, meaning that the log odds of voting for the Liberal party of people with higher education level is 0.584 higher than that of voters with a lower education level under certain age and sex.
Besides, we can see that the p-values for each each coefficients is small except β_2, meaning the sex seems not significantly related to probability of voting for the Liberal party.

Table 7: The results of the logistic model for the Conservative party

Coefficients	Corresponding predictor	Estimate	Standard Error	Statistics	P-value
$\hat{\beta}_0$	NA	-2.1691604	0.1401433	-15.478163	0.0e+00
$\hat{\beta}_1$	x_{age}	0.0094775	0.0022332	4.243938	2.2e-05

Coefficients	Corresponding predictor	Estimate	Standard Error	Statistics	P-value
$\hat{\beta}_2$	x_{Male}	0.5841310	0.0780334	7.485655	0.0e+00
$\hat{\beta}_3$	$x_{Uncompleted}$	0.4449879	0.0751775	5.919161	0.0e+00

The table 7 shows the results of the logistic model of voting for the Conservative party, we can also identity our regression equation, which is:

Equation 4:

$$log(\frac{\hat{p}_c}{1-\hat{p}_c}) = -2.169 + 0.009x_{age} + 0.584x_{Male} + 0.444x_{Uncompleted}$$

Similarly, the equation 4 is the true logistic model of voting for the Conservative party, we can interpret the model:
• The intercept of our model is -2.169, meaning that the log odds of a Female voter with a age of 0 and completed her university will be -1.327.
• Every one unit increase in a participant's age, there will be 0.009 increase in log odds of voting for the Conservative party.
• The difference in log odds of voting for the Liberal party between different sex (Male and Female) is 0.584, meaning that the log odds of Male is 0.584 higher than the Female under a certain age and education level.
• The difference in log odds of voting for the Conservative party between different education level (completed and uncompleted university) is 0.444, meaning that the log odds of voting for the Conservative party for people with lower education level is 0.444 higher than that of voters with a higher education level under certain age and sex.
If we look on p-values, we can see that the p-values of each coefficients are very small, meaning that each predictor is significantly related to the response.

Comparing the two models, we can find that the value of log odds will be higher for the voter with a higher education level for the Liberal party, whereas it will be higher for voters whose education level is relatively low for the Conservative party. This results match the Figure 1 and Figure 2, people with a higher education level seem to vote the Liberal party more than the Conservative party.

After we have the regression equations of each model, we can use the methods of poststratification to predict the overall popular vote for each party. The results is shown below.

The Table 8 shows the first six cells in the Poststratification.

Table 8: The first six cells of Poststratification

Age	Sex	Completion of University	Population	Vote for the Liberal	Vote for the Conservative
18	Female	Uncompleted	30	0.1462856	0.1745701
18	Male	Uncompleted	41	0.1377811	0.2749907
19	Female	Completed	1	0.2365374	0.1203533
19	Female	Uncompleted	61	0.1473101	0.1759399
19	Male	Completed	1	0.2241637	0.1970306
19	Male	Uncompleted	57	0.1387557	0.2768842

From Table 8, we can see that the each cell contains at least one different values of variable. For example, the first cell contains all females who is 18 years old and did not complete their university. We can see

that there are 30 females in each cell, and the probability of these 30 females voting for he Liberal party is 0.1462856 and 0.1745701 for the Conservative party.
There are many other cells and each cell contain different age, sex and status of completion of university.

The Table 9 shows the overall popular vote for the two parties.

Table 9: The predicted overall popular vote for the Liberal and the Conservative party

Party	Overall Popular Vote ($\hat{y}^{PS}$)
The Liberal Party	0.2143414
The Conservative Party	0.2572216

The table 9 shows the predicted overall popular vote for the Liberal and the Conservative party from our analysis. We can see that the predicted overall popular vote for the Liberal and the Conservative are about 0.214 and 0.257, respectively. This means the Conservative party will be the winner in the next Canadian federal election, and the Liberal will lose their power, Justin Trudeau can not have his fourth term as a prime minister.
Besides, the result we have of the overall popular vote for the Liberal and the Conservative party make senses. It matches the results in the 2021 Canadian federal election and also consistent with our hypothesis in the Introduction section, which the populat vote of the Conservative party is higher than the Liberal party.

Finally, we can now answer our research question. We can conclude that the Conservative party will be the winner in 2025 Canadian federal election based on the data of GSS 2017 and CES 2019.

Conclusions

In conclusion, this assignment aims to predict the overall popular vote for the Liberal party and the Conservative party and the party between these two parties more likely to be the winner of the 2025 Canadian federal election. My hypothesis is: **Based on the outcome of the 2021 federal election [4], I predict that the Conservative party will again yield the highest popular vote in the 2025 Canadian federal election.** and I also assume that **the party with a higher popular vote is the winner**. In order to predict the overall popular vote for each party, I build a logistic multiple regression model for each party and implement the methodology of poststratification to find the estimate of each cell. You can find more details in the section "Methods."

The results from my analysis are surprising and match the results from the last Federal Election and my hypothesis. The overall popular vote for the Liberal and Conservative party is 0.214074 and 0.2570859, meaning that 21.4074% and 25.70859% of voters are voting for the Liberal and Conservative party. And the winner will be the Conservative party, meaning that Justin Trudeau can not win his fourth term as the prime minister of Canada. More surprisingly, the gap between the two parties' popular vote is higher than in the 2021 Canadian federal election [4] (32.62% vs. 33.74% and 21.4074% vs. 25.70859%). However, even though our analysis has successfully predicted the overall popular vote and the winner in the 2025 Federal Election, there is still some weaknesses/limitation that can be improved.

The first thing is that the data sets I chose are not the latest ones: the Canada Election survey was taken in 2019, and the General Social survey was taken in 2017. Therefore the data we have may not reflect how the people think about which party they will vote now, especially since we are now living in a post-pandemic

society. The second thing is that the multiple logistic model we built for the Liberal party is not well-fitted. The p-value of the coefficient of "Male" is not so small and the assumption of linearity seems not satisfied. Therefore the results we calculate from our models may be less convincing. The third thing is that there are one-third observations in CES 2019 do not share their voting preference, and I did not remove these observations as it may be unethical. The impact is that the values of the overall popular vote from our analysis will be smaller than the true values. You can see that the overall popular vote for the two parties from our analysis is around 20%, but popular votes for the two parties in the 2021 Federal election are above 30% [4].

If you want to go further and deeper, such as predicting the 2029 Canadian federal election after the 2025 federal election, there are several ways that can make the prediction more accurate. The first is that you can make your model to be more complicated, meaning that you can include more variables to predict the popular vote for a different party. Besides, it is more appropriate to use the latest data sets (both general survey and Canada election survey data) as the people may change their voting decision after the global pandemic (COVID). For example, some people felt disappointed about the responses to COVID made by the government (i.e., the Liberal Party). They may consider voting for other parties (e.g., the Conservative Party) in the next Federal Election. Lastly, if the model is more complicated, this means there will be more cells in post-stratification. This can make our results more accurate, and the predicted overall popular vote will be more convincing.

To sum up, this assignment predicts the overall popular vote for the Liberal and Conservative party for the 2025 Canada Federal Election. And the results show that the winner will be the Conservative party due to a higher predicted overall popular vote. However, there are still some limitations and weaknesses in my analysis, and it may be enhanced in future work.

Bibliography

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

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

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

4. Wikimedia Foundation. (2022, November 30). 2021 Canadian federal election. Wikipedia. Retrieved December 1, 2022, from https://en.wikipedia.org/wiki/2021_Canadian_federal_election

5. Majority and minority governments. (n.d.). Retrieved December 1, 2022, from https://learn.parl.ca/understanding-comprendre/en/how-parliament-works/majority-and-minority-governments/

6. Welcome to the 2019 Canadian election study. Canadian Election Study. (n.d.). Retrieved December 1, 2022, from http://www.ces-eec.ca/

7. Government of Canada, S. C. (2019, February 6). General Social Survey - Family (GSS). Surveys and statistical programs. Retrieved December 1, 2022, from https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4501#a1

8. Technology, A. K. through. (n.d.). Data Centre. CHASS Data Centre. Retrieved December 1, 2022, from https://datacentre.chass.utoronto.ca/

9. Stringer, A. G. and A. (2021, January 20). Probability, statistics, and data analysis. Chapter 16 Short tutorial on pulling data for Assignment 1. Retrieved December 1, 2022, from https://awstringer1.github.io/sta238-book/section-short-tutorial-on-pulling-data-for-assignment-1.html#section-canadian-election-study

10. Canada, E. (n.d.). Home. – Elections Canada. Retrieved December 1, 2022, from https://www.elections.ca/content.aspx?section=med&dir=c76%2Fcitizen&document=index&lang=e

11. John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

12. 2019 Canadian Election Study - Phone Survey Technical Report.pdf - Harvard Dataverse. (n.d.). Retrieved December 1, 2022, from https://dataverse.harvard.edu/file.xhtml?persistentId=doi%3A10.7910%2FDVN%2F8RHLG1%2F1PBGR3&version=2.0

13. Yihui Xie (2022). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.40.