Informative Title Name

STA304 - Assignment 2

Eric Zhu, Yumeng Wang, Rafee Rahman

November 24, 2022

Abstract

Introduction

The Canadian Constitution requires federal elections to be held at least every five years. By convention, Canada has been holding federal elections every four years, although in some cases they may break this pattern as the governor general can dissolve parliament and call an election [1].

Canada is divided into 338 constituencies, also known as ridings [1]. Canada runs on a political party system where each party has one leader, the candidate running for Prime Minister, and representatives of that party attempting to win over a riding. On election day, registered voters will vote in the polling divisions within their riding, and select one member of parliament representing the party of their choice. The leader of the Canadian party that wins over the majority of ridings (170 of the 338) will become the Prime Minister.

With the next Canadian federal election taking place in 2025, the goal of our report is to predict the overall popular vote using a regression model with post-stratification. Post-stratification is a way to adjust the weighted totals within mutually exclusive cells so that they equal the known population totals [2]. This helps account for differences between two populations, for example, some surveyed groups of people, and the true census population. By post-stratifying, we will obtain better results and make better inferences about the true population.

The most popular vote is determined by the political party that has the highest number of votes, and this outcome will be the party that is most likely to win the election. This might not always be the case, however, because ridings vary in population. The party with the highest number of votes may not always win because another party might have won more ridings with narrow margins [3].

change this to reflect the GAM model description Using logistic regression models and survey data about the party choice of each voter, we will use predictor variables including province, sex, visible minority, income, and age to make our predictions about whether a party is voted for or not. In the post-stratification model, we will weigh the predicted outcome for each cell that holds the variables of interest, by the proportion of the population. This will help adjust the data correctly to make correct inferences for the actual population.

Based on the percentage of popular votes being slightly higher for the Conservatives than the Liberal party during the 2021 and 2019 federal elections [4], **we will hypothesize that the overall popular vote will remain the same and go to the Conservative Party**.

Data

The Canadian Election Study is an annual study of voting trends, demographics, and other preferences that are assumed to be related to Canadian voters' political behavior. The major goals of CES are to give a

comprehensive account of the election, to emphasize the key factors that influence people's voting decisions, to show what changes and remains constant throughout the campaign and from election to election, and to draw attention to the similarities and differences between voting and elections in Canada and other countries. According to the dataset, 37822 people were surveyed and 620 variables were measured. The campaign survey began on September 13, 2019, and finished on October 21, 2019, while the post-election survey started on October 24, 2019, and ended on November 11, 2019.

The population is all Canadian citizens and permanent residents aged 18 or older. The sampling frame is the list of all Canadians and permanent residents older than 18, the framed population is the eligible residents who could access the survey, and the sampled population is the Canadians who saw the survey and completed it. The campaign survey and post-election survey are delivered online and collected by Advanis.Inc and presented on Qualtrics online platform.

Qualtrics gathered a sample from various panels for CES. They use stratified sampling, which divides people into groups according to province/territory and ensures that each region is balanced regarding gender and age.

The General Social Survey (GSS) gathers comprehensive socio-demographic data, including information on age, sex, education, religion, ethnicity, income, etc. The General Social Survey's two main goals are to give information on specific social policy concerns and to collect data on social trends in order to identify changes in Canadians' living conditions and well-being across time. All non-institutionalized individuals and non-residents of First Nation communities who are 15 years of age or older and reside in one of Canada's ten provinces make up the target population for GSS.

Questionnaires are distributed in order to gather data for the census. Similar to CES, a probability sampling-based stratified design provides a framework for the regular sample. At the provincial level, the stratification is conducted out. One household member at least 15 years old is chosen randomly to provide the information.

Data Cleaning & Variables

For our model, we initially started looking for predictor variables by seeing which ones were common between the CES dataset and the GSS dataset. Doing so would allow for the post-stratification of these variables. We found that there were several variables in common between the datasets, which included age, language, birthplace, citizenship status, province, the ownership of a home, income, education, voting choice, marital status, health, and sex. For example, we introduce new variables like "education" and change the value of the original cps19 education to one of the following values: 1 to 4 for "No degree/diploma", 5 for "High School", 6 and 7 for "College/Trade", 8 and 9 for "Bachelor", and 10 to 11 for "Postgrad".

<Summary table of variables.>

Table 1: CES 2019 Summary Statistics

education	language	age_category	income_category	gender	province	health	citizenship	[illegible]	marital_status	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]favour
Bachelor :1981	English :3725	55-64 :1286	$100,000 to $124,999: 714	Female:2527	Ontario :2171	Don't Know : 256	No : 24	Don't know/Prefer not to say: 8	Divorced: 425	No residence :1327	Min. :0.000	Min. : 0.00	Min. : 0.00	Min. : 0.00	Min. : 0	Min. : 0.00	Min. : 0.0
College/Trade :1647	English and French : 653	65-74 :1222	$125,000 and more :1201	Male :2785	Quebec : 845	Excellent : 947	Yes:5288	No : 725	Don't know/Prefer not to answer: 17	Own residence:3985	1st Qu.:0.000	1st Qu.: 17.00	1st Qu.: 10.00	1st Qu.: 23.00	1st Qu.: 0	1st Qu.: 23.50	1st Qu.: 0.0

education	language	age_category	income_category	gender	province	health	citizenship	born_in_canada	marital_status	votes	liberal_favour	conservative_favour	ndp_favour	bloc_favour	green_favour	ppc_favour
High School : 656	French : 644	45-54 : 902	$25,000 to $49,999 : 851		British Columbia: 741	Good/Very Good:2592	Yes :4579		Living with a partner : 707	Median :1.000	Median : 55.00	Median : 41.00	Median : 51.00	Median : 10	Median : 51.00	Median : 15.0
No degree/diploma: 157	Neither English nor French: 290	35-44 : 795	$50,000 to $74,999 :1127		Alberta : 683	Poor/Fair :1517			Married :2761	Mean :1.196	Mean : 48.72	Mean : 44.83	Mean : 48.05	Mean : 20	Mean : 46.43	Mean : 23.3
Postgrad : 871		25-34 : 665	$75,000 to $99,999 :1014		Maintoba : 258				Never Married : 981	3rd Qu.:2.000	3rd Qu.: 78.00	3rd Qu.: 77.00	3rd Qu.: 71.00	3rd Qu.: 31	3rd Qu.: 69.00	3rd Qu.: 40.0
		75-84 : 304	Less than $25,000 : 405		Saskatchewan : 211				Separated : 169	Max. :5.000	Max. :100.00	Max. :100.00	Max. :100.00	Max. :100	Max. :100.00	Max. :100.0
		(Other): 138			(Other) : 403				Widowed : 252							

Table 2: GSS Summary Statistics

education	language	age_category	income_category	gender	province	health	citizenship	born_in_canada	marital_status	owns_house
College/Trade :3539	Both English and French :1698	15-24: 762	$100,000 to $ 124,999: 300	Female:5420	Ontario :2655	Don't Know : 43	No : 477	Don't know/Prefer not to say: 29	Divorced : 981	No residence :2938
High School :2976	Don't know : 3	25-34:1085	$125,000 and more : 356	Male :4646	Quebec :1930	Excellent :3030	Yes:9589	No :1925	Living with a partner: 943	Owns a residence:7128
No degree/diploma:2500	English only :7259	35-44:1318	$25,000 to $49,999 :3165		British Columbia:1228	Good/Very Good:4642		Yes :8112	Married :4310	
Postgrad :1051	French only :1079	45-54:1376	$50,000 to $74,999 :1645		Alberta : 816	Poor/Fair :2351			Never Married :2314	
	Neither English nor French: 27	55-64:2101	$75,000 to $99,999 : 747		New Brunswick : 709				Separated : 374	

education	language	age_category	income_category	gender	province	health	citizenship	born_in_canada	marital_status	owns_house
		65-74:1958	Less than $25,000 :3853		Nova Scotia : 701				Widowed :1144	
		75-84:1466			(Other) :2027					

<Describe cleaning process?>

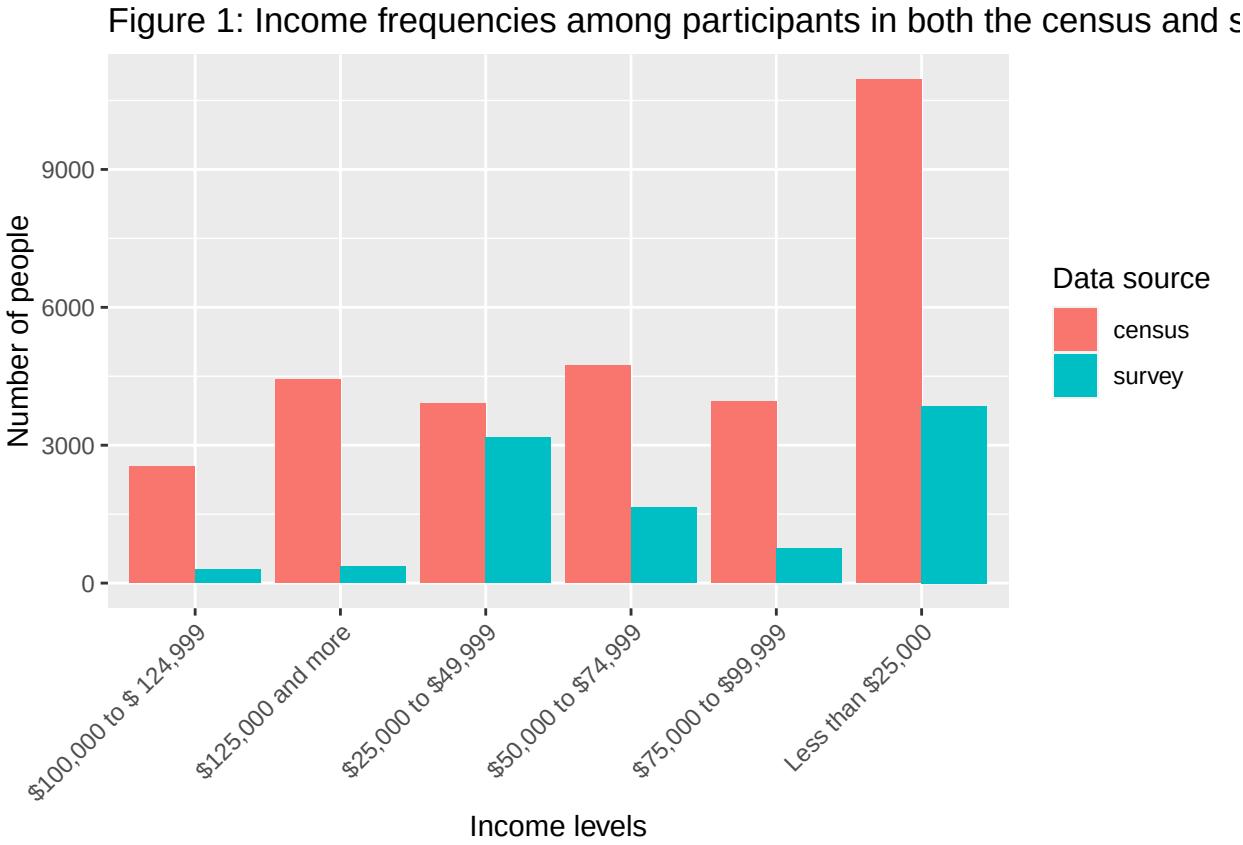

Figure 1 tells us some information about our survey data. We can see that there were more participants in the survey data with income levels of "25,000 to 74,999 and"Less than 25,000" than the rest of the categories. Similarly, in the census data, the frequency of the income level "Less than 25,000" is also much higher than the rest of the categories. This tells us that our data and post-stratification will result with many perspectives of voting opinions that come from this income category. It could be a good idea to use income as a predictor, as these sparsity's between the frequencies of the categories could present an effect on the voting outcomes and post-stratification results.

Justifying predictor variables

As the purpose of our paper is to determine the most popular vote, we wanted to choose variables we thought would have some effect on the voting outcome. To justify which predictor variables would be the best choices, we will explore some of them in more detail.

A survey done prior to the 2020 Federal Election by EKOS Research Associates surveyed people in different groups (n=3006) to measure vote intention by demographics [6]. One of the questions asked was "If a federal

election was held tomorrow, which party would you vote for?". The answers to this question were then analyzed based on several demographics. We will use the demographics that provided the highest variation in voting intention, as these would likely have a strong effect as a predictor for who someone might vote for.

Age

In the study, the liberal party had the highest percentage of people who intended to vote for them. In the age group 35-49, the percentages were 39% [6] for liberals and 34%[6] for conservatives. This was a smaller difference than the other groups, as the percentage of people intending to vote liberal was much higher than people who voted intended to vote conservative for people aged 50 and over. In addition, the second study that mentioned indicated that voter turnout rates were much higher for people over 50 [5]. This could indicate that age does have some influence on peoples voting choices. There are also many societal factors that may push a certain age group to vote for a specific party, such as students voting for a party that would fund their studies.

Additionally, I will do a two group hypothesis test to answer the following question: Does the proportion of people aged 15-24 who intend to vote for the liberal party differ between the people aged 65-74 who intend to vote for the liberal party? I will use the CES 2019 data to do this.

Null hypothesis: There is no difference between the proportion of liberal votes for people aged 15-24 and the proportion of liberal votes for people aged 65-74: $H_0 : p_{young} - p_{old} = 0$

Alternative hypothesis: There is a difference between the proportion of liberal votes for people aged 15-24 and the proportion of liberal votes for people aged 65-74: $H_0 : p_{young} - p_{old} \neq 0$ ($\alpha = 0.05$)

where, p_{young} represents the proportion of liberal votes for people aged 15-24,and p_{old} represents the proportion of liberal votes for people aged 65-74.

I will conduct a z test for the difference between the two proportions. This test assumes z to be approximately normally distributed under the null hypothesis. The test statistic will determine the p-value which will tell us whether or not we should reject the null hypothesis. It follows,

$$Z = \frac{\hat{p}_{old} - \hat{p}_{young}}{\hat{p}(1-\hat{p})(\frac{1}{n_{old}} + \frac{1}{n_{young}})}, \text{where } \hat{p} = \frac{votes_{old} + votes_{young}}{n_{young} + n_{old}}$$

where $votes_{old}$ represents the total votes of people aged 65-74, and $votes_{young}$ represents the total votes of people aged 15-24.

Table 3: Proportion of intended liberal votes by age group from the CES 2019 data

Age range	Number of obersvations	Total who would vote liberal	Proportion
15-24	2378	759	0.319
25-34	5457	1800	0.330
35-44	5349	1821	0.340
45-54	4971	1660	0.334
55-64	5945	2049	0.345
65-74	5033	1856	0.369
75-84	1216	455	0.374
85+	192	74	0.385

From this data, we can see that $\hat{p}_{old} = 0.374$, $\hat{p}_{young} = 0.319$, $n_{young} = 5457$ and $n_{old} = 2378$. A built in function called *prop.test*() was used. This function takes in two vectors, where each vector contains the groups proportion and number of observations.

After conducting a two proportion Z test at a 95% confidence level, there was strong evidence against the null hypotheses ($z = 4.195$, $p = 2.7331632 \times 10^{-5}$). We can interpret that there is a difference between the proportion of liberal votes for people aged 15-24 and the proportion of liberal votes for people aged 65-74. Similarly, using the formula below, the confidence interval was calculated.

$$(\hat{p}_{\text{old}} - \hat{p}_{\text{young}}) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_{old}(1-\hat{p}_{old})}{n_{old}} + \frac{\hat{p}_{young}(1-\hat{p}_{young})}{n_{young}}}$$

The calculated interval was (-0.129, -0.047). We are 95% confident that the true difference of the proportion of liberal votes between the two groups lie in this range. As the interval does not contain 0, it is evident that some difference occurs, and this gives further evidence that age is a good predictor for voting outcomes.

Sex

According to their analysis of voting intention by gender, the percentage of men who would vote liberal was 35% [6], and the percentage of men who would vote conservative was 37% [6]. However, the percentage of women who would vote liberal was 47% [6], and the percentage of women who would vote conservative was 23% [6]. This data indicates some evidence for the gender divide within Canada, as there is only a 2% difference in male voting intentions as opposed to a 24% difference in female voting intentions for the two largest parties in Canada. Hence, we will use the 'sex' variable as one of our predictors as it could possibly have an impact on voting outcomes.

Province

The study also explored the voting intention by province. Although some provinces had similar voting intentions for the two major parties of Canada, conservative, and liberal, there were clear differences with other provinces. For example, in British Columbia, there was an equal percentage of people who had the intention of voting for the liberal and conservative parties. Precisely, 33% [6] of British Colombians intended to vote for the liberal party, and 33% [6] of British Colombians intended to vote for the conservative party. In contrast, Alberta has 55% [6] of individuals intending to vote for the conservative party, and 22% [6] intending to vote for the liberal party. We can see that there is a clear difference in the political landscape between these two provinces, as Alberta highly favors the Conservative party, while British Columbia has no favorite. Overall, every province besides Alberta and Saskatoon favored the liberals, but it was clear that conservatives were favored extremely in these two provinces. Through these results, it is evident that province would be a good variable to look at when it comes to predicting who someone would vote for.

Education

In the study, the voting intentions of individuals were looked at between 3 education groups. These groups were "High school", "College", and "University". The voting intention of high school and college students for the conservative and liberal parties was of similar proportions. However, for universities, there was a clear distinction, as 46% [6] of people in this group said they would vote for the liberal party, while 26% [6] said they would vote for the liberal party. This distinction indicates that education would be a good predictor for voter outcomes.

Income

The study also examines social class in four groups, "Poor", "Working class", "Middle class", and "Upper class". Each class favors the liberal party, however, the middle and upper classes favor the liberals more than the poor and working class. They also have a lower proportion of people who would vote conservative. In the upper class, 44% would vote liberal and 22% would vote conservative [6]. In the working class, 36% would

vote liberal and 32% would vote conservative. This distinction indicates that it could be a good predictor variable.

By examining a subset of the total variables we found in more detail, we have conslusively chosen sex, province, age, education, and income to be random effects in our final model.

Methods

<Include some text introducing the methodology, maybe restating the problem/goal of this analysis.>

Model Specifics

<I will (incorrectly) be using a linear regression model to model the proportion of voters who will vote for Donald Trump. This is a naive model. I will only be using age, which is recorded as a numeric variable, to model the probability of voting for Donald Trump. The simple linear regression model I am using is:>

$$y = \beta_0 + \beta_1 x_{age} + \epsilon$$

<Where y represents the .... β_0 represents....>

Post-Stratification

<In order to estimate the proportion of voters.....>

<To put math/LaTeX inline just use one set of dollar signs. Example: $\hat{y}^{PS}$ >

$$include.your.mathematical.model.here.if.you.have.some.math.to.show$$

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

Results

<Here you present your results. You may want to put them into a well formatted table. Be sure that there is some text describing the results.>

<Note: Alternatively you can use the `knitr::kable` function to create a well formatted table from your code. See here: https://rmarkdown.rstudio.com/lesson-7.html.>

<Remember you can use `r` to use inline R code.>

<Include an explanation/interpretation of the visualizations. Make sure to comment on the appropriateness of the assumptions/results.>

Conclusions

<Here you should give a summary of the Hypotheses, Methods and Results>

<Highlight Key Results.>

<Talk about big picture.>

<Comment on any Weaknesses.>

<End with a concluding paragraph to wrap up the report.>

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)

(Will cite properly at the end)

1 https://www.thecanadianencyclopedia.ca/en/article/electoral-systems 2 https://www.surveypractice.org/article/2809-post-stratification-or-non-response-adjustment 3 https://www.cicnews.com/2021/08/how-canadas-electoral-system-works-0819016.html#gs.j2j5wh 4 https://www.sfu.ca/~aheard/elections/1867-present.html