Informative Title Name

STA304 - Assignment 2

Eric Zhu, Yumeng Wang, Rafee Rahman

November 24, 2022

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 (The Canadian Encyclopedia).

Canada is divided into 338 constituencies, also known as ridings. 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 (Kolenikov, 2016). 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 (Cicnews.com).

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 (Heard), **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 Process

During the cleaning process, our goal was to clean any irrelevant values and extract variables that were in both datasets for our post-stratification analysis. We began with cleaning the CES 2019 Online Survey dataset. We used the documentation (CES Documentation) to figure out what the original variables represented and what the factors for the responses meant. After this process, we had to change variables in the GSS census data as well to ensure that the common variables had the same factors.

After identifying common variables that matched the GSS Census data, we used R functions to mutate the data into categories. One example was changing the original cps19 education data. For education, the values 1 to 11 originally represented different education levels. We used the mutate() function to change these numbers into categories. We changed the numbers 1 to 4 to "No degree/diploma", 5 to "High School", 6 and 7 to "College/Trade", 8 and 9 to "Bachelor", and 10 to 11 to "Postgrad". This process was done for other variables as well, such as provinces, incomes, health, and age. We also did this same process in the GSS data, to ensure that the categories within the common variables of the census were the same. For example, the original GSS data had the following categories: "Less than", "High school", "College", "Bachelor's degree (e.g. B.A., B.Sc., LL.B.)", and "University certificate, diploma or degree above the bach") ~ "Postgrad". We changed these factors to match what we had in the cps19 data by mutating these categories respectively to "No degree/diploma", "High school", "College/Trade", "Bachelor", and "Postgrad". These categorical variables could then be used as random effects to determine who the respondent would vote for, and would also help when creating the post-stratified results.

We also changed these categories into factors by using the following line, mutate_if(is.character, as.factor). This automated the process of turning the variable options into factors in order to do computations. Additionally, many cells within variables originally had NA values. After we selected our variables of interest and mutated the factors respectively, we made sure do use the drop_na() to remove NA rows within the data frame.

Summary of Data and 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. The table below shows all the variables of the CES and GSS data, along with the number of observations for each response option.

Table 1: CES 2019 Summary Statistics

education	language	age_category	income_category	gender	province
Bachelor , n=1981	English , n=3725	55-64 , n=1286	$100,000 to $ 124,999, n= 714	Female, n=2527	Ontario , n=2171
College/Trade , n=1647	English_and_French , n= 653	65-74 , n=1222	$125,000 and more , n=1201	Male , n=2785	Quebec , n= 845
High School , n= 656	French , n= 644	45-54 , n= 902	$25,000 to $49,999 , n= 851		British Columbia, n= 741
No degree/diploma, n= 157	Neither English nor French, n= 290	35-44 , n= 795	$50,000 to $74,999 , n=1127		Alberta , n= 683
Postgrad , n= 871		25-34 , n= 665	$75,000 to $99,999 , n=1014		Maintoba , n= 258
		75-84 , n= 304	Less than $25,000 , n= 405		Saskatchewan , n= 211
		(Other), n= 138			(Other) , n= 403

Table 2: CES 2019 Summary Statistics

health	citizenship	born_in_canada	marital_status	owns_house	vote_choice
Don't Know , n= 256	No , n= 24	Don't know/Prefer not to say, n= 8	Divorced , n= 425	No residence , n=1327	Min. , n=0.000
Excellent , n= 947	Yes, n=5288	No , n= 725	Don't know/Prefer not to answer, n= 17	Own a residence, n=3985	1st Qu., n=0.000
Good/Very Good, n=2592		Yes , n=4579	Living with a partner , n= 707		Median , n=1.000
Poor/Fair , n=1517			Married , n=2761		Mean , n=1.196
			Never Married , n= 981		3rd Qu., n=2.000
			Separated , n= 169		Max. , n=5.000
			Widowed , n= 252		

Table 3: CES 2019 Summary Statistics

health	citizenship	born_in_canada	marital_status	owns_house	vote_choice
Don't Know , n= 256	No , n= 24	Don't know/Prefer not to say, n= 8	Divorced , n= 425	No residence , n=1327	Min. , n=0.000
Excellent , n= 947	Yes, n=5288	No , n= 725	Don't know/Prefer not to answer, n= 17	Own a residence, n=3985	1st Qu., n=0.000
Good/Very Good, n=2592		Yes , n=4579	Living with a partner , n= 707		Median , n=1.000
Poor/Fair , n=1517			Married , n=2761		Mean , n=1.196
			Never Married , n= 981		3rd Qu., n=2.000
			Separated , n= 169		Max. , n=5.000
			Widowed , n= 252		

There are 5312 observations overall after missing data removal, most of the samples are Canadian citizens. The survey received replies from 50 percent males and 50 percent women, as they would have expected. The poll includes 2785 males and 2527 females, with the age groups of 55 to 64 (1286) and 65 to 74 (1222) having the highest proportions, which is 47.21% of the overall sample population, followed by middle-aged people. 1981 individuals—or 37.3% of the total—have bachelor's degrees, followed by 1647 individuals who attended colleges. People who speak English as their first language account for 70% of the sample (3725), while native French speakers account for 644. There are also persons who can speak both English and French (653), as well as a small number of people who speak neither English nor French as their first language. The largest number of participants in the survey was in Ontario, with 2171 people, accounting for 40.9% of the total. Although the income distribution is fairly equal, we still recognize that 10% of the population earns less than $25,000 a year. The samples are primarily older adults and middle-aged adults, therefore their annual income of $25,000 does indicate that they are impoverished. The majority of people (2592) believe they are in very good health, yet 1517 still believe they are in bad health. The limitation would be that people might exaggerate to obtain more compensation or to persuade the government to treat the issue more seriously, which could lead to the health statistics not being accurate. More information about the relationship between these predictors and voting preferences is provided in the following section.

Table 4: GSS Summary Statistics

education	language	age_category	income_category	gender
College/Trade , n=3539	Both English and French , n=1698	15-24, n= 762	$100,000 to $ 124,999, n= 300	Female, n=5420
High School , n=2976	Don't know , n= 3	25-34, n=1085	$125,000 and more , n= 356	Male , n=4646
No degree/diploma, n=2500	English only , n=7259	35-44, n=1318	$25,000 to $49,999 , n=3165	
Postgrad , n=1051	French only , n=1079	45-54, n=1376	$50,000 to $74,999 , n=1645	
	Neither English nor French, n= 27	55-64, n=2101	$75,000 to $99,999 , n= 747	
		65-74, n=1958	Less than $25,000 , n=3853	

education	language	age_category	income_category	gender
		75-84, n=1466		

Table 5: GSS Summary Statistics

province	health	citizenship	born_in_canada	marital_status	owns_house
Ontario , n=2655	Don't Know , n= 43	No , n= 477	Don't know/Prefer not to say, n= 29	Divorced , n= 981	No residence , n=2938
Quebec , n=1930	Excellent , n=3030	Yes, n=9589	No , n=1925	Living with a partner, n= 943	Owns a residence, n=7128
British Columbia, n=1228	Good/Very Good, n=4642		Yes , n=8112	Married , n=4310	
Alberta , n= 816	Poor/Fair , n=2351			Never Married , n=2314	
New Brunswick , n= 709				Separated , n= 374	
Nova Scotia , n= 701				Widowed , n=1144	
(Other) , n=2027					

The cleaned GSS has 10066 observations. The behavior of each predictor, such as gender, education, and wealth, is similar to the General Election Study. It is important to point out that while 4642 people believe they have good health, 3030 believe they have exceptional health, and 2351 believe they have poor health, accounting for 23% of the overall population as contrasted to 30% from the GES, respectively. The General Social Study is less biased than the General Election Study because people are unaware of its objective.

Figure 1: Income frequencies among participants in both the census and survey data.

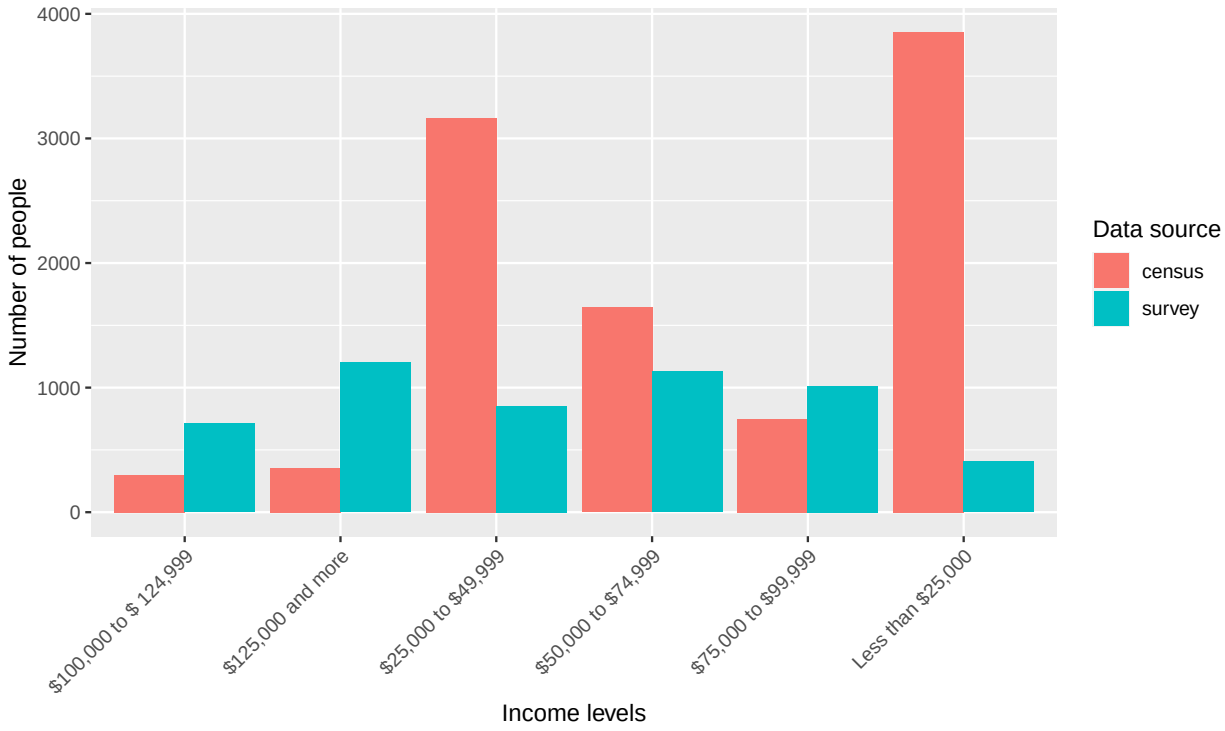

Figure 1 tells us some information about our survey data. We can see that there were more participants in census data with income levels of "25,000 to 74,999 and"Less than 25,000" than the rest of the categories. In contrast, within the survey data, the frequency of the income level "Less than 25,000" is the lowest. We can see that using income as a random effect will more accurately represent the population once it is post-stratified.

Figure 2: Health conditions among participants in both the census and survey data.

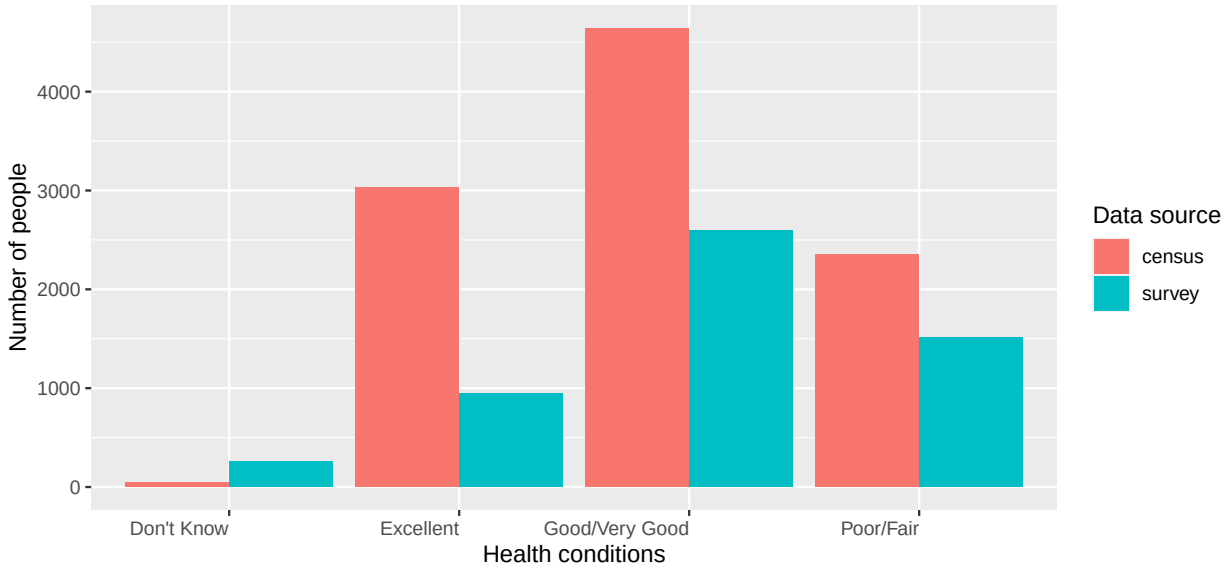

Figure 2 indicates the health conditions of the census and survey participants. It can be seen that in the census data, the highest reported health condition was Good/Very Good. This was also true in the survey data. One difference however is the census data having more reported conditions as Excellent than

Poor/Fair conditions, while the survey data has the opposite. This information tells us more about the sample population, and it will be interest to post-stratify to match the census population weights.

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 (Ekos Politics, 2020). 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% (Ekos Politics, 2020) for liberals and 34% (Ekos Politics, 2020) 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, a second study that mentioned that voter turnout rates were much higher for people over 50 (Sharanjit and Sébastien, 2012). 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 6: 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	95	20	0.211

Age range	Number of obersvations	Total who would vote liberal	Proportion
25-34	665	178	0.268
35-44	795	244	0.307
45-54	902	324	0.359
55-64	1286	439	0.341
65-74	1222	432	0.354
75-84	304	127	0.418
85+	43	16	0.372

From this data, we can see that $\hat{p}_{old} = 0.354$, $\hat{p}_{young} = 0.211$, $n_{young} = 95$ and $n_{old} = 1222$. 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 ($z = 2.715$, p =0.0066179). 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 (0.051, 0.235). 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% (Ekos Politics, 2020), and the percentage of men who would vote conservative was 37% (Ekos Politics, 2020). However, the percentage of women who would vote liberal was 47% (Ekos Politics, 2020), and the percentage of women who would vote conservative was 23% (Ekos Politics, 2020). 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% (Ekos Politics, 2020)of British Colombians intended to vote for the liberal party, and 33% (Ekos Politics, 2020) of British Colombians intended to vote for the conservative party. In contrast, Alberta has 55% (Ekos Politics, 2020) of individuals intending to vote for the conservative party, and 22% (Ekos Politics, 2020) 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% (Ekos Politics, 2020) of people in this group said they would vote for the liberal party, while 26% (Ekos Politics, 2020) 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 (Ekos Politics, 2020). In the working class, 36% would vote liberal and 32% would vote conservative (Ekos Politics, 2020). This distinction indicates that it could be a good predictor variable.

Language

Languages are said to have an influence on one's culture. People have argued that moral identity is a cultural construct (Jia, Krettenauer, 2017) and that it is context-dependent which ties it to different social and cultural obligations. As Canada has two official languages, English and French, the moral identity between these groups of people could have an influence on their voting choice. For example, in Quebec, French is their first language and is used to communicate. In other provinces and regions, English is used. Due to the fact that language has a large influence on culture, there are evidently some cultural differences between French Canada and English Canada (Santa Fe Relocation, 2019). These differences may lead to individuals having different perspectives from a moral standpoint on issues within Canada, and thus could be a great predictor of voting outcomes.

By examining a subset of the total variables we found in more detail, we have conclusively found sex, province, age, education, income, and language to be some of the better predictors to use as 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

@misc{canadian election study, title={Welcome to the 2019 Canadian election study}, url={http://www.ces-eec.ca/}, journal={Canadian Election Study}}

@misc{cicnews.com, url={https://www.cicnews.com/2021/08/how-canadas-electoral-system-works-0819016.html#gs.j2j5wh}, journal={Cicnews.com}}

@misc{ekos politics_2020, title={Update on the political landscape and the issues of race, policing, and the three MS in the Canada-china affair}, url={https://www.ekospolitics.com/index.php/2020/06/update-on-the-political-landscape-and-the-issues-of-race-policing-and-the-three-ms-in-the-canada-china-affair/}, journal={EKOS Politics}, year={2020}, month={Jun}}

@misc{government of canada_2021, title={General Social Survey - Social Identity (SI)}, url={https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5024}, journal={Surveys and statistical programs}, author={Government of Canada, Statistics Canada}, year={2021}, month={Sep}}

@misc{heard, title={Canadian election results by party 1867 to 2021}, url={https://www.sfu.ca/~aheard/elections/1867-present.html}, journal={Canadian Election Results: 1867-2021}, author={Heard, Andrew}}

@misc{kolenikov_2016, title={Post-stratification or non-response adjustment?: Published in survey practice}, url={https://www.surveypractice.org/article/2809-post-stratification-or-non-response-adjustment}, journal={Survey Practice}, author={Kolenikov, Stas}, year={2016}, month={Aug}}

@misc{philippe j. fournier june 28_2020, title={The biggest divide in Canadian politics? men vs. women.}, url={https://www.macleans.ca/politics/ottawa/the-biggest-divide-in-canadian-politics-men-vs-women/}, journal={Macleans.ca}, author={Philippe J. Fournier June 28, 2020}, year={2020}, month={Jun}}

@misc{the canadian encyclopedia, title={Canadian electoral system}, url={https://www.thecanadianencyclopedia.ca/en/article/electoral-systems}, journal={The Canadian Encyclopedia}}