New possibility to the 2019 Canadian Federal Election

The Liberal Party of Canada is still leading

Zhipeng Zhou

Dcember 23, 2020

Github repo: https://github.com/Ryanzzp/STA304-Final.git

Abstract

The 2019 Canadian Federal Election had passed more than one year. Although Justin Trudeau's Liberals are still the largest party, they have lost their majority (Clarke & Levett, 2019). In this report, we build a multilevel regression with post-stratification (MRP) model based on the Canadian Election Study (Stephenson et al., 2020) and Canadian General Social Survey (Technology, CHASS) to identify how the 2019 Canadian Federal Election would have been different if all eligible voters had voted. By applying the MRP model, we find the victory of Justin Trudeau in 2019 is not just luck. If every eligible voter participated into the election, the Liberal party of Canada maybe still hold a majority of the seats in the House of Commons.

Keywords

Canada, federal election, public opinion, Liberal, Conservative, social survey, multilevel regression with post-stratification

Introduction

Prime Minister Justin Trudeau won a second term in Canada's national election in 2019, but it seems the nation is more divided. In 2015 Federal Election, Trudeau's Liberals had 184 seats in the 338-seat House of Commons with 39.47% popular vote, which made Trudeau win a majority government. However, Trudeau's Liberals lost their majority. In 2019 Federal Election, the Liberals had 157 seats and 33.12% popular vote, 14 short of the 170 needed for a majority. And the Conservatives had 121 seats with 34.34% this time which significantly narrowed the seats gap to the Liberals (Cecco, 2019). What's more, there was 65.95% of eligible voters getting to the polls (CBC news, 2019). Although it was not as high as in 2015 when Justin Trudeau first swept yo power, it was still a strong turnout. But not every eligible voters participated the election makes us wonder what may happen if everyone gets to the poll, especially the gap between the Liberal and the conservative already getting narrowed.

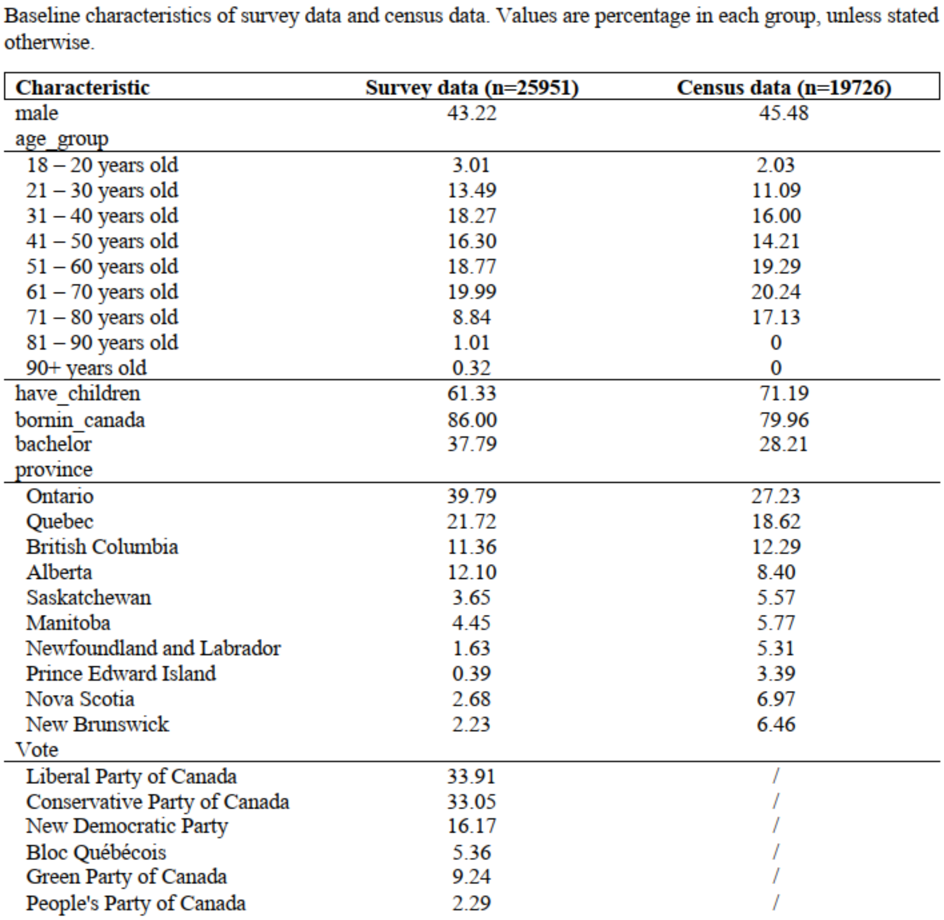
In order to verify our conjecture, we will apply the multilevel regression with post-stratification model (MRP model) based on some datasets. MRP model is a statistical technique which can be used to correct the estimate differences between sample population and target population (Little, 1993). Since it's now impossible to ask everyone to vote in 2019 Federal Election, a MRP model based on a collected survey data for 2019 Federal Election and a Canadian general social survey for the nation can be used to estimate the result if everyone voted in 2019 Federal Election.

Two data sets will be used to build the MRP model. In the Methodology section, I describe the source and the characteristics of both data can explain why I use these data sets. And the model that was used to estimate the popularity for different parties also mentioned in the Methodology section. Result of the popularity for different parties are provided in the Result section, and the inferences of these data along with conclusions are presented in Conclusion Section. And, finally, in the Discussion section, the weakness for this report and the potential further analysis are provided.

Methodology

Data

To build the MRP model, we collected 2019 Canadian Election Study (CES) online survey data set to be the survey data and 2017 Canadian General Social Survey (GSS) data set to be the census data. In this case, the target population for our report is all eligible voters in 2019 Canada Federal Election. The sample frame is CES and GSS. And the sample population are all the people who participated in the relative CES and GSS survey. For the datasets, There are 37822 observations with 620 variables in the original CES data set, where two-wave panel was composed, rolling-cross section during the campaign and a post-election wave. In order to make the data meet the reality and easy, we first cleaned the data. Since the model is used to estimate the popularity for different parties, the post-election follow-up survey data is removed from the survey data set. All variables with over 50% missing value rate and observations with missing value are also removed to reduce the influence of missing values. And all ages were grouped into certain age group, 18-20, 21-30, 31-40, and so on, which can be easily discussed in post-stratification data. There are 10 Canadian provinces, with three territories to the north. According to the 2019 CES Codebook (Stephenson et al., 2020), the territories is too sparse to be representative. And to match the variables in census data, all observations from territories are also removed. Eventually, the best subsets regression was applied to select the predictor variables. Then 25951 observations were selected to the survey data (Figure 1). Similar process was taken to clean the census data. There are 20602 observations with 461 variables in the original GSS dataset. To match the variables in survey, most veriables were removed and only leave 6 variables. To reduce the influence of noisy, only the observations without missing values were kept. Eventually, there are 19726 observations were selected to the census data (Figure 1).



Baseline characteristics of survey data and census data. Values are percentage in each group, unless stated otherwise.

Characteristic	**Survey data (n=25951)**	**Census data (n=19726)**
male	43.22	45.48
age_group		
18 – 20 years old	3.01	2.03
21 – 30 years old	13.49	11.09
31 – 40 years old	18.27	16.00
41 – 50 years old	16.30	14.21
51 – 60 years old	18.77	19.29
61 – 70 years old	19.99	20.24
71 – 80 years old	8.84	17.13
81 – 90 years old	1.01	0
90+ years old	0.32	0
have_children	61.33	71.19
bornin_canada	86.00	79.96
bachelor	37.79	28.21
province		
Ontario	39.79	27.23
Quebec	21.72	18.62
British Columbia	11.36	12.29
Alberta	12.10	8.40
Saskatchewan	3.65	5.57
Manitoba	4.45	5.77
Newfoundland and Labrador	1.63	5.31
Prince Edward Island	0.39	3.39
Nova Scotia	2.68	6.97
New Brunswick	2.23	6.46
Vote		
Liberal Party of Canada	33.91	/
Conservative Party of Canada	33.05	/
New Democratic Party	16.17	/
Bloc Québécois	5.36	/
Green Party of Canada	9.24	/
People's Party of Canada	2.29	/

Figure 1: Baseline characteristics of the data

male tells whether the voter is male or not. *have_children* tells whether the voter has at least one child or not. *bornin_canada* represents whether the voter was born in Canada or not. And the variable *bachelor* represents whether the voter gained a bachelor degree or not. According to the baseline characteristic of datasets, there is no significant odd value, except the census datasets have no voter voter 80 years old. And the vote result of popularity for different parties is roughly the same as the real result for the 2019 Canadian Federal Election. Therefore, these cleaned data can be used to build the target multilevel regression with post-stratification model.

Model

In order to estimate the popularity for different parties in 2019 Canadian Federal Election if every eligible voters participated the election. We will build a multilevel regression with post-stratification model. More formally, applying MRP in our setting comprises two steps. First, we fit six logistic multilevel regression models for six different parties to obtain estimates for sparse post-stratification cells; second, we average over the cells, weighting the values by a measure of forecasted voter turnout, to get the overall estimates.The post-stratification technique can reduce the bias of our models. Since the population of voter for 2019 Canadian Federal Election are huge and various, applying the model to the whole population can lead to great bias. The post-stratification technique divide the whole sample population into several small groups and each group has their own feature. And the bias for each group can be smaller than consider them as a whole. Then we add the result for each group up, weight them by size, and eventually calculate the average value for the whole population. In this case, we are not summarize the whole population, but also take feature of different groups into consideration.

Logistic multilevel regression models

According to the 2019 Canada's electoral map (Bogart & Tahirali, 2019), we can find that the popularity for different parties vary from province to province, which means the region may have influence on the popularity. Therefore, we use multilevel regression model to be the first step of MRP model, since the multilevel model estimates group effects simultaneously with the effects of group-level predictors. And there are 6 different parties in the survey data. Then we apply logistic multilevel regression model for each part. The reason why we use logistic regression model is logistic regression is a classification algorithm, which works well when the target variable is categorical in nature. The predictor variables are selected by the best subsets regression. According to the best subsets regression, the voters whose age are less than 30 performs a really different liking from those who older than 30. Therefore, the variable *age_under_*30 was chosen rather than the age_group. There are 5 fixed variables and 1 random variable are selected. The random variable is *province*, and the fixed variables are *male*, *bachelor*, *bornin_canada*, *have_children*, and *age_under_*30.

Before providing the models, the variables are listed below:

β_j - the random intercept term for party j (L: Liberal; C: Conservation; G: Green; N: NDP; P: People; B: Bloc)

x_1 - whether the voter is male (male is 1; otherwise 0)

x_2 - whether the voter has bachelor degree (has bachelor degree is 1; otherwise 0)

x_3 - whether the voter was born in Canada (Born in Canada is 1; otherwise 0)

x_4 - whether the voter has children (has children is 1; otherwise 0)

x_5 - whether the voter's age under 30 (under 30 is 1; otherwise 0)

r_j - the slope for the province for party j

W_j - the value of the province (i.e. which province a respondent is from)

1. Liberal Party of Canada

$$\text{Level 1: } \ln \frac{y_L}{1-y_L} = \beta_L - 0.11 \cdot x_{1L} + 0.4 \cdot x_{2L} - 0.22 \cdot x_{3L} - 0.16 \cdot x_{4L} - 0.26 \cdot x_{5L}$$

$$\text{Level 2: } \beta_L = -0.47 + r_L \cdot W_L$$

Province	r_L
Alberta	-0.8291288
British Columbia	-0.1186256
Manitoba	-0.1457643
New Brunswick	0.2340905
Newfoundland and Labrador	0.6747400
Nova Scotia	0.5278237
Ontario	0.2199252
Prince Edward Island	0.3581645
Quebec	0.1278303
Saskatchewan	-1.0401537

Table 1: Liberal Party of Canada r_L

2. Conservative Party of Canada

$$\text{Level 1: } \ln \frac{y_C}{1-y_C} = \beta_C + 0.44 \cdot x_{1C} - 0.31 \cdot x_{2C} - 0.16 \cdot x_{3C} + 0.43 \cdot x_{4C} - 0.38 \cdot x_{5C}$$

$$\text{Level 2: } \beta_C = -0.9 + r_C \cdot W_C$$

Province	r_C
Alberta	1.31043257
British Columbia	-0.06830830
Manitoba	0.41824894
New Brunswick	-0.21091179
Newfoundland and Labrador	-0.42695585
Nova Scotia	-0.53714849
Ontario	-0.04128941
Prince Edward Island	-0.54229862
Quebec	-0.81343647
Saskatchewan	0.93138399

Table 2: Conservative Party of Canada r_C

3. New Democratic Party

$$\text{Level 1: } \ln \frac{y_N}{1-y_N} = \beta_N - 0.5 \cdot x_{1N} - 0.09 \cdot x_{2N} + 0.23 \cdot x_{3N} - 0.32 \cdot x_{4N} + 0.64 \cdot x_{5N}$$

$$\text{Level 2: } \beta_N = -1.61 + r_N \cdot W_N$$

Province	r_N
Alberta	-0.3186241787
British Columbia	0.3693909802
Manitoba	0.1560634782
New Brunswick	-0.5889873340
Newfoundland and Labrador	0.4214687688
Nova Scotia	0.0009313613
Ontario	0.1887576876
Prince Edward Island	-0.2768936347
Quebec	-0.3848501246
Saskatchewan	0.4624638708

Table 3: New Democratic Party r_N

4. Bloc Québécois

$$\text{Level 1: } \ln \frac{y_B}{1-y_B} = \beta_B + 0.16 \cdot x_{1B} - 0.23 \cdot x_{2B} + 1.52 \cdot x_{3B} + 0.2 \cdot x_{4B} - 0.8 \cdot x_{5B}$$

$$\text{Level 2: } \beta_B = -0.19 + r_B \cdot W_B$$

Province	r_B
Alberta	-0.0219023194
British Columbia	-0.0196470434
Manitoba	-0.0082983470
New Brunswick	-0.0042866926
Newfoundland and Labrador	-0.0032765739
Nova Scotia	-0.0051456824
Ontario	-0.0624436787
Prince Edward Island	-0.0007730896
Quebec	16.2667868918
Saskatchewan	-0.0070773316

Table 4: Bloc Québécois r_B

5. Green Party of Canada

$$\text{Level 1: } \ln\frac{y_G}{1-y_G} = \beta_G - 0.24 \cdot x_{1G} + 0.01 \cdot x_{2G} + 0.21 \cdot x_{3G} - 0.2 \cdot x_{4G} + 0.34 \cdot x_{5G}$$

$$\text{Level 2: } \beta_G = -2.27 + r_G \cdot W_G$$

Province	r_G
Alberta	-0.78569700
British Columbia	0.49892879
Manitoba	-0.11239534
New Brunswick	0.89195713
Newfoundland and Labrador	-1.00609565
Nova Scotia	0.40116536
Ontario	-0.08606451
Prince Edward Island	1.07617928
Quebec	-0.18937805
Saskatchewan	-0.63005516

Table 5: Green Party of Canada r_G

6. People's Party of Canada

$$\text{Level 1: } \ln\frac{y_P}{1-y_P} = \beta_P + 0.45 \cdot x_{1P} - 0.44 \cdot x_{2P} + 0.3 \cdot x_{3P} + 0.06 \cdot x_{4P} + 0.36 \cdot x_{5P}$$

$$\text{Level 2: } \beta_P = -4.2 + r_P \cdot W_P$$

Province	r_P
Alberta	0.010269356
British Columbia	-0.009049538
Manitoba	-0.026610316
New Brunswick	0.019881967
Newfoundland and Labrador	-0.012853393
Nova Scotia	0.004512183
Ontario	0.034394384
Prince Edward Island	-0.006484948
Quebec	-0.020670903
Saskatchewan	0.007771218

Table 6: People's Party of Canada r_P

Post-stratification

The post-stratification averages over the cells, weighting the values by a measure of forecasted voter turnout, to get the overall estimates. So to perform the post-stratification, we first need to generate the cells. For the census dataset we have, we have 6 variables. So we generate the cells by considering all possible combinations of *male* (2 categories), *province* (10 categories), *bachelor* (2 categories), *bornin_canada* (2 categories), *have_children* (2 categories), and *age_under*_30 (2 categories). And since the size of each cell can be really small, which makes that cell meaningless, we only keep cells with size over 10 observation. Therefore, there are 195 cells left.

The post-stratification estimate is defined by

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

Where $\hat{y_j}$ is the estimate in each cell. And N_j is the population size of the j^{th} cell. Then we calculate the overall estimates for all 6 different parties by plug the logistic multilevel regression models into the cells we generated above.

Results

Eventually, we get the estimations of the popularity for 6 different parties. We estimate that the proportion of voters in favour of voting for Liberal Party of Canada to be 0.341, the proportion of voters in favour of voting for Conservative Party of Canada to be 0.339, the proportion of voters in favour of voting for New Democratic Party to be 0.15, the proportion of voters in favour of voting for Bloc Québécois to be 0.046, the proportion of voters in favour of voting for Green Party of Canada to be 0.098, and the proportion of voters in favour of voting for People's Party of Canada to be 0.023. And all these estimates are based off our post-stratification analysis of the proportion of voters in favour of their own party modeled by a multilevel logistic regression model, which accounted for the variable list, *male* (2 categories), *province*, *bachelor*, *bornin_canada*, *have_children*, and *age_under_*30.

Liberal	Conservative	NDP	Bloc	Green	People
0.341	0.339	0.15	0.046	0.098	0.023

Table 7: Estimation result

According to Figure 2, we can find that the estimation for the popularity for different parties is almost the same as survey baseline data and the real election result. For the Liberal Party, Conservative Party, and New Democratic Party result, the error between our estimation and the reality is less than 1%. However, the error in Bloc Québécois, Green Party, and People's Party are slightly larger, but still within 3%.

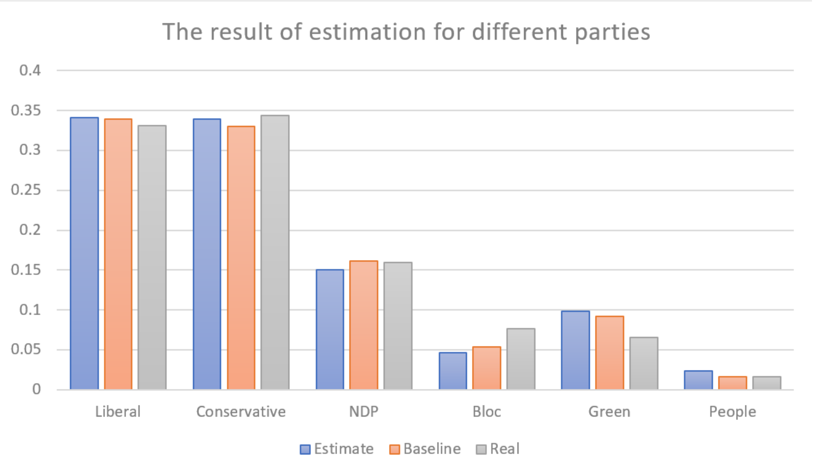


Figure 2: The result of estimations for different parties

Discussion

Summary

In this report, we gathered data from 2019 Canadian Election Study - online survey and 2017 Canadian General Social Survey. The cleaned dataset of CES was used to fit 6 different multilevel logistic regression models for 6 different major Canadian parties. And the GSS data formed the census data, which was used to apply the post-stratification model. And eventually, the multilevel regression with post-stratification model based on the Canadian Election Study and Canadian General Social Survey datasets helps us to estimate the popularity for different parties when every eligible voter participated into the election.

Conclusions

According the Figure 2, we find that the participation of all eligible voter in 2019 Canadian Federal Election doesn't make too much difference from the original result. In 2019 Canadian Federal Election, Trudeau's Liberals won the election with a lower vote share than the Conservative Party of Canada. However, the estimation we get in Result section tells a different story. Although the vote shares are still quite close between the Liberal Party and the Conservative Party, the Liberal Party has more vote share. Which mean if all eligible voters participated the election, Trudeau may repeat his tremendous victor in 2015.

The mathematical notation of model can always tell some interest things. As we know, the coefficient before the variable can decide the correlation between the independent variable and the dependent variable. The more far away from 0, the larger the correlation. Compare with other coefficients, the coefficient before the value of Quebec in the logistic multilevel regression model for Bloc Québécois quite stands out, whose value is around 16, with other coefficients' value are only around ±1. The estimation confirms the suspicion. According to Figure 3, we can find that the vote share rate is extremely high in Quebec province. The reason why it happened may because its Quebec nationalism policy.

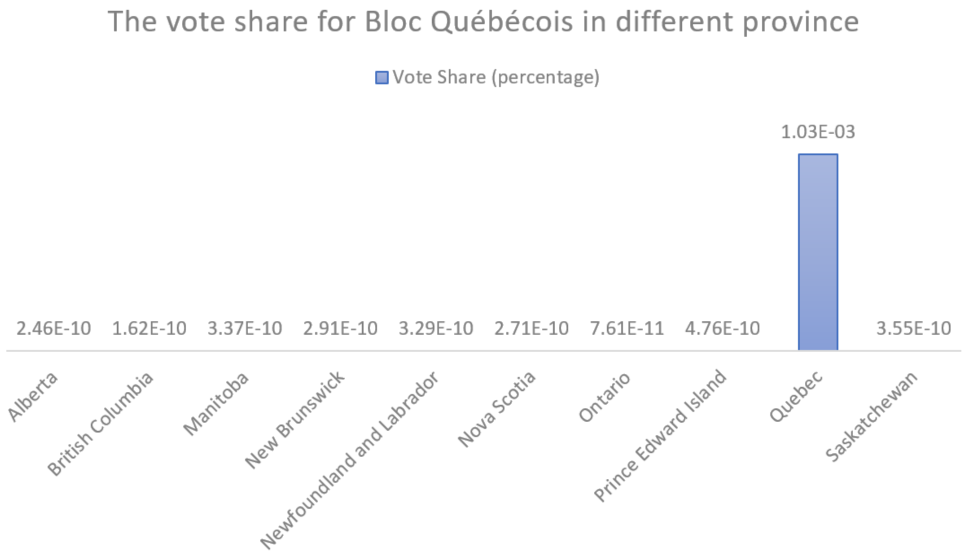


Figure 3: The vote share for Bloc Québécois in different province

Weakness & Next Steps

Although the result for the estimations are roughly performed as expected, there are make weakness can be corrected. The source data has disappointed in many ways.

1. According to Figure 1, we know that the size of census dataset is smaller than the survey data. In MRP model, we divide the census data into different cells, use the fitted model in each cell, and eventually average over the cells. In general, the larger the census data population and the number of cells, the better the effect of the post-stratification model. In the GSS census data, we can only split it into 195 cells which is quite little. So the effect of post-stratification model is also not really good.

   Solution (Next Step): Try to find other census datasets with larger size

2. There are too many missing values in the CES and GSS datasets, which may lead to a large error in the estimation. In general, there are only two ways to deal with the missing value, remove it directly or try to make it up. If the size of datasize was not large, removing it directly can make the sample size insufficient. And making it up always introduce new errors.

   Solution (Next Step): Try to find other datasets with less missing value

3. Missing age groups. According to Figure, we can find that there is no observations over 80 years old, which is impossible. According to the population estimates from Statistics Canada(Government of Canada, 2020), there are over 1,500,000 people over 80 years old, which counts 4% of the overall population. Therefore, the observations with over 80 years old should not be ignored.

   Solution (Next Step): Try to find other census datasets with complete age group

4. Although the territories is too sparse to be representative, the eligible voters in territories still belongs to the target population. So we should also not ignore them.

5. Except the data source, the way how we deal with the missing value is also not good. In this report, all observations with missing values were removed directly, which can be one of the reason why we don't have sufficient data in census dataset. And removing the observations directly can also lead to the missing of important variables.

   Solution (Next Step): Instead of removing the observations with missing value directly, try to make them up. For example, replace the missing data by mean substitution or multiple Imputation.(Kang H, 2013)

6. In this report, to reduce the analysis time, we only used several important variables to build model. And the variables are selected by the best subsets regression, which can also lead to a large error.

   Solution (Next Step): We can try some more and some other variables to build the model. And we can also divide the census data in different ways to get moreproper cells.

References

1. Clarke S, Levett C. Canada election 2019: full results [Internet]. The Guardian. Guardian News and Media; 2019 [cited 2020Dec23]. Available from: https://www.theguardian.com/world/2019/oct/22/canada-election-2019-full-results

2. Stephenson LB, Harell A, Rubenson D, Loewen PJ. 2019 Canadian Election Study - Online Survey [Internet]. Harvard Dataverse. Harvard Dataverse; 2020 [cited 2020Dec23]. Available from: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FDUS88V

3. Technology AKthrough. [Internet]. Computing in the Humanities and Social Sciences. [cited 2020Dec23]. Available from: http://www.chass.utoronto.ca/

4. Cecco L. Canada elections: Trudeau wins narrow victory to form minority government [Internet]. The Guardian. Guardian News and Media; 2019 [cited 2020Dec23]. Available from: https://www.theguardian.com/world/2019/oct/22/canada-elections-justin-trudeau-wins-narrow-victory-to-form-minority-government

5. Canadian election drew nearly 66% of registered voters | CBC News [Internet]. CBCnews. CBC/Radio Canada; 2019 [cited 2020Dec23]. Available from: https://www.cbc.ca/news/canada/voter-turnout-2019-1.5330207

6. Little, R. J. (1993). Post-stratification: a modeler's perspective. Journal of the American Statistical Association, 88, 1001–1012.

7. Bogart N, Tahirali J. How Canada's electoral map changed after the vote [Internet]. Federal Election 2019. CTV News; 2019 [cited 2020Dec23]. Available from: https://election.ctvnews.ca/how-canada-s-electoral-map-changed-after-the-vote-1.4652484

8. Government of Canada, Statistics Canada. [Internet]. Population estimates on July 1st, by age and sex. Government of Canada, Statistics Canada; 2020 [cited 2020Dec23]. Available from: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000501

9. Kang H. The prevention and handling of the missing data. Korean J Anesthesiol. 2013;64(5):402-406. doi:10.4097/kjae.2013.64.5.402