Prediction of 2019 Canadian Federal Election Result Useing Logistic Regression with Post-Stratification

Minhui Yu

December 21st 2020

Prediction of 2019 Canadian Federal Election Result Useing Logistic Regression with Post-Stratification

Name Author: Minhui Yu

Date: December 21st 2020

GitHub Repo

Code and data supporting this analysis is available at: https://github.com/mawymaggie/STA304PS3

I. Abstract

The study examined if the result of 2019 Canada Federal Election will change when "everyone" had voted. We first clean survey data and census data, then build logistic model and use post-stratification. We obtained estimated vote percentage of each party and found that the result of the election will not change if "everyone" had voted. The conclusion is even though some voter abstained, their behavior won't influence overall perspective of the result, but the specific percentage will fluctuate.

II. Key Words

Logistic Model, Post-Stratification, 2019 Canadian Federal Election, Liberal Party and predication.

III. Introduction

Statistics is widely used in various disciplines, and with the accelerating process of digitization, people increasingly hope to be able to summarize some empirical laws from a large amount of data to provide some basis for subsequent decision-making. Logistic regression with post-stratification is one of the method to achieve this aim. And it is used for correcting model estimates for known differences between a sample population (the population of the data you have), and a target population (a population you would like to estimate for). Therefore use statistics to build our project is appropriate.

Elections and voting are very important to most countries, and Canada is no exception. In our project, we will build logistic model and use post-stratification to identify how the 2019 Canadian Federal Election would have been different if 'everyone' had voted. Here we have two kinds of data, one is survey data and another one is census data. Survey data is from Canadian Election Survey and census data is from 2017

General Social Survey (GSS). And we assume that there is no significant change in population from 2016 to 2019.

We obtain the vote percentage of Liberal Party is 33.098%, the vote percentage of Conservative Party is 33.895%, the vote percentage of New Democratic Party is 14.991%, the vote percentage of Bloc Québécois is 5.078%, the vote percentage of Green Party is 9.942% and the vote percentage of People's Party is 2.411%. We found that these percentages makes the result of election remain the same, but have difference in each of them. The conclusion is there is a certain percentage of voters with the same background, so those who abstain will not affect the final result.

The first part is Abstract, the second part is keywords, and the third part is introduction. I will describe the two data in part 4, the model made and the post-stratification in part 5. The result part (part 6) will give the explanation of the table and plot in this report. Discussion and limitation will be show in part 7. And the last part is reference.

IV. Data

This study has two data set, one is survey data from the 2019 Canadian Election Study (http://www.ces-eec.ca/) and download it via the cesRR package (https://hodgettsp.github.io/cesR/), and another one is 2017 Canadian General Social Survey (GSS).

(1) Survey Data: There are 620 variable in the raw data, and I choose:

- cps19_votechoice: the choice of voter; fist drop the answer with no idea or NA, then I use this variable create 6 column stands for 6 party, e.g. column vote_liberal gives whether voter will vote for liberal party, if yes the value is 1, otherwise is 0; I choose this variable because we need the vote choice as for our study, and this variable is the important one in our study; the reason I didn't choose other variables is they cannot give the specific outcome like this;
- cps19_citizenship: the citizenship of the voter; this variable can filter the voter who don't have the right to vote for the election, such as if the voter's citizenship is permanent resident, then this voter can not vote for the election, so we need drop these voter;
- cps19_province: the province the voter in; since Canada has 10 province and voter lives in different province have different life, so they may have different thought on which party is appropriate for Canada; here I drop "Yukon" since census data doesn't contain it;
- cps19_education: the education level of the voter; with different education voter may has different policy in fiance, tax and so on, and each party has their own focus on their policies such as fiance, immigration, ...; difference in education level may influence voter perfer differnet policy;
- cps19_gender: the gender of the voter; Canada is becoming more and more respectful of feminism, so I am interested in whether gender will influence the choice of voter; here I drop "both";
- cps19_yob: the birth year of the voter; I create a new column called age_group with the birth year of voter and divided them in different age group; the reason I choose this variable is I think the length of the experience may affects voter's choices;

(2) Census Data: There are 81 variables in the raw data, and I choose age, sex, education and province which are correspond to the choices we made in survey data. After cleaning, I create a column n, which summarize the number of voter has same education level, same sex, live in same province, and in same age group. Thus there are only 603 rows in the final census data.

- Here are the example of what survey data and census data look like:

Table 1: Example of the Survey Data

votechoice	vote_liberal	vote_conservative	vote_ndp	vote_bq	vote_green	vote_people	sex	age_group	education	province
Green Party	0	0	0	0	1	0	Female	20 to 39	above bachelor	Quebec
Liberal Party	1	0	0	0	0	0	Female	Under 20	less than university	Ontario
Conservative Party	0	1	0	0	0	0	Male	20 to 39	less than university	Ontario
Liberal Party	1	0	0	0	0	0	Female	20 to 39	less than university	Ontario
Liberal Party	1	0	0	0	0	0	Female	20 to 39	less than university	Ontario
Liberal Party	1	0	0	0	0	0	Male	Under 20	high school	Ontario

Table 2: Example of the Census Data

age_group	sex	education	province	n
20 to 39	Female	above bachelor	Alberta	29
20 to 39	Female	above bachelor	British Columbia	44
20 to 39	Female	above bachelor	Manitoba	11
20 to 39	Female	above bachelor	New Brunswick	15
20 to 39	Female	above bachelor	Newfoundland and Labrador	16
20 to 39	Female	above bachelor	Nova Scotia	19

V. Model / Methodology

There are six main party (Liberal Party, Conservative Party, New Democratic Party, Bloc Québécois, Green Party and People's Party) and some other party (which we don't talk about them in our study, since the sum of their voted percentage is lower than 1%) participated in the 2019 Canadian Federal Election. We are interested in the vote percentage of the six party, so here we build six model for six party. We implement logistic regression model on the proportion of voters who will vote for each party using RStudio. There are four predictors in each model: age group, province, sex, and education. Since the response variable is binary, it is appropriate to adapt a logistic regression model.

- Here are boxplots for Voter's Choice and different explanatory variables in survey data.

Figure 1 – Bar Plot for Voter's Choice and Age Group

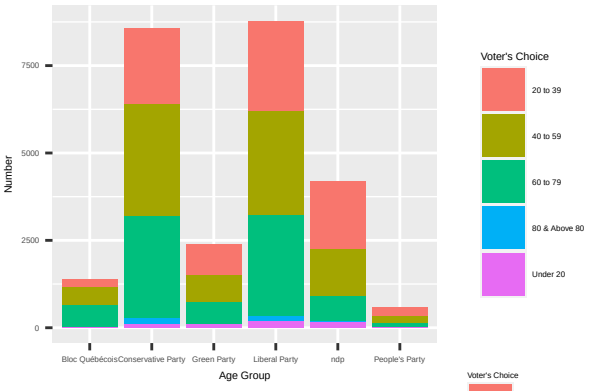


Figure 2 – Bar Plot for Voter's Choice and Sex

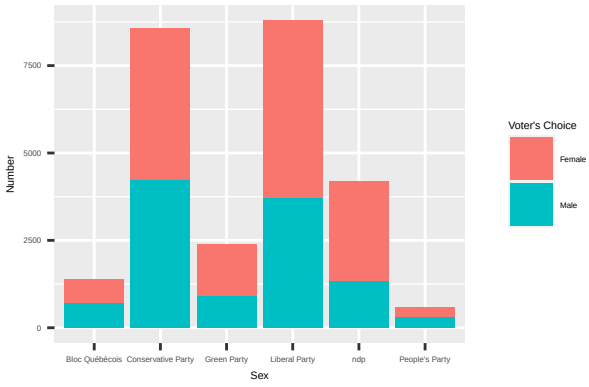


Figure 3 – Bar Plot for Voter's Choice and Province

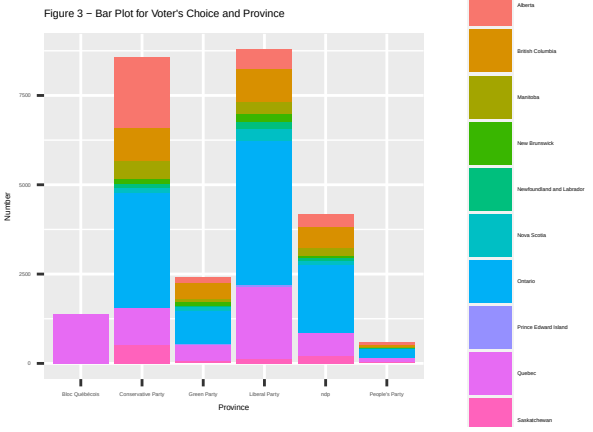
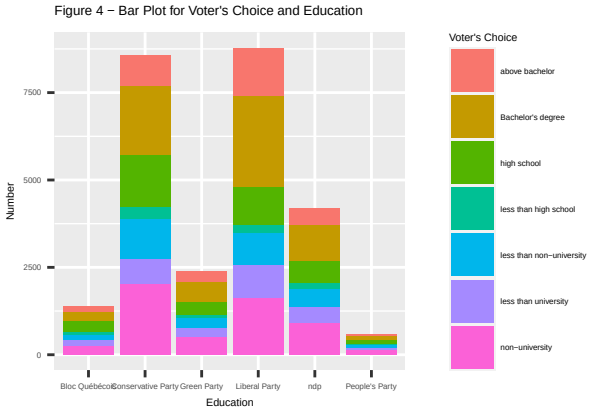


Figure 4 – Bar Plot for Voter's Choice and Education



The number of people of different age groups participating in voting varies greatly, and the choices they make are also different. We divide age into five age groups because the range of age is large and there is only one data has numerical age. By grouping age, it's easier for us to visualize the relationship among age groups. Since sex has two type – male and female (we drop both in data part since there is no both in census data), then it is categorical variable. We choose sex because from the bar plot above we found that same party obtain different proportion in male and female.The province voter living is different, they will have different choice, such as, almost all voter in Quebec vote for Bloc Québécois, however, just few voter in Ontario vote for Bloc Québécois. Education is an important variable, it influence the vote percentage. For example, from the bar plot, bachelor more likely to vote Liberal Party. These four predictor are all categorical.

- The estimated formula of logistic regression for Liberal Party's model is $log(\hat{p}^l/(1-p^l)) = \hat{\beta}_0^l + \hat{\beta}_1^l x_1 + \hat{\beta}_2^l x_2 + ... + \hat{\beta}_5^l x_5 + \hat{\beta}_6^l x_6 + ... + \hat{\beta}_{11}^l x_{11} + \hat{\beta}_{12}^l x_{12} + ... + \hat{\beta}_{20}^l x_{20}$, where $\hat{p}^l$ represents the probability that voter vote for Liberal Party.
- The estimated formula of logistic regression for Conservative Party's model is $log(\hat{p}^c/(1-\hat{p}^c)) = \hat{\beta}_0^c + \hat{\beta}_1^c x_1 + \hat{\beta}_2^c x_2 + ... + \hat{\beta}_5^c x_5 + \hat{\beta}_6^c x_6 + ... + \hat{\beta}_{11}^c x_{11} + \hat{\beta}_{12}^c x_{12} + ... + \hat{\beta}_{20}^c x_{20}$, where $\hat{p}^c$ represents the probability that voter vote for Conservative Party.
- The estimated formula of logistic regression for New Democratic Party's model is $log(\hat{p}^{nd}/(1-\hat{p}^{nd})) = \hat{\beta}_0^{nd} + \hat{\beta}_1^{nd} x_1 + \hat{\beta}_2^{nd} x_2 + ... + \hat{\beta}_5^{nd} x_5 + \hat{\beta}_6^{nd} x_6 + ... + \hat{\beta}_{11}^{nd} x_{11} + \hat{\beta}_{12}^{nd} x_{12} + ... + \hat{\beta}_{20}^{nd} x_{20}$, where $\hat{p}^{nd}$ represents the probability that voter vote for New Democratic Party.
- The estimated formula of logistic regression for Bloc Québécois's model is $log(\hat{p}^b/(1-\hat{p}^b)) = \hat{\beta}_0^b + \hat{\beta}_1^b x_1 + \hat{\beta}_2^b x_2 + ... + \hat{\beta}_5^b x_5 + \hat{\beta}_6^b x_6 + ... + \hat{\beta}_{11}^b x_{11} + \hat{\beta}_{12}^b x_{12} + ... + \hat{\beta}_{20}^b x_{20}$, where $\hat{p}^b$ represents the probability that voter vote for Bloc Québécois Party.
- The estimated formula of logistic regression for Green Party's model is $log(\hat{p}^g/(1-\hat{p}^g)) = \hat{\beta}_0^g + \hat{\beta}_1^g x_1 + \hat{\beta}_2^g x_2 + ... + \hat{\beta}_5^g x_5 + \hat{\beta}_6^g x_6 + ... + \hat{\beta}_{11}^g x_{11} + \hat{\beta}_{12}^g x_{12} + ... + \hat{\beta}_{20}^g x_{20}$, where $\hat{p}^g$ represents the probability that voter vote for Green Party.
- The estimated formula of logistic regression for People's Party's model is $log(\hat{p}^r/(1-\hat{p}^r)) = \hat{\beta}_0^r + \hat{\beta}_1^r x_1 +$

$\hat{\beta}_2^r x_2 + ... + \hat{\beta}_5^r x_5 + \hat{\beta}_6^r x_6 + ... + \hat{\beta}_{11}^r x_{11} + \hat{\beta}_{12}^r x_{12} + ... + \hat{\beta}_{20}^r x_{20}$, where $\hat{p}^r$ represents the probability that voter vote for People's Party.

In above arguments, x_1 represent sex is male, x_2 to x_5 represent different age groups, x_6 to x_{11} represent voters' different educational attainment, and x_{12} to x_{20} represent different province in the Canada. They are all dummy variable for example $x_1 = 1$ if the gender of voter is female, and $x_1 = 0$ if the voter is male. Similar for x_2 to x_{20} In addition, when the voter is in age group 20-39, the education is above bachelor, lives in Alberta and is a female, the odds of this voter vote for Liberal Party is $exp(\hat{\beta}_0^l)$. Take the estimated coefficient $\hat{\beta}_2^l$ as an example, if the voter is a female, education is above bachelor, lives in Alberta but she is in age group 40 - 59, odd of this voter vote for Liberal Party will be change to $exp(\hat{\beta}_2^l)$. Other $\hat{\beta}_i^j$ for i={1, 2, ..., 20} and j = {l,c,nd,b,g,r} have the similar interpretation. The logistic model is practical because the data are extracted from national survey and census, which are real life data and we want the probability. The four predictors we selected provide sufficient information on voters' demographics, building a logistic regression model on these predictors would give a relatively valid result.

After these, we use post-stratification which is one of the stratification method in which a sample is first taken from a population using simple random sampling. Then the cells in the sample are stratified according to some characteristics. It can be applied when the population information is incomplete. And in general, its estimation efficiency is better than simple random sampling. Since we want to predict the vote percentage for each party, we start the post-stratification. Our choice of cells are based on different age group, sex, education and province. We generate the cells by considering all possible combinations of sex (2 categories), age (5 categories), education (7 categories) and province (10 categories), thus partitioning the data into 700 cells. However there are only 603 cells actually, this is because some combination do not exist such as female voter who lives in Alberta and in age group 80 & above 80 don't have education background above Bachelor. We choose such cells because our model above contained all of the explanatory variable in cells. The main formula in this part is $\hat{y}^{ps} = (\sum N_j * \hat{y}_j)/\sum N_j$. The process of post-stratification is first calculate the estimate of each cell. Since our model is a logistic model, we can not get estimate directly, for instance, we get $a_i = log(estimate_i)$. Then use $estimate_i = e^{a_i}/(1 + e^{a_i})$ get the real estimate. Now we can multiply n_i (the size of this cell) by estimate, and do this for all cells and add these solution to our census_data.csv as a new column. In order to estimate the proportion of voters in cells, we will sum the value of $estimate_i * n_i$ and divide this sum by the entire population size. And we do this process for the six model we made above and then get the vote percentage for each party.

VI. Result

- Here are the baseline characteristics of survey data:

(1) Age Group: 20-39
(2) Education: above Bachelor
(3) Province: Alberta
(4) Sex: Female

- Here are the post-stratification estimates:

$$\hat{y}_l^{ps} = 0.3310$$
$$\hat{y}_c^{ps} = 0.3390$$
$$\hat{y}_{nd}^{ps} = 0.1499$$
$$\hat{y}_b^{ps} = 0.5078$$
$$\hat{y}_g^{ps} = 0.9942$$
$$\hat{y}_r^{ps} = 0.2411$$

Table 3: Summary of Vote Liberal Party Logistic Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.27156	0.06234	-20.39644	0.00000
sexMale	-0.11861	0.02767	-4.28743	0.00002
age_group40 to 59	0.08078	0.03365	2.40036	0.01638
age_group60 to 79	0.26748	0.03483	7.68007	0.00000
age_group80 & Above 80	0.28377	0.10647	2.66528	0.00769
age_groupUnder 20	0.25865	0.09899	2.61281	0.00898
educationBachelor's degree	-0.06558	0.04413	-1.48602	0.13727
educationhigh school	-0.62813	0.05137	-12.22685	0.00000
educationless than high school	-0.76018	0.08228	-9.23909	0.00000
educationless than non-university	-0.49010	0.05382	-9.10660	0.00000
educationless than university	-0.25088	0.05532	-4.53465	0.00001
educationnon-university	-0.49213	0.04699	-10.47246	0.00000
provinceBritish Columbia	0.71018	0.06189	11.47509	0.00000
provinceManitoba	0.67494	0.08030	8.40493	0.00000
provinceNew Brunswick	1.06356	0.09857	10.79026	0.00000
provinceNewfoundland and Labrador	1.50596	0.10904	13.81073	0.00000
provinceNova Scotia	1.34741	0.09015	14.94577	0.00000
provinceOntario	1.06373	0.05130	20.73626	0.00000
provincePrince Edward Island	1.18727	0.21268	5.58247	0.00000
provinceQuebec	0.95142	0.05485	17.34653	0.00000
provinceSaskatchewan	-0.26217	0.10482	-2.50103	0.01238

Base on the process in post-stratification, we estimate that the percentage of voter who would vote for Liberal Party ($\hat{y}_l^{ps}$) to be 33.10%. We estimate that the percentage of voter who would vote for Conservative Party ($\hat{y}_c^{ps}$) to be 33.90%. We estimate that the percentage of voter who would vote for New Democratic Party ($\hat{y}_{nd}^{ps}$) to be 14.99%. We estimate that the percentage of voter who would vote for Bloc Québécois ($\hat{y}_b^{ps}$) to be 5.07%. We estimate that the percentage of voter who would vote for Green Party ($\hat{y}_g^{ps}$) to be 9.94%. We estimate that the percentage of voter who would vote for People's Party ($\hat{y}_r^{ps}$) to be 2.41%. These result is from our post-stratification analysis of the vote percentage for each party by a logistic model, which accounted for age group, sex, province and education.

- Here is the table of our model outcomes:

For the six table below shows the coefficients of our model. The first column is the name of variable, the second column is the value of $\hat{\beta}_i^j$ where i=$\{0, 1, \ldots, 20\}$, the third column is the standard error of estimates and j = {l,c,nd,b,g,r} and the last column is the p-value (if p-value is smaller than 0.05 we will reject $H_0 = 0$, and the t test is significant).

When the voter is in age group 20-39, the education is above bachelor, lives in Alberta and is a female, the odds of this voter vote for Liberal Party is $exp(-1.27156)$. Take the estimated coefficient $\hat{\beta}_2^l$ as an example, if the voter is a female, education is above bachelor, lives in Alberta but she is in age group 40 - 59, odd of this voter vote for Liberal Party will be change to $exp(0.08078)$. Other $\hat{\beta}_i^j$ for i=$\{1, 2, \ldots, 20\}$ and j = {l,c,nd,b,g,r} have the similar interpretation.

- The estimated formula of logistic regression for Liberal Party's model is $log(\hat{p}^l/(1-\hat{p}^l)) = \hat{\beta}_0^l + \hat{\beta}_1^l x_1 + \hat{\beta}_2^l x_2 + ... + \hat{\beta}_5^l x_5 + \hat{\beta}_6^l x_6 + ... + \hat{\beta}_{11}^l x_{11} + \hat{\beta}_{12}^l x_{12} + ... + \hat{\beta}_{20}^l x_{20}$, where $\hat{p}^l$ represents the probability that voter vote for Liberal Party.

Table 4: Summary of Vote Conservative Party Logistic Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.22040	0.05928	-3.71821	0.00020
sexMale	0.42079	0.02864	14.69181	0.00000
age_group40 to 59	0.42838	0.03546	12.08198	0.00000
age_group60 to 79	0.47585	0.03672	12.95826	0.00000
age_group80 & Above 80	0.67098	0.10738	6.24872	0.00000
age_groupUnder 20	-0.63300	0.12218	-5.18107	0.00000
educationBachelor's degree	0.18606	0.04999	3.72179	0.00020
educationhigh school	0.54448	0.05438	10.01270	0.00000
educationless than high school	0.42003	0.08092	5.19059	0.00000
educationless than non-university	0.46596	0.05734	8.12578	0.00000
educationless than university	0.12077	0.06199	1.94818	0.05139
educationnon-university	0.46751	0.05100	9.16605	0.00000
provinceBritish Columbia	-1.39236	0.05538	-25.14072	0.00000
provinceManitoba	-0.92145	0.07173	-12.84614	0.00000
provinceNew Brunswick	-1.52356	0.10115	-15.06247	0.00000
provinceNewfoundland and Labrador	-1.74145	0.12119	-14.36975	0.00000
provinceNova Scotia	-1.88301	0.10076	-18.68797	0.00000
provinceOntario	-1.36715	0.04362	-31.34311	0.00000
provincePrince Edward Island	-1.91017	0.25210	-7.57690	0.00000
provinceQuebec	-2.15262	0.05183	-41.52970	0.00000
provinceSaskatchewan	-0.36748	0.07650	-4.80336	0.00000

Table 5: Summary of Vote New Democratic Party Logistic Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.48171	0.07736	-19.15376	0.00000
sexMale	-0.44582	0.03695	-12.06457	0.00000
age_group40 to 59	-0.56431	0.04018	-14.04496	0.00000
age_group60 to 79	-1.06017	0.04763	-22.25947	0.00000
age_group80 & Above 80	-1.34247	0.18971	-7.07635	0.00000
age_groupUnder 20	0.08966	0.10491	0.85458	0.39278
educationBachelor's degree	-0.03037	0.06125	-0.49591	0.61996
educationhigh school	0.09976	0.06798	1.46747	0.14225
educationless than high school	0.24712	0.09652	2.56019	0.01046
educationless than non-university	0.10271	0.07144	1.43770	0.15052
educationless than university	0.19450	0.07322	2.65654	0.00789
educationnon-university	0.12249	0.06289	1.94775	0.05144
provinceBritish Columbia	0.75289	0.07287	10.33248	0.00000
provinceManitoba	0.54594	0.09610	5.68068	0.00000
provinceNew Brunswick	-0.36593	0.16079	-2.27582	0.02286
provinceNewfoundland and Labrador	0.78570	0.13181	5.96094	0.00000
provinceNova Scotia	0.38276	0.12000	3.18964	0.00142
provinceOntario	0.54388	0.06154	8.83826	0.00000
provincePrince Edward Island	-0.15872	0.34215	-0.46389	0.64272
provinceQuebec	-0.01508	0.07056	-0.21367	0.83080
provinceSaskatchewan	0.81057	0.09663	8.38808	0.00000

Table 6: Summary of Vote Bloc Québécois Logistic Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-22.40496	514.11493	-0.04358	0.96524
sexMale	0.07353	0.06436	1.14239	0.25329
age_group40 to 59	0.66206	0.08876	7.45900	0.00000
age_group60 to 79	1.07985	0.08887	12.15091	0.00000
age_group80 & Above 80	0.83651	0.28232	2.96299	0.00305
age_groupUnder 20	-0.84993	0.40161	-2.11630	0.03432
educationBachelor's degree	0.02479	0.11531	0.21495	0.82980
educationhigh school	0.44220	0.11621	3.80510	0.00014
educationless than high school	0.77415	0.16792	4.61019	0.00000
educationless than non-university	0.10063	0.13221	0.76114	0.44658
educationless than university	0.17096	0.13267	1.28857	0.19755
educationnon-university	0.16944	0.11886	1.42554	0.15400
provinceBritish Columbia	-0.05940	739.26011	-0.00008	0.99994
provinceManitoba	-0.05159	995.12305	-0.00005	0.99996
provinceNew Brunswick	0.00535	1309.24544	0.00000	1.00000
provinceNewfoundland and Labrador	0.00674	1499.51140	0.00000	1.00000
provinceNova Scotia	-0.05009	1210.47236	-0.00004	0.99997
provinceOntario	0.03442	587.11868	0.00006	0.99995
provincePrince Edward Island	-0.05726	2946.98288	-0.00002	0.99998
provinceQuebec	20.41677	514.11492	0.03971	0.96832
provinceSaskatchewan	-0.02083	1068.43529	-0.00002	0.99998

Table 7: Summary of Vote Green Party Logistic Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.60830	0.10654	-24.48226	0.00000
sexMale	-0.22988	0.04544	-5.05844	0.00000
age_group40 to 59	-0.30149	0.05253	-5.73917	0.00000
age_group60 to 79	-0.40961	0.05643	-7.25863	0.00000
age_group80 & Above 80	-0.58838	0.20051	-2.93449	0.00334
age_groupUnder 20	0.55815	0.12469	4.47631	0.00001
educationBachelor's degree	-0.18928	0.07297	-2.59386	0.00949
educationhigh school	-0.20394	0.08229	-2.47830	0.01320
educationless than high school	-0.17363	0.12512	-1.38762	0.16525
educationless than non-university	-0.06874	0.08507	-0.80805	0.41906
educationless than university	-0.09307	0.08902	-1.04550	0.29579
educationnon-university	-0.12406	0.07553	-1.64254	0.10048
provinceBritish Columbia	1.32505	0.09967	13.29499	0.00000
provinceManitoba	0.71729	0.13469	5.32546	0.00000
provinceNew Brunswick	1.72520	0.13343	12.92936	0.00000
provinceNewfoundland and Labrador	-0.33714	0.28484	-1.18364	0.23656
provinceNova Scotia	1.24330	0.13866	8.96624	0.00000
provinceOntario	0.72203	0.09169	7.87486	0.00000
provincePrince Edward Island	2.07188	0.24306	8.52419	0.00000
provinceQuebec	0.62956	0.09818	6.41228	0.00000
provinceSaskatchewan	0.17723	0.16627	1.06588	0.28648

Table 8: Summary of Vote People's Party Logistic Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.84711	0.18206	-21.13087	0.00000
sexMale	0.50849	0.08492	5.98785	0.00000
age_group40 to 59	-0.29635	0.09535	-3.10793	0.00188
age_group60 to 79	-0.94114	0.11814	-7.96655	0.00000
age_group80 & Above 80	-0.64903	0.36442	-1.78099	0.07491
age_groupUnder 20	-0.72197	0.32932	-2.19232	0.02836
educationBachelor's degree	-0.20391	0.16124	-1.26461	0.20601
educationhigh school	0.47303	0.16074	2.94275	0.00325
educationless than high school	0.88981	0.19970	4.45582	0.00001
educationless than non-university	0.40247	0.17011	2.36592	0.01799
educationless than university	0.21543	0.18613	1.15741	0.24710
educationnon-university	0.30189	0.15519	1.94531	0.05174
provinceBritish Columbia	-0.10701	0.17544	-0.60996	0.54189
provinceManitoba	-0.51669	0.27793	-1.85904	0.06302
provinceNew Brunswick	0.44483	0.25199	1.76529	0.07751
provinceNewfoundland and Labrador	-0.70753	0.46557	-1.51969	0.12859
provinceNova Scotia	0.12067	0.27276	0.44239	0.65820
provinceOntario	0.03811	0.13366	0.28512	0.77555
provincePrince Edward Island	-12.81793	239.10970	-0.05361	0.95725
provinceQuebec	-0.07581	0.14837	-0.51097	0.60938
provinceSaskatchewan	0.10759	0.23450	0.45881	0.64637

VII. Discussion

A. Summary

The purpose of our study is to predict if the result of 2019 Canadian Federal Election will change if "everyone" had voted. We find survey data from the 2019 Canadian Election Study (http://www.ces-eec.ca/) and download it via the cesRR package (https://hodgettsp.github.io/cesR/), and our census data is 2017 Canadian General Social Survey (GSS). After cleaning these two data set, we use survey data to build our logistic model for the six party and do post-stratification base on census data. Now we obtain the estimated vote percentage for each party if "everyone" had voted in the 2019 Canadian Federal Election.

B. Conclusion

For each of the predictive variables we used in our prediction, trends in each characteristic of voters can be clearly seen. Most voter who are under 20 will choose to vote for Liberal Party, most female voter will vote for Liberal Party, most voter live in Quebuc will vote for Bloc Québécois and most voter whose education background is non-university will vote for Conservaative part. We obtain the estimated vote percentage of Liberal Party is 33.10%, the estimated vote percentage of Conservative Party is 33.90%, the estimated vote percentage of New Democratic Party is 14.99%, the estimated vote percentage of Bloc Québécois is 5.08%, the estimated vote percentage of Green Party is 9.94% and the estimated vote percentage of People's Party is 2.41%.

I found the real value on Wikipedia – 33.1% for Liberal Party, 34.3% for Conservative Party, 16.0 for New Democratic Party, 7.6% for Bloc Québécois, 6.5% for Green and 0.4% for other party. Here we can conclude that whether "everyone" voted, Conservative Party has the highest vote percentage. However, this percentage is lower than 50% and is very close to the vote percentage for Liberal Party. Base on our results, if "everyone" had voted, the difference between Conservative Party and Liberal Party will decrease, which

means the vote percentage of Conservative Party will decrease. Maybe most of those who did not vote for the 2019 Canadian federal election will vote for Liberal Party. The vote percentage of Bloc Québécois and New Democratic will decrease, and the vote percentage of Green and People's Party will increase. If we sum these six percentage, we can find there is still 0.58% remains, which means 0.58% will vote for other parties.

C. Weakness & Next Steps

Here in our study, when we clean data, we fold some gourps such as education background is Phd. These would decrease the accuracy of our prediction in the post-stratification. From table 2-7, we observed that some p-value is larger than 0.05 (can not reject $H_0 = 0$), which means they are not good for our model. So next we may use AIC / BIC forward or backward Elimination to pick the appropriate predictor. Another point is our census data has many variable which seems can be predictor. However, our survey data has few variables, and some variables we can not use in our study. So if possible, we can find more data to build our model. Our model in this study is logistic model, since some parameter may vary at more than one level after we add more predictor, so we can try Multilevel Regression model.

VIII. References

1. Datasets

Beaupre, Pascale. *General Social Survey Cycle 31: Families, 2020.* Statistics Canada Minister of Industry [distributor]. Web. April 2020.

Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, "2019 Canadian Election Study - Online Survey", https://doi.org/10.7910/DVN/DUS88V, Harvard Dataverse, V1 Documentation for the 2019 CES Online Survey can be accessed from here: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DUS88V

2. Software

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

3. Packages

Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.0. https://CRAN.R-project.org/package=broom

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

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.

Hao Zhu, kableExtra: Construct Complex Table with 'kable' and Pipe Syntax https://cran.r-project.org/web/packages/kableExtra/kableExtra.pdf

4. Websites

How to change font size of table in Rmarkdown, LaTeX and .pdf? https://stackoverflow.com/questions/44490209/how-to-change-font-size-of-table-in-rmarkdown-latex-and-pdf

Results of the 2019 Canadian federal election https://en.wikipedia.org/wiki/Results_of_the_2019_Canadian_federal_election