Justin Trudeau would still win 2019 Canadian Federal Election even if "everyone" had voted*

Wen Wang

22 December, 2020

Abstract

This paper investigated whether Justin Trudeau or Andrew Scheer would win the 2019 Canadian Federal Election if "everyone" had voted by using MRP technique based on GSS and CES datasets. The result is that Justin Trudeau would still win even if everyone had voted. The analysis is important since it might encourage citizens to participate in the election period to choose the governing party they want, especially when there is only a small difference in the win rate of two parties, which might result in a tremendous change in the future of a country.

keywords: 2019 Canadian Federal Election, Multilevel Regression Model, Post-Stratification, Logistic Regression Model, everyone voted

1. Introduction

The voting process is crucial since it would determine the Prime Minister/President of a country. However, this is not true that everyone eligible to vote had voted in the elections and this phenomenon would probably affect the final results. According to an article in The Economist, if "everyone" had voted in the 2016 US presidential election, Hillary Clinton would probably be president.[The Economist]. Based on the information on Statistics Canada, just over three-quarters (77%) of Canadians reported voting in the 2019 federal election [Statistics Canada], so it could be inferred that Andrew Scheer would probably win the election if "everyone" had voted.

So, how to make "everyone" vote? It is time-consuming and cost-consuming and almost impossible to do a survey among the target population, all Canadian citizens who are above 18 in this case since not everyone would like to participate in the survey. One statistical way to solve this is through multilevel regression and post-stratification (MRP). The central idea of MRP is to partition the data into thousands of demographic cells, estimate voter intent at the cell level using a multilevel regression model, and finally aggregate the cell-level estimates in accordance with the target population's demographic composition.[et al., 2015]. In this report, I will use MRP to identify whether Andrew Scheer or Justin Trudeau would win the 2019 Canadian federal election if "everyone" had voted.

The procedure is as follows. In the Methodology section (Section 2), two data sets (survey data and census data) will be introduced to investigate how MRP could be used to predict whether Andrew Scheer or Justin Trudeau would win the 2019 Canadian federal election if "everyone" had voted. Results of the MRP analysis are provided in the Results section (Section 3), and summary and conclusion of the MPR analysis along with Weaknesses and Next Steps to improve the analysis are presented in Conclusion section (Section 4).

*Code and data are available at: https://github.com/wangw218/STA304_FinalProject

2. Methodology

2.1 Data

The survey data is from the 2019 Canadian Election Study - Online Survey [Stephenson et al., 2020a]. This online sample was for Canadian Election Study and was composed of a two-wave panel with a modified rolling-cross section during the campaign period and a post-election recontact wave. We only focus on the campaign period survey (CPS) in this report. The CPS contains an online sample of 37,822 members of the Canadian general population through Qualtrics, with targets stratified by region and balanced on gender and age within each region. The target population is all Canadian citizens or permanent residents who are age 18 or over. The frame population is Canadian citizens or permanent residents who are age 18 or over and saw the survey. The sample population is Canadian citizens or permanent residents who are age 18 or over and filled in the survey.[Stephenson et al., 2020b]

The census data is from the General social survey on Family (cycle 31), 2017 [Faculty of Arts & Sciences, b]. The GSS program, established in 1985, conducts telephone surveys across the ten provinces. The GSS is recognized for its regular collection of cross-sectional data that allows for trend analysis, and its capacity to test and develop new concepts that address current or emerging issues. For the census data survey, the target population is all people in Canada who are above 15 years old, excluding those from Yukon, Northwest Territories, and Nunavut. Full-time residents of institutions are also excluded. The frame population is people who are above 15 years old and could be contacted by telephones. The sampled population is people who are above 15 years old and took the survey by telephone. Each record in the survey frame was assigned to a stratum within its province. A simple random sample without replacement of records was next performed in each stratum. There are 20,602 respondents and 461 variables in the census data.[Faculty of Arts & Sciences, a]

Figure 1 shows the number of supporters of each party in survey Data. From the plot, we could see the Liberal Party and the Conservative Party have the most number of supporters while the number of supporters of the Liberal Party is slightly larger than the Conservative Party's. There are a lot of people who don't know or prefer not to answer which party they would like to vote and many NA data.

After comparing and analyzing the survey data and census data, I found age, household income, sex, Province are the common and meaningful variables between these two datasets in investing the result of whether Andrew Scheer or Justin Trudeau would win the 2019 Canadian federal election if "everyone" had voted.

I did some data cleaning to match the variables' names and values between the two original survey data and census data files. For both of the cleaned census data and survey data, they contain the following variables:

"age_group" with range ("45 to 59","60+", "Age 18 to 29", "Age 30 to 44"); "sex" with range (Female, Male); "household_income" with range ("$25,000 to $49,999", "$75,000 to $99,999", "$100,000 to $124,999", $50,000 to $74,999","Less than $25,000" , "$125,000 and more");

"province" with range ("Quebec","Manitoba","Ontario","Alberta","Nova Scotia","British Columbia", "Saskatchewan", "Prince Edward Island", "New Brunswick", "Newfoundland and Labrador") for census data and three more territories: Yukon, Northwest Territories and Nunavut, for survey data. The distribution of the province about survey and census data can be found in Appendix A.

In the survey data, there are two more variables, which are called "voteLib" and "voteCon". Both of them are binary variables, the value 1 indicating the respondent would like to vote for the party. The values of these two variables are gotten based on the value of "cps19_votechoice" variable in the raw census dataset. If the value of cps19_votechoice is "Liberal Party", then "voteLib" is 1, otherwise, it is 0; Similarly, if the value of cps19_votechoice is "Conservative Party", then "voteCon" is 1, otherwise, it is 0.

The overview of our cleaned datasets after removing the NA rows is shown in survey_data Table 1 for survey data and Table 2 for census data.

The key feature of this study is it has a very large survey data size. After cleaning the useless and NA data, it still has 21513 samples. However, our cleaned census data size is only 20178, which is slightly lower than our survey data size. This means our census data is not that good. This is reasonable since our census data

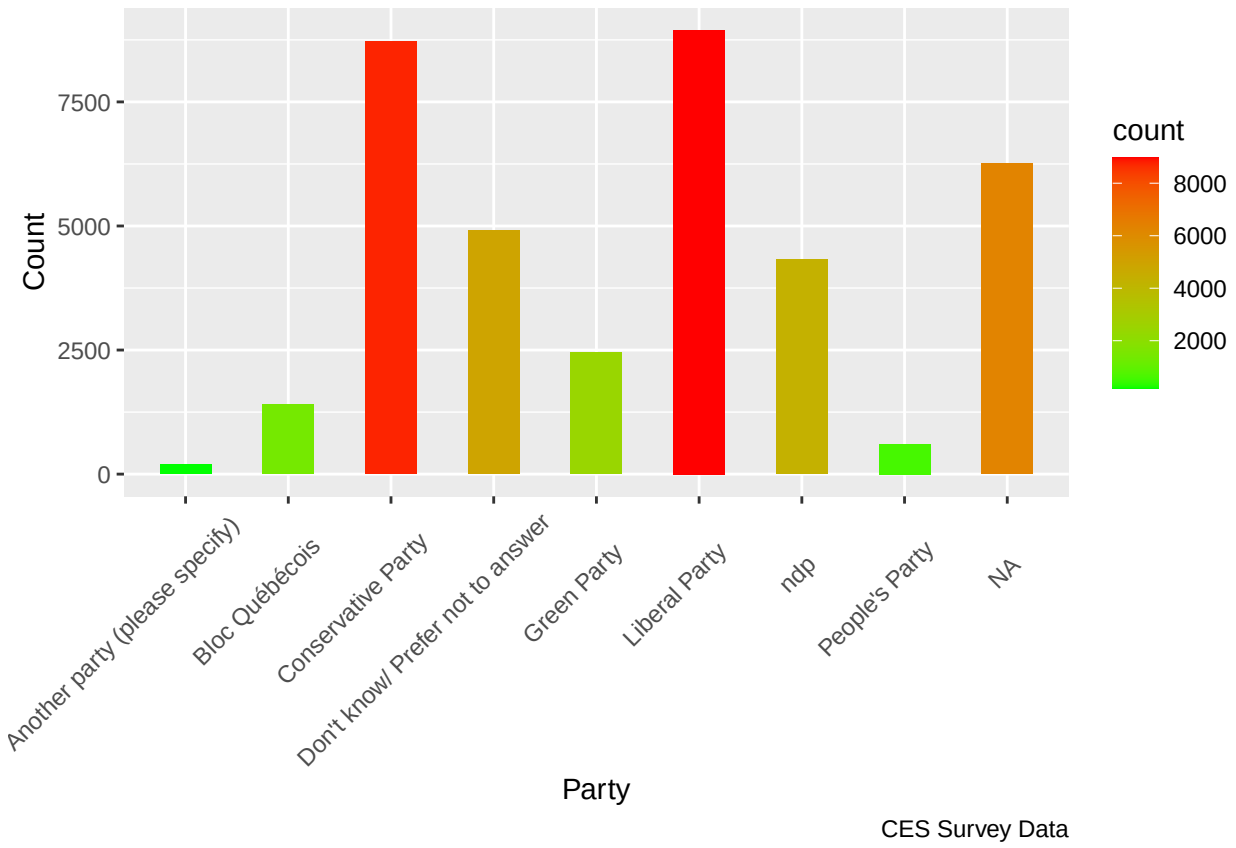

Figure 1: The number of supporters of each Party in survey Data

Table 1: Overview of Survey Data

age_group	sex	household_income	province	voteLib	voteCon
Age 18 to 29	Female	$50,000 to $74,999	Ontario	0	0
Age 18 to 29	Female	$25,000 to $49,999	Ontario	1	0
Age 18 to 29	Male	Less than $25,000	Ontario	1	0
Age 18 to 29	Female	Less than $25,000	Ontario	0	0
Age 18 to 29	Male	$50,000 to $74,999	Ontario	0	1
Age 18 to 29	Male	$125,000 and more	Ontario	1	0

Table 2: Overview of Census Data

age_group	sex	household_income	province
45 to 59	Female	$25,000 to $49,999	Quebec
45 to 59	Male	$75,000 to $99,999	Manitoba
60+	Female	$75,000 to $99,999	Ontario
60+	Female	$100,000 to $ 124,999	Alberta
Age 18 to 29	Male	$50,000 to $74,999	Quebec
60+	Female	$50,000 to $74,999	Quebec

was collected in 2017 and our survey data was collected in 2019 and after 2 years, the population size in Canada would increase. The strength of this study is that our survey data has a very large size, which should be very representative to predict the outcome of the 2019 Canadian Federal Election. However, based on Figure 1, there are almost 5000 people who don't know or prefer not to say which party he/she would like to vote in the survey. They account for approximately 23% of our cleaned survey data size. In the real election, their choices might be determined and influential, which might result in the inaccuracy of our prediction.

2.2 Model

The model I used in this report is MRP as mentioned in the Introduction by running RStudio [RStudio Team, 2020] with R installed [R Core Team, 2020]. I use the survey data to fit the multilevel regression model with a binomial family and then apply the fitted model to the cleaned census data by using Post-Stratification.

I will build two similar multilevel regression models, one for the Liberal Party and another for the Conservative Party and apply each model to the census data to get the estimates of the probabilities of voting for the two parties among the 10 provinces in census data. Then I will use the estimated probabilities of voting for the two parties in the 10 provinces to get the estimates of seats in each province for the Liberal Party and the Conservative Party. In this report, I assume the party which got a higher number of seats would become the governing party. In our analysis, the leader of this party would win the 2019 Canadian federal election if "everyone" had voted. More information about the Canada Election rule can be found in [Qualter et al.]

Model Specifics

I will introduce the model of the Liberal Party first. I used a logistic regression model as my level 1 model since "voteLib" are binary variables. Since "province" is in Group level based on what we learned in the class, a level 2 model would be built to account for different intercepts for level 1 model when provinces are different. The level 1 and level 2 models about the Liberal Party are shown as Equation (1) and Equation (2) respectively:

$$
\begin{aligned}
log(\frac{\hat{P}}{1-\hat{P}}) = \beta_{0j} &+ \beta_1 x_{age\_group60+,j} + \beta_2 x_{age\_group18-29,j} + \beta_3 x_{age\_group30-44,j} \\
&+ \beta_4 x_{sexMale,j} \\
&+ \beta_4 x_{household\_income\$125,000+,j} + \beta_5 x_{household\_income\$25,000-\$49,999,j} \\
&+ \beta_6 x_{household\_income\$50,000-\$74,999,j} + \beta_7 x_{household\_income\$75,000-\$99,999,j}
\end{aligned} \tag{1}
$$

$$
\beta_{0j} = r_{00} + r_{01} W_j \tag{2}
$$

In the level 1 model (Equation (1)), $\hat{P}$ is the dependent variable and represents the probability of voting for the Liberal Party. There are 3 independent variables, age_group, sex and household_income. Since they are all categorical, dummy variable encoding is used for all of them, resulting in the Equation (1)). j is the j province. β_{0j} is the intercept, which is the dependent variable in the level 2 model.

β_{0j} shows the log odds of the probability of voting for the Liberal Party when age_group is "45 to 60", Sex is Female and household income is "$100,000 to $124,999".

For other βs, β_1, β_2, β_3, β_4, they are all slope parameters. Each of them shows the change in log odds of the probability of voting for the Liberal Party when the responding x value changes from 0 to 1. Actually, these x values can only be 0 or 1 since they are dummy variables. More specifically, $x_{age\_group60+,j}$, $x_{age\_group18-29,j}$ and $x_{age\_group30-44,j}$ can only have one variable be 1 at the same time since a person can only in one age group every time or they can be all 0, which indicates the age group is "45 to 60". Same logic applies to the variables about household income.

In the level 2 model (Equation (2)),β_{0j} is the predicated intercept of the level 1 model, and each value for j corresponds to a different province. r_{00} is the intercept of the level 2 model. r_{01} is the slope parameter. W_j is the independent variable that influencing province.

The multilevel regression model about Conservative Party and the understanding is the same as Liberal Party's, except now in level 1, the dependent variable $\hat{P}$ represents the probability of voting for the Conservative Party.

Post-Stratification

In this section, I would do post-stratification: apply the models I built to the census data. Post-Stratification is the practice of partitioning data into thousands of demographic cells, and the final estimation is calculated by the weighted estimate for each cell. This technique could reduce the bias from non-probability based sampling. To begin the post-stratification process, for each province, I partition the data into demographic cells by age_group, sex and household income. This is also the reason why when I do the data cleaning process, I make these variables to be categorical.

To do Post-Stratification on the Liberal Party's model: the post-stratification formula (Equation (3)) would be used to calculate the probability of voting for the Liberal Party for each province:

$$
\hat{y}^{PS} = \frac{\Sigma N_j \hat{y}_j}{\Sigma N_j} \tag{3}
$$

In Equation (3), $\hat{y}^{PS}$ is the estimated probability of voting for the Liberal Party in a particular province; $\hat{y}_j$ is the estimate of the probability of voting for the Liberal Party in each cell of that province; N_j is the number of voters of the j^{th} cell for that province.

Post-Stratification on the Conservative Party's model would use the same equation as the Liberal Party's except $\hat{y}^{PS}$ is now the estimated probability of voting for the Conservative Party and $\hat{y}_j$ is the estimate of the probability of voting for the Conservative Party in each cell for a particular province.

Using Post-Stratification, we would get the estimated probability of voting for these two parties. To figure out whether Andrew Scheer or Justin Trudeau would win the 2019 Canadian federal election if "everyone" had voted, we need to look at the total number of seats these two parties got. I would estimate the number of seats each party got from a province by the product of the probability of voting for the party in that province and the number of seats that province have. Then the total number of seats each party got is the sum of the seats they got in all provinces. The results would be shown in Result section.

Model Check

Before moving to the Result Part, I would discuss if the models I chose are good and valid. I will verify this by using the ROC curve technique. If the area under curve (AUC), is greater than 0.5, I could consider my models are good [GraphPad].

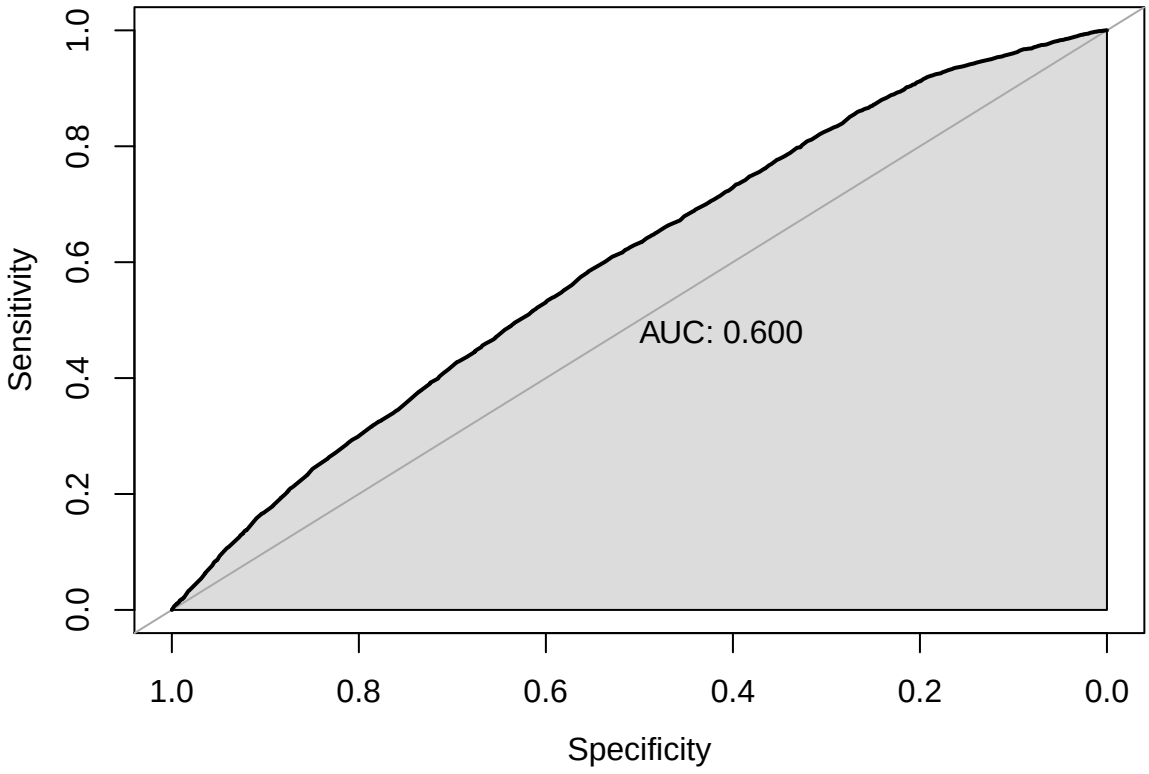

Figure 2: ROC curve of Liberal Party Model

From the Figure 2 and Figure 3, we could see the AUC for both of them are much greater than 0.5. Therefore, I could conclude my models are good and valid.

More explanation and analysis about the coefficients of the two models could be found in Appendix A.

Alternative Model

It should be acceptable by just using a logistic regression model with the 4 variables: province,age_group,sex,household_incom all be categorical. However, the drawback of the logistic regression model is that it does not take into account that different provinces might result in different intercepts as seen in level 1 of the multilevel

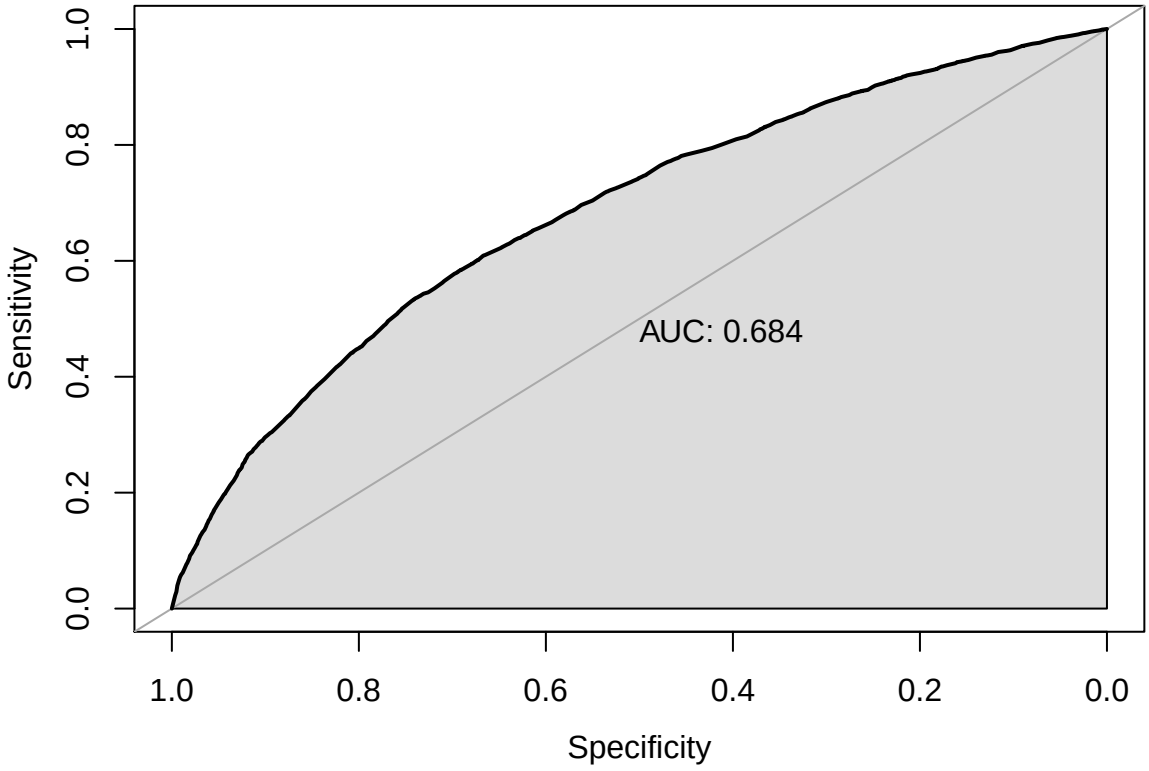

Figure 3: ROC curve of Conservative Party Model

Table 3: Probability of voting the Liberal Party among each province

province	probability
Alberta	0.1576650
British Columbia	0.2760662
Manitoba	0.2854885
New Brunswick	0.3505286
Newfoundland and Labrador	0.4127339
Nova Scotia	0.3892143
Ontario	0.3516403
Prince Edward Island	0.3610852
Quebec	0.2989038
Saskatchewan	0.1377931

Table 4: Probability of voting the Conservative Party among each province

province	probability
Alberta	0.5594167
British Columbia	0.2831276
Manitoba	0.3727373
New Brunswick	0.2627144
Newfoundland and Labrador	0.2117861
Nova Scotia	0.1770610
Ontario	0.2740196
Prince Edward Island	0.1317391
Quebec	0.1523329
Saskatchewan	0.4793926

regression model (Equation (1)). A single logistic regression model may not as accurate as the multilevel regression model. However, it would run faster than multilevel regression model, especially when there are a lot of level 2 variables in the multilevel regression model. It would be time-efficient. In our case, level 2 only contains the "Province" variable, the multilevel regression model running time is not long and it would be logically more accurate than a single logistic regression model, so I chose multilevel regression model as my final model.

3. Results

Table 3 shows the Probability of voting for the Liberal Party among each province.

Table 4 shows the Probability of voting for the Conservative Party among each province.

Based on Elections Canada [Canada], the number of seats each province/Territory has is as Table 5

Notice in our census data, the three territories are excluded, therefore the number of seats each party got is estimated only by the 10 provinces.

The total number of seats each party got is shown in Figure 4:

According to Figure 4, the estimated result is that the number of seats the Liberal Party got would be 104 and the number of seats the Conservative Party got would be 97. This is based off the number of seats in each province and and the post-stratification analysis of the proportion of voters in each province in favour of the Liberal Party and the Conservative Party modeled by two multilevel regression models with a binomial family, where the level 1 logistic model accounted for sex, age_group, household_income and level 2 model accounted for province.

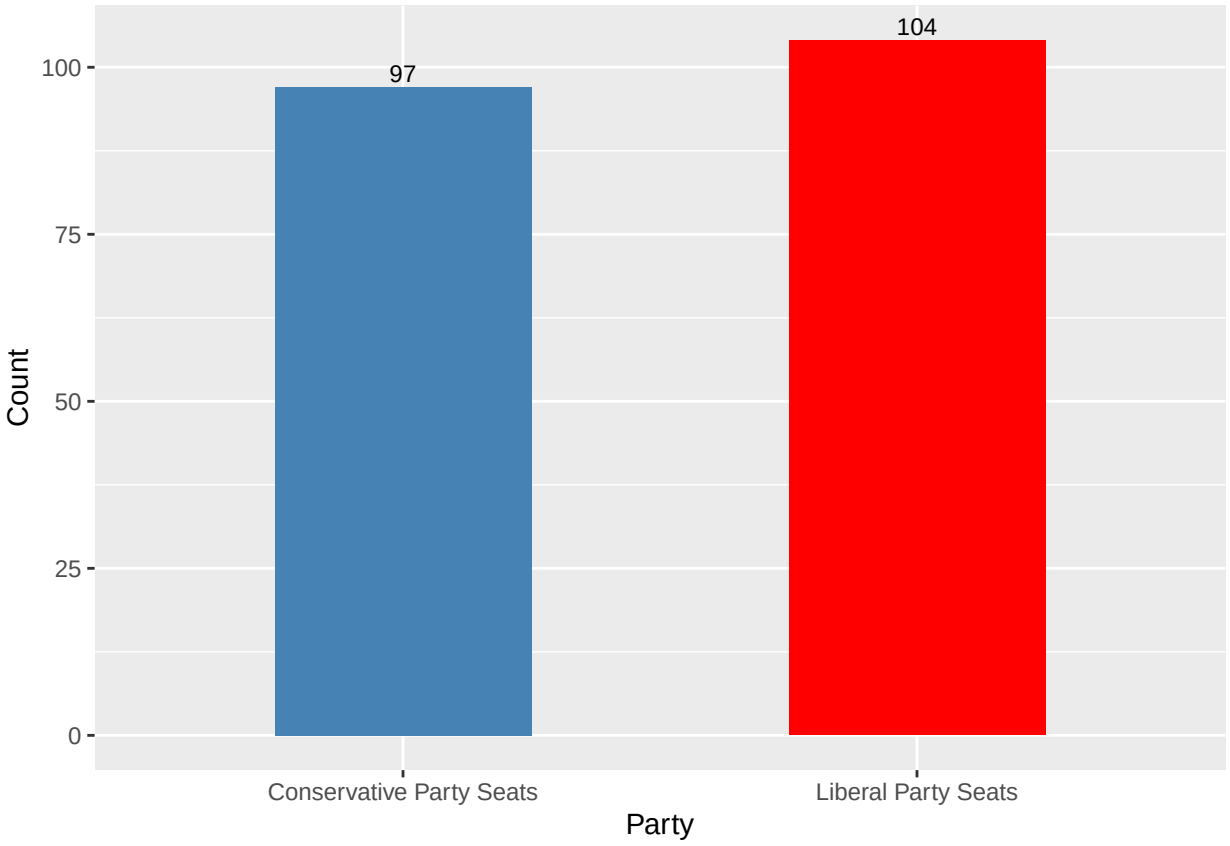

Figure 4: Number of Seats of Liberal Party and Conservative Party

Table 5: Allocation of Seats in the House of Commons

Province/Territory	Total Seats
Alberta	34
British Columbia	42
Manitoba	14
New Brunswick	19
Newfoundland and Labrador	7
Nova Scotia	11
Ontario	121
Prince Edward Island	4
Quebec	78
Saskatchewan	14
Yukon	1
Northwest Territories	1
Nunavut	1

It is clear Justin Trudeau would still win the 2019 Canadian federal election even if "everyone" had voted since the Liberal Party has 7 more seats than the Conservative Party. The result still would not change even if we assume the Conservative Party won the 3 seats in Yukon, Northwest Territories and Nunavut.

4. Discussion

In the paper, I investigated if the 2019 Canadian federal election result would change if "everyone" had voted, specifically, if Andrew Scheer would win the election instead of Justin Trudeau, by using MRP technique. I chose the 2019 Canadian Election Study - Online Survey as my survey data and the General social survey on Family (cycle 31), 2017 as my census data. First, I cleaned the data in order to do further analysis. I chose four common and meaningful variables between the survey and census data: sex, household income, age, province to do the analysis. In order to do post-stratification, I classified the continuous variables, household income and age, to be categorical with the same range. In addition, I made two new binary variables, voteLib and voteCon, based on the variable "cps19_votechoice" with the question being "Which party do you think you will vote for?" in Survey data. "voteLib" stands for "voting for Liberal Party" and value 1 means the respondent would like to vote for it and 0 means not. voteCon has the same logic but it's for the Conservative Party. The overview of the cleaned survey and census data can be found in Table 1and Table 2. The cleaned survey data has 21513 samples while the cleaned census data size is only 20178. Obviously, this is a drawback.

After the data were cleaned, I built two multilevel Regression models using a binomial family based on the survey data to predict the probability of voting for the Liberal Party and the Conservative Party. Then I use these two models to do post-stratification based on province in census data, splitting cells based on sex, household income and age group. After Post-Stratification, we would get the estimated probability of voting for these two parties among the 10 provinces, as shown on Table 3 and Table 4. Then I estimated the number of seats each party got from every province by the product of the probability of voting for the party in that province and the number of seats that province has according to the data on Table 5. Then the total number of seats each party got is the sum of the seats they got in all provinces. Finally, I got the result that Liberal Party would get 104 seats and Conservative Party would get 97 seats (Figure 4), so Justin Trudeau would still win the 2019 Canadian federal election even if "everyone" had voted.

In our analysis, Justin Trudeau would still win the 2019 Canadian federal election even if "everyone" had voted but the Liberal Party only have 7 more seats than Conservative Party. However, in other cases, the governing party might change and this might have a great influence on the country and the world. Like the example in the Introduction section, if "everyone" had voted in the 2016 US presidential election, Hillary Clinton would probably be president [The Economist] and the situation about COVID-19 in America might not be such terrible [Boot], 17,515,091 confirmed cases with 313,748 deaths up to 21st December 2020

according to World Health Organization [Organization]. This report would encourage every eligible citizen in each country to vote during the election period.

4.2 Weaknesses

Generally, there are four weaknesses in our model and prediction.

First, the cleaned survey dataset has 21513 samples while the cleaned census data size is only 20178, which means the cleaned census data size is too small. In addition, the census data was collected in 2017 [Faculty of Arts & Sciences, b], which would have differences with the population size in 2019 and would increase the bias.

Second, there are almost 5000 people (Figure 1) who don't know or prefer not to say which party he/she would like to vote in the survey data, which account for approximately 23% of the cleaned survey data size. In the real election, their choices might be determined and influential, which might result in the inaccuracy of our prediction.

Third, the census data excludes people from Yukon, Northwest Territories, and Nunavut. When we do post-stratification on the census data, we could not get the estimates of the probabilities of voting for the Liberal Party and the Conservative Party among these three territories, which would result in the lower estimates of total seats each party got, although these three territories only have 3 seats and would not change our final prediction result in this case since the Liberal Party have 7 more seats than Conservative Party.

Lastly, our models only use 4 variables: sex, age group, household income and province. This is not a large number of explanatory variables. If more variables like education and race could be used in our models, the prediction result would be more accurate and less biased.

4.3 Next Steps

There are 3 next steps we could do to improve our analysis.

First, find a greater census dataset that was collected in 2019 and contains all regions of Canada to do the Post-Stratification to increase the accuracy of the prediction.

Second, include more explanatory variables in our models to make the models better.

Third, try to learn deeper knowledge about statistics and see if can build more proper models.

Appendix A

Figure 5 and Figure 6 show the total observations of different provinces in raw survey Data and census data respectively, where we could see in the survey data there are three more territories: Yukon, Northwest Territories and Nunavut and the respondents in these three territories are much less than those in the other 10 provinces.

Appendix B

From Table 6, most variables about household_income are significant since their p values are small. SexMale is not significant since it has a large p value but the intercept has a very small p value, we could infer SexFemale should be significant. In addition, we could infer and conclude age_group from 45 to 59 should be significant. Overall, most variables are significant. This multilevel regression model is a good model for predicting the probability of voting for the Liberal Party.

From the Table 7, most variables about age_group, sex and household_income are significant since their p values are small except "household_income$50,000 to $74,999" and "household_income$75,000 to $99,999".

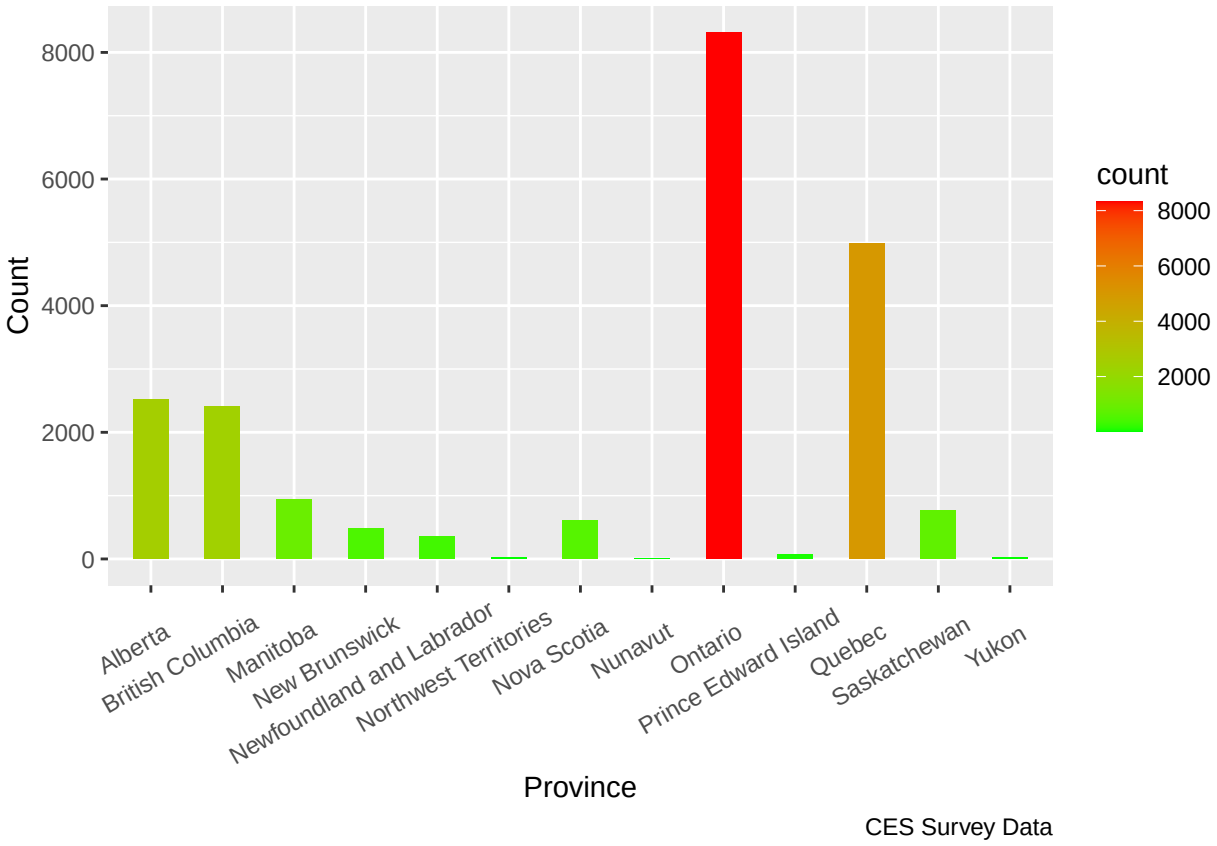

Figure 5: Total observations of different provinces in survey Data

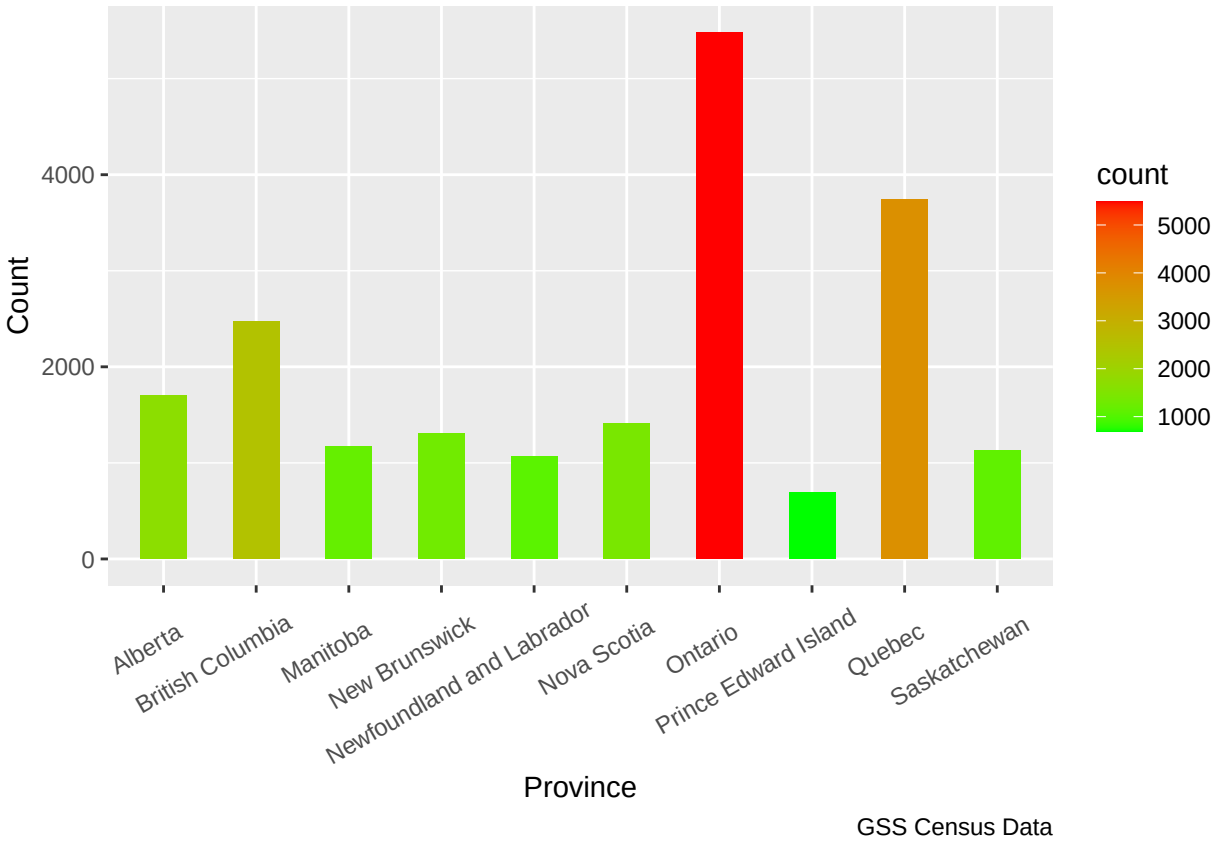

Figure 6: Total observations of different provinces in census Data

Table 6: Summary of Cofficient of the Liberal Party model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.7916928	0.1509637	-5.2442608	0.0000002
age_group60+	0.2106690	0.0395035	5.3329145	0.0000001
age_groupAge 18 to 29	0.0548459	0.0541253	1.0133138	0.3109103
age_groupAge 30 to 44	0.0454360	0.0412764	1.1007738	0.2709951
sexMale	-0.0119054	0.0309918	-0.3841469	0.7008696
household_income$125,000 and more	0.0145640	0.0549699	0.2649452	0.7910517
household_income$25,000 to $49,999	-0.2704240	0.0570307	-4.7417286	0.0000021
household_income$50,000 to $74,999	-0.2125261	0.0544808	-3.9009362	0.0000958
household_income$75,000 to $99,999	-0.0936475	0.0556284	-1.6834469	0.0922886
household_incomeLess than $25,000	-0.2917659	0.0658863	-4.4283230	0.0000095

Table 7: Summary of Cofficient of the Conservative Party model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.1781299	0.2050423	-5.7457883	0.0000000
age_group60+	0.0717732	0.0405719	1.7690344	0.0768881
age_groupAge 18 to 29	-0.3936207	0.0593472	-6.6325081	0.0000000
age_groupAge 30 to 44	-0.1771601	0.0428926	-4.1303212	0.0000362
sexMale	0.4725797	0.0323415	14.6121744	0.0000000
household_income$125,000 and more	0.1262472	0.0571163	2.2103536	0.0270806
household_income$25,000 to $49,999	-0.1738590	0.0603997	-2.8784718	0.0039961
household_income$50,000 to $74,999	-0.0003091	0.0568675	-0.0054363	0.9956625
household_income$75,000 to $99,999	-0.0108914	0.0583160	-0.1867653	0.8518447
household_incomeLess than $25,000	-0.3915402	0.0717034	-5.4605524	0.0000000

This multilevel regression model is a good model for predicting the probability of voting for the Conservative Party.

Appendix C

In this report, we used R [R Core Team, 2020], Rstudio [RStudio Team, 2020] and R packages: tidyverse [Wickham et al., 2019],lme4 [Bates et al., 2015], knitr [Xie, 2019] [Xie, 2015] [Xie, 2014], cesR [Hodgetts and Alexander, 2020], labelled [Larmarange, 2020] and pROC [Robin et al., 2011].

References

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

Max Boot. This wouldn't have happened if hillary clinton had won. URL https://www.washingtonpost.com/opinions/2020/03/21/this-wouldnt-have-happened-if-hillary-clinton-had-won/.

Elections Canada. House of commons seat allocation by province. URL https://www.elections.ca/content.aspx?section=res&dir=cir/red/allo&document=index&lang=e.

W. Wang et al. Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31 (3):980–991, 2015. doi: https://doi.org/10.1016/j.ijforecast.2014.06.001.

University of Toronto Faculty of Arts & Sciences. 2017 general social survey (gss): Families cycle 31 user's guide, a. URL https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf.

University of Toronto Faculty of Arts & Sciences. Canadian general social surveys (gss), b. URL https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm.

GraphPad. Interpreting logistic roc curves. URL https://www.graphpad.com/guides/prism/8/curve-fitting/reg_logistic_roc_curves.htm.

Paul A. Hodgetts and Rohan Alexander. *cesR: Access the CES Datasets a Little Easier.*, 2020. R package version 0.1.0.

Joseph Larmarange. *labelled: Manipulating Labelled Data*, 2020. URL https://CRAN.R-project.org/package=labelled. R package version 2.7.0.

World Health Organization. United states of america situation. URL https://covid19.who.int/region/amro/country/us.

Terence H. Qualter, John M. Wilson, Norman Ward, John C. Courtney, and Garth Stevenson. Canadian electoral system. URL https://www.thecanadianencyclopedia.ca/en/article/electoral-systems.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77, 2011.

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

Statistics Canada. Reasons for not voting in the federal election, october 21, 2019. URL https://www150.statcan.gc.ca/n1/daily-quotidien/200226/dq200226b-eng.htm.

Laura B Stephenson, Allison Harell, Daniel Rubenson, and Peter John Loewen. 2019 Canadian Election Study - Online Survey, 2020a. URL https://doi.org/10.7910/DVN/DUS88V.

Laura B Stephenson, Allison Harell, Daniel Rubenson, and Peter John Loewen. 2019 canadian election study - online survey technical report and codebook.pdf. In *2019 Canadian Election Study - Online Survey*. Harvard Dataverse, 2020b. doi: 10.7910/DVN/DUS88V/HRZ21G. URL https://doi.org/10.7910/DVN/DUS88V/HRZ21G.

The Economist. If everyone had voted, hillary clinton would probably be president. URL https://www.economist.com/graphic-detail/2019/07/06/if-everyone-had-voted-hillary-clinton-would-probably-be-president.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

Yihui Xie. knitr: A comprehensive tool for reproducible research in R. In Victoria Stodden, Friedrich Leisch, and Roger D. Peng, editors, *Implementing Reproducible Computational Research*. Chapman and Hall/CRC, 2014. URL http://www.crcpress.com/product/isbn/9781466561595. ISBN 978-1466561595.

Yihui Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition, 2015. URL https://yihui.name/knitr/. ISBN 978-1498716963.

Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2019. URL https://yihui.name/knitr/. R package version 1.23.