Liberal Might Win 31.80% of Popular Votes with 100% Turnout Rate

Yijie Zhao

22 December 2020

Abstract

Turnout rate has been always playing an important role in democratic election system. In this paper, we are going to figure out what outcome would occur if every qualified voters voted in 2019 Canadian Federal election. We will use multilevel regression with post-stratification techniques to estimate the vote share that Liberal would win if the turnout rate is 100%. We will also discuss about the importance of turnout rate at last.

Keywords: Federal Election; Liberal; Turnout; Multilevel Regression with Post-stratification.

1 Introduction

In 2019 Canadian federal election, Trudeau won his re-election and Liberal Party continued to be the party in power. However, even though the Liberal has won the federal election, they did not achieve a complete victory. They won less vote share of popular votes than the Conservative. Also, they did not won majority seats. The end result was that the Liberal party formed a minority government. If the turnout rate was 100% in 2019 Canadian federal election, will the results be different for Liberal?

In this paper, I use two sets of data, one is 2019 CES survey data, which is completely about the 2019 Canadian federal election; the other is 2016 Canadian census data, which should be representative enough to represent the population of Canada. I will first load and clean these two datasets, and choose some demographic variables that I think matters the vote intention of individuals. I will talk about the source, survey design and import features of these two datasets in the data section. I will also draw some plots of same variables in different dataset to compare their distributions. In the model section, I will use MRP technique to build a multilevel logistics regression model based on the 2019 CES survey data and then apply this model to predict the vote intention of individuals in census data. I will compare the results with 2019 election results. At last, I will discuss further more about the use of MRP and the importance of turnout rate for democratic election system. Also, I will think about the weakness of data and model, to try to find which works can be improved in the future.

2 Data

In order to figure out the influence of voter turnout on the 2019 Canadian federal election, I chose two sets of data for analyzing and modeling. One is 2019 Canadian Election Study (CES) survey data (Stephenson (2020)) and the other is 2016 Canadian census data (2016 (n.d.)). In this paper, I used 2019 CES data as individual level survey data, to build model for predicting the polling of popular votes. Meanwhile, I used 2016 Canadian census data as post-stratification data, to make better estimates by adjusting the outcomes of model prediction with the population distribution.

In the following data section, I first discussed about the key points of these two data sets in the first two parts, for example, contents of survey design and data collection. Then I drew some plots of each data, to make key features of these data more intuitively visualized.

2.1 Individual Level Survey Data

For individual-level survey data, I used the 2019 CES survey data released by Canadian Election Study in August 17, 2020. Canadian Election Study is a study group led by Laura Stephenson et al, focusing on the research of Canadian election activities, since 1965. This study group makes surveys on every election year and collect all aspects of data related to Canadian election from their target population. I obtained the datasets of 2019 CES survey data from loading a R package named cesR (Hodgetts and Alexander (2020)). The CES team made two modes of survey this time, one is online survey with sample size of 37,822 and the other is phone survey with sample size of 4,021. In this analysis paper, I chose the online survey (Stephenson (2020)), because it contains a much larger sample size of data than the phone survey, which can enhance representativeness of population.

The population of this 2019 CES survey is all Canadian citizens and permanent residents, with age of 18 or older. Around the period of election, they made a two-wave online survey. The first wave started from September 13 to October 21, 2019, which was during the campaign period, with 37,822 interviews. The second wave started from October 24 to November 11, 2019, which was after the federal election, with 10,337 interviews. They used stratified sampling by region and sampled 37,822 respondents through Qualtrics, which provided them with the frame list of target population. As for the response rate and data quality, they got 37,822 responses in total for campaign period survey before the federal election, and among these responses, about 89.64% of responses are with high quality.

I chose the data of campaign period online survey as my individual level survey data. The first reason is that the campaign period online survey sampled much more respondents than the post election survey. Second, I intended to build a prediction model for voting in 2019 federal election. Therefore, I need a response variable which represents an intention of voting or a preference of different Canadian political parties, not a variable of voting results after election.

After datasets selection, I used R (R Core Team (2020)) and R package tidyverse (Wickham et al. (2019)), labelled (Larmarange (2020)) to write scripts for cleaning this survey data. First, I removed all samples who did not have Canadian citizenship, which means they didn't have rights to vote in federal election. Among the 620 variables of this survey data, I focused on demographic variables. I selected some demographic variables and used them to create new categorical variables as explanatory variables for this analysis paper. They reason why I adjusted these original variables is because I need to make each pair of homogeneous variables exactly the same in survey data and the following census data, so that I could build prediction model by using survey data and apply model to census data successfully.

The adjusted explanatory variables are showed as follows:

- age_group;
- gender;
- education;
- working_status;
- income;
- marital;
- province

I also created a new binary categorical variable named vote_Liberal, valuing 1 with "Yes" and 0 with "No".

2.2 Post-stratification Data

For post-stratification data, I used the data of "Census of Canada, 2016: individual public use microdata file" (2016 (n.d.)), which is the most recent census data. The 2016 Canadian census data was collected and released by Statistics Canada in 2016. Statistics Canada holds this census program every five years. I obtained the public used microdata file of 2016 census data through library website of University of Toronto. I visited the website of CHASS data centre and downloaded the data file of 2016 census program.

The target population of this 2016 census data is people lived in Canada on the reference date, May 10, 2016. The frame is a list of household addresses of individuals and households. 2016 census data collection used the wave methodology. There were 3 waves in total, which aimed at reminding people to complete survey and reducing the cases of non-response. There were three options to complete census survey, responding online, writing mailed paper questionnaire and mailing back, and contacting help line. In the past, there were one short mandatory questionnaire and one long voluntary questionnaire. Since 2011, Statistics Canada started to use only one mandatory short form and move most of questions of long form to a voluntary survey called National Household Survey. I chose the PUMF of 2016 census data. This data file was sampled in two-phase to the respondents of long form census, by systematic sampling methods. The total sample size is 930,421, which represents 2.7% of target population.

As for the data quality of this census file, the sampling errors still exist obviously because this data file sampled from long form respondents, i.e., sample from samples. In this analysis paper, I assumed this census data is representative for population. As for non-response in census program, Statistics Canada changed the mandatory survey in 2015. Therefore, the response rate of 2016 census survey are extremely high, 96.6%.

I used R (R Core Team (2020)) to write scripts for cleaning this census data, the same as how I cleaned and prepared the survey data in the previous part. I adjusted demographic variables I chose to match the explanatory variables in survey data. The adjusted explanatory variables are showed as follows:

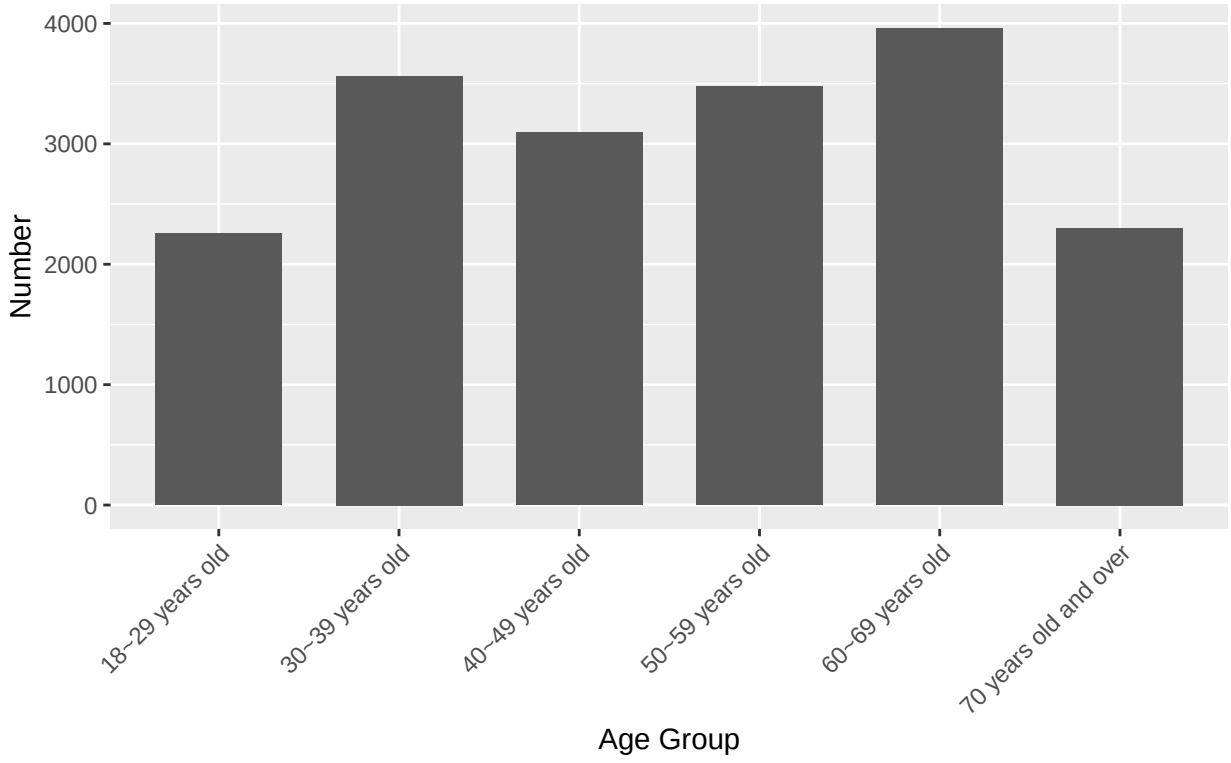
- age_group;
- gender;
- education;
- working_status;
- income;
- marital;
- province

Besides, as the census data was collected in 2016, I should assume that there are no remarkable differences of the target population, from 2016 to 2019, in terms of population and other variables I listed in data section.

2.3 Data Visualization

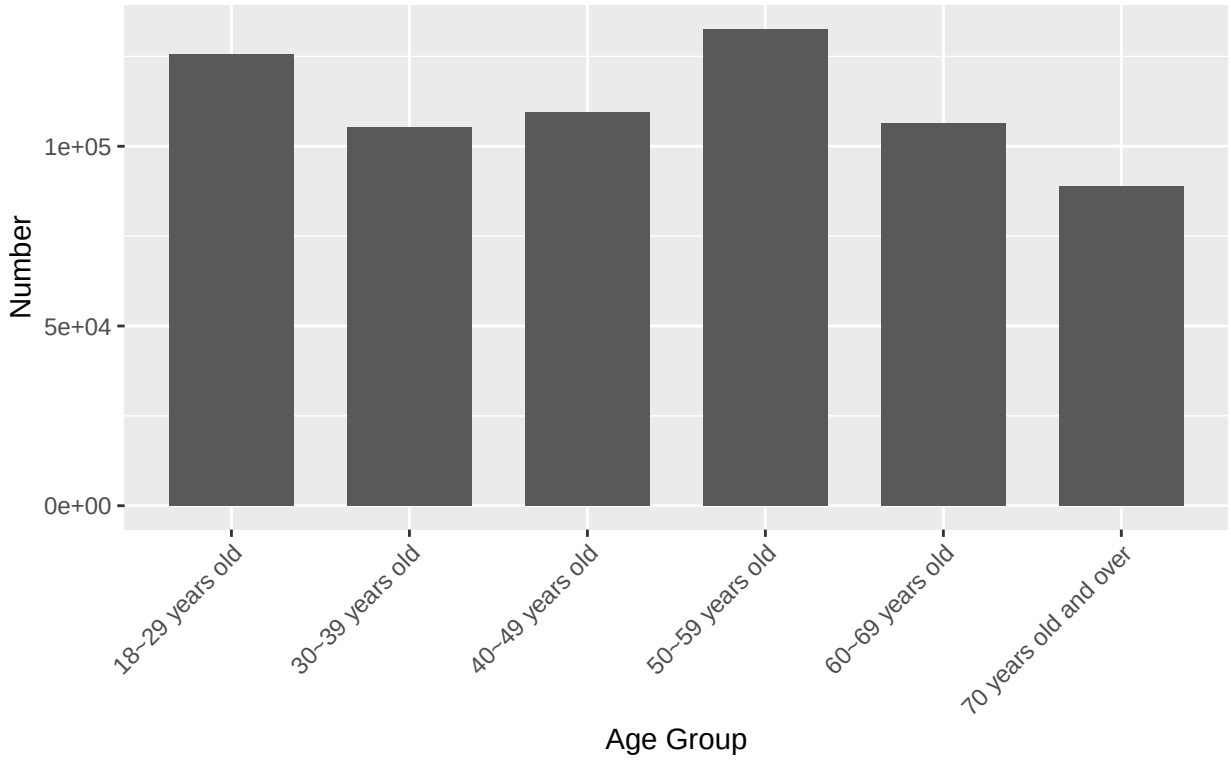
After data cleaning, there are 18645 observations of 8 variables in the cleaned CES dataset and 668655 observations of 7 variables in the cleaned census dataset. Before building MRP model, I need to compare the differences of variable distributions of survey data and census data, in a more intuitive way. Therefore, I drew 7 pairs of plots for 7 explanatory variables that I had interested in, by using R package tidyverse (Wickham et al. (2019)) and ggplot2 (Wickham (2016)).

Figure 1. Age Distribution of 2019 CES Survey Data



Source: Canadian Election Study

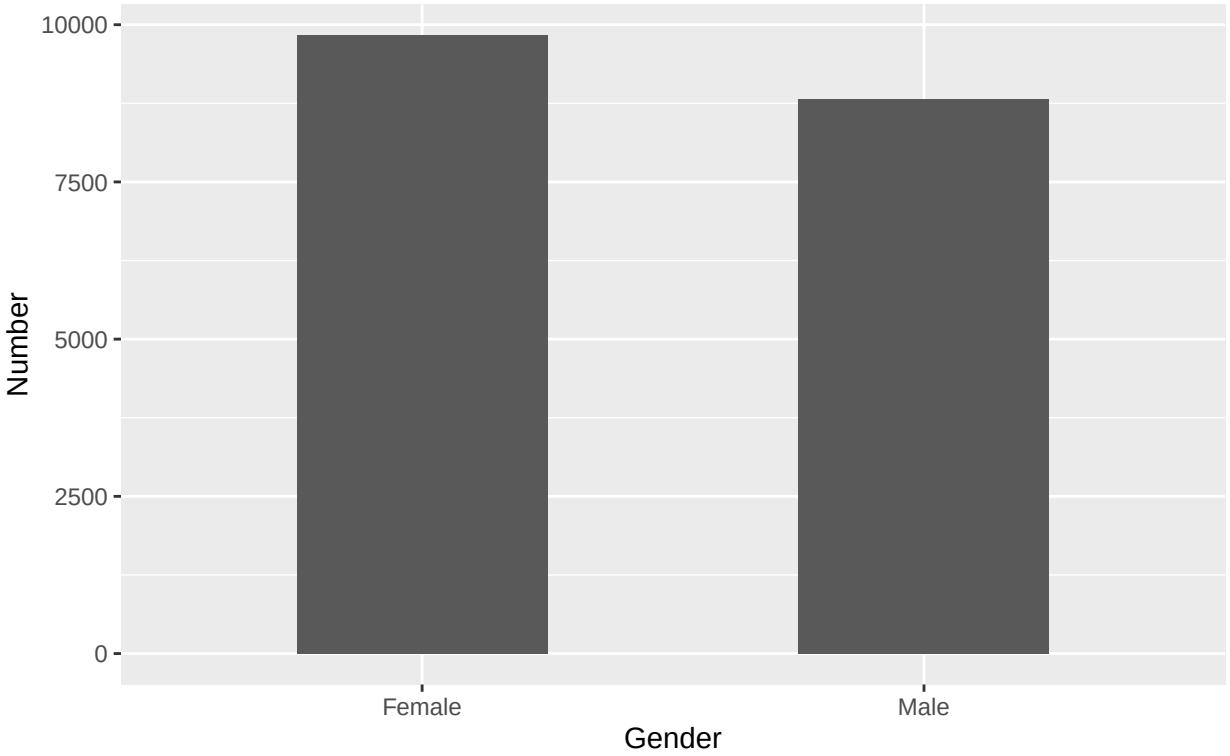
Figure 2. Age Distribution of 2016 Census Data



Source: Statistics Canada

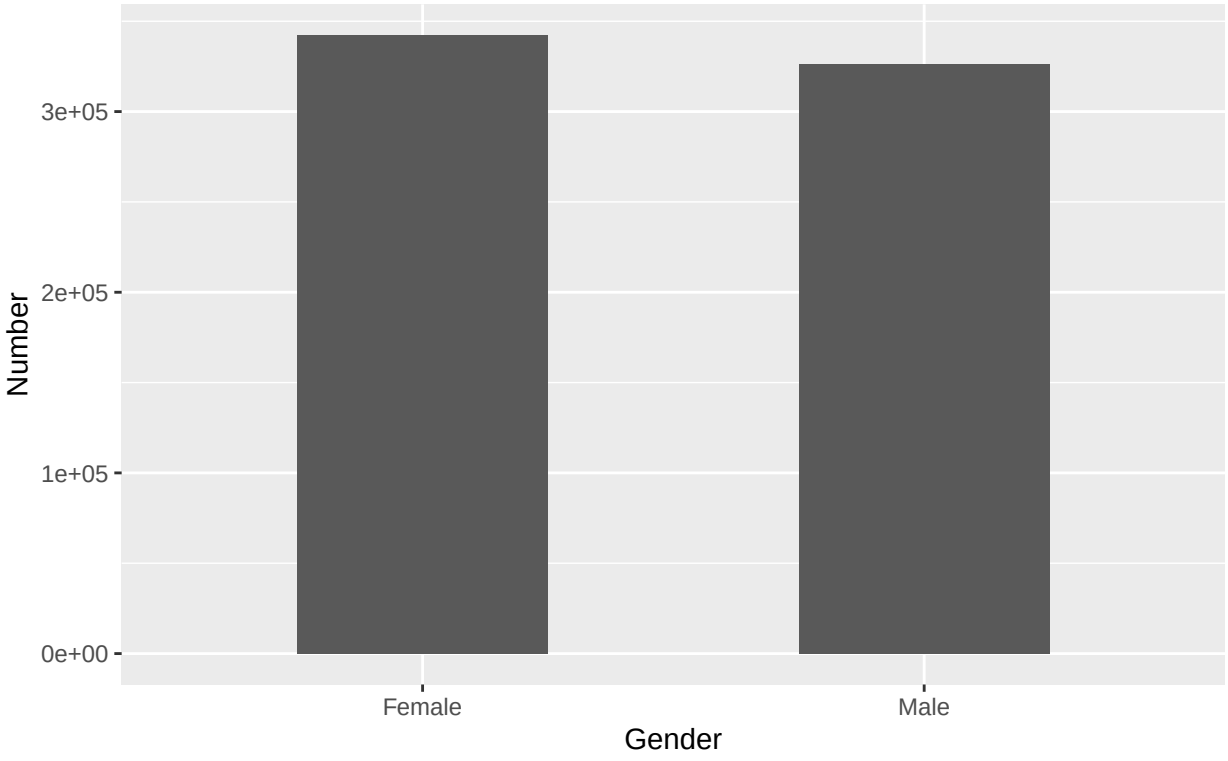
From Figure 2, you can find that in census data, the age groups of 18 to 29 years old and 50 to 59 years old have the largest two portion of population. However, comparing Figure 1 with Figure 2, the proportion of 18 to 29 year-old respondents is the smallest in survey data. That indicates the young people under 30 years old are under sampled in CES survey data, compared to census data.

Figure 3. Gender Distribution of 2019 CES Survey Data



Source: Canadian Election Study

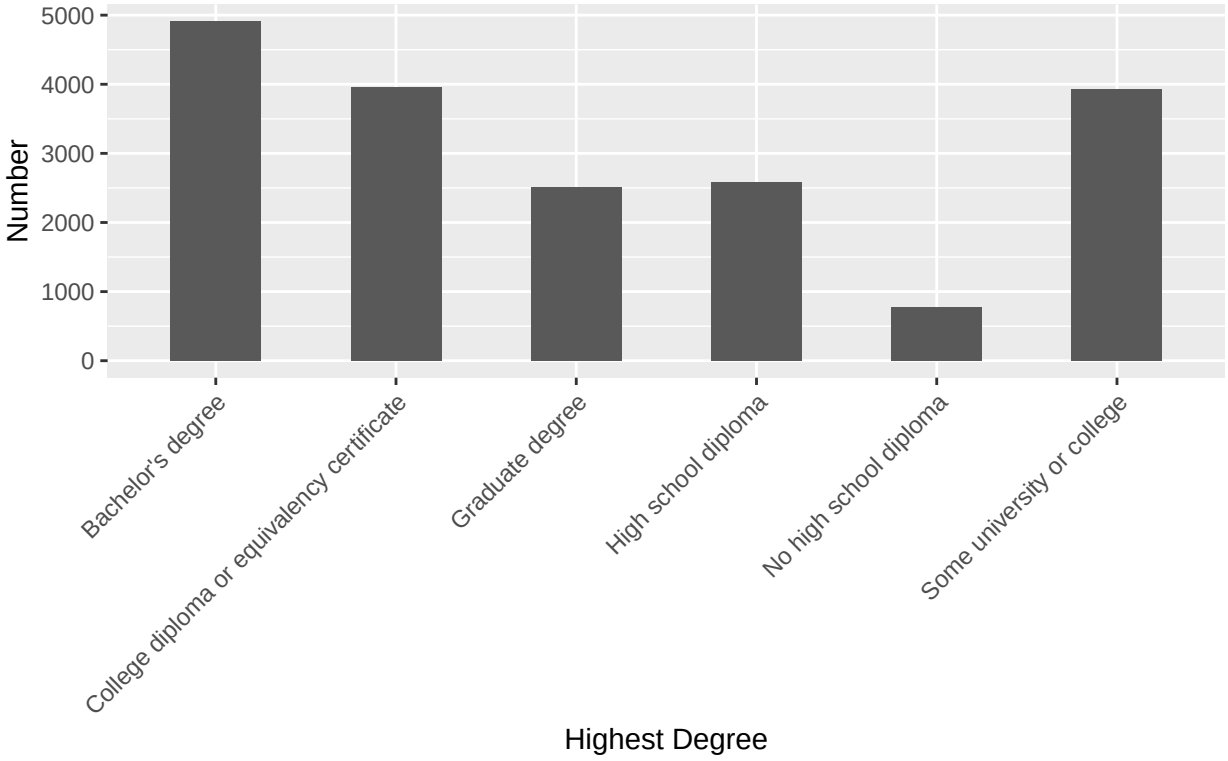
Figure 4. Gender Distribution of 2016 Census Data



Source: Statistics Canada

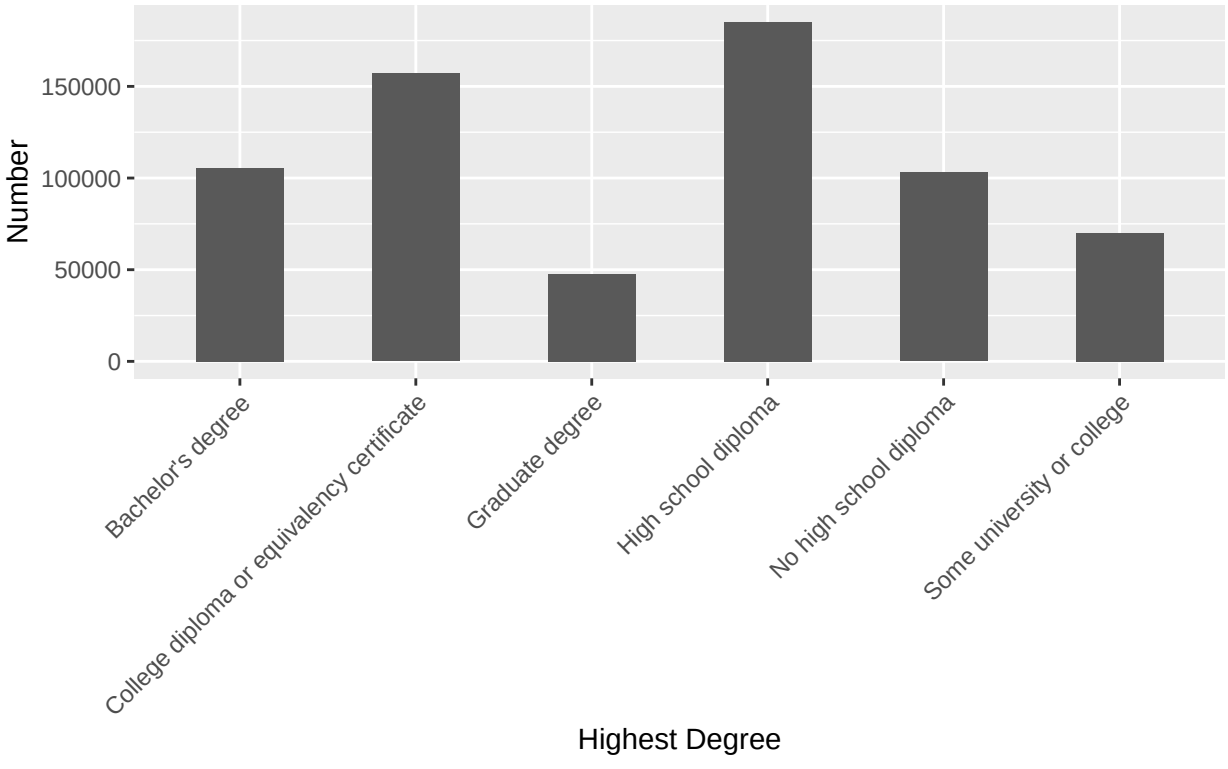
From Figure 3 and Figure 4, it is obvious that the ratio of gender is more unbalanced in survey data than in census data. Female are over sampled and male are under sampled in 2019 CES survey data, compared with 2016 census data.

Figure 5. Education Distribution of 2019 CES Survey Data



Source: Canadian Election Study

Figure 6. Education Distribution of 2016 Census Data



Source: Statistics Canada

From Figure 5 and Figure 6, we can figure that the distributions of respondents' highest degrees are totally different in two data sets. Compared to census data, people with and without high school diploma as highest degrees are both under sampled in survey data. Besides, people with bachelor's degree are obviously over sampled in CES survey data.

Figure 7. Working Status Distribution of 2019 Survey Data

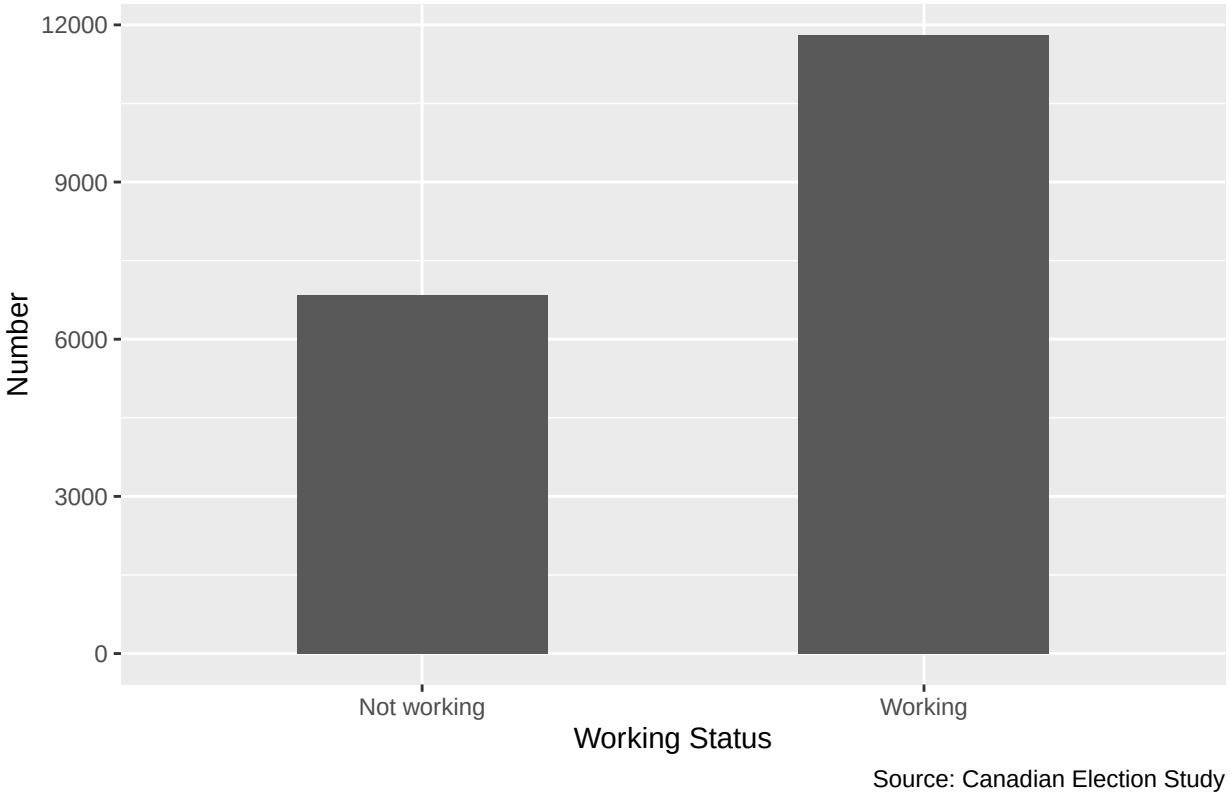
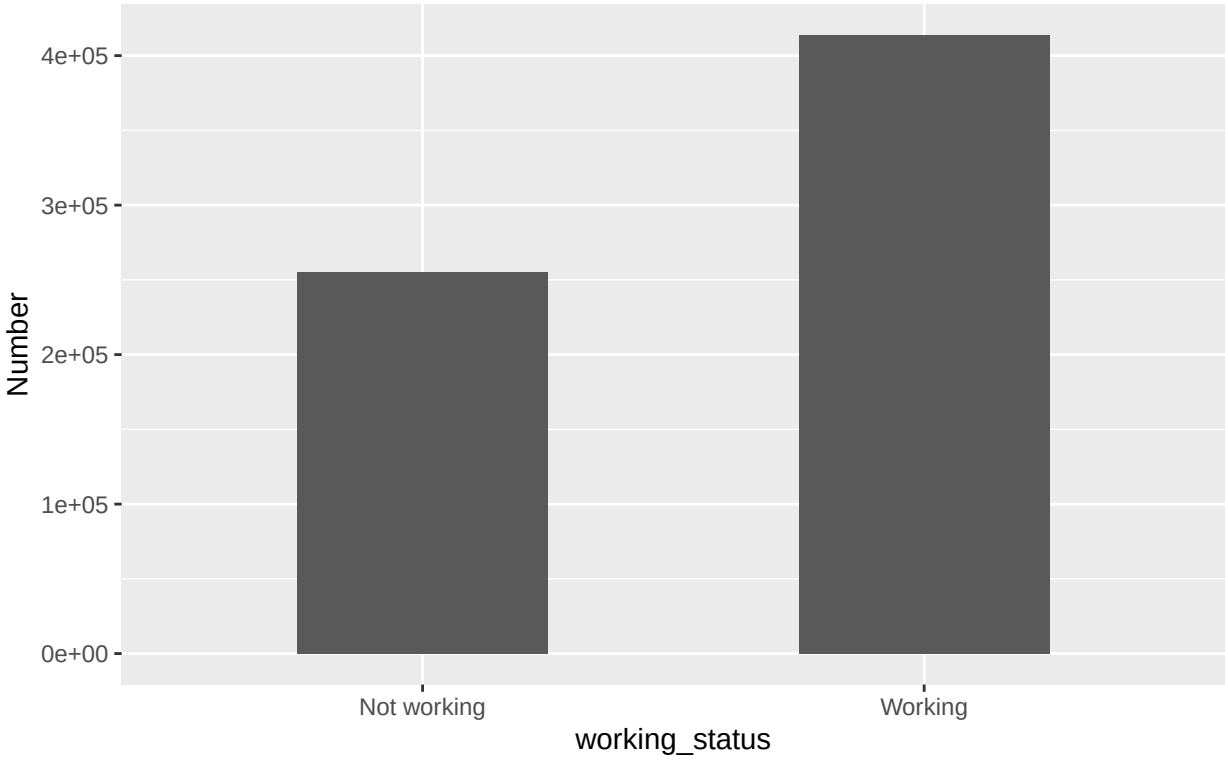
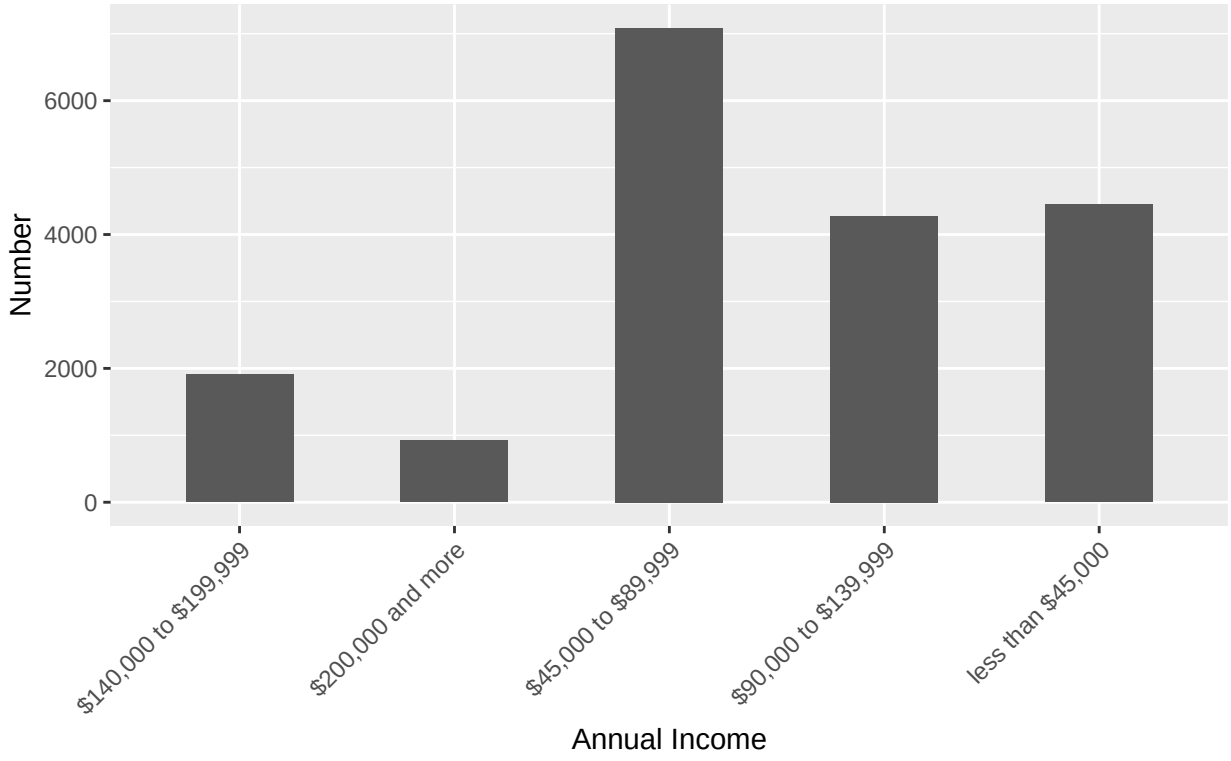


Figure 8. Working Status Distribution of 2016 Census Data



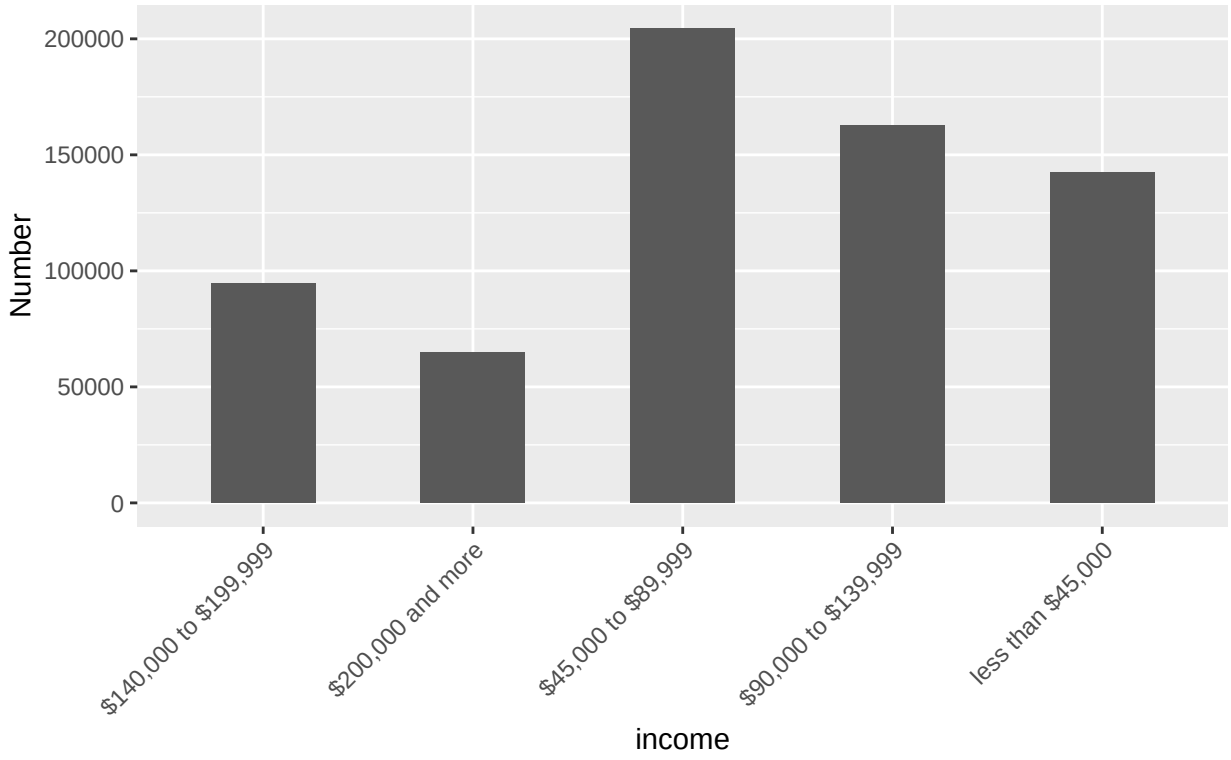
From Figure 7 and Figure 8, it is showed that the distributions of whether people worked are almost identical in survey data and census data. The number of people that are currently not working is almost half of the number of people working by the reference date.

Figure 9. Income Distribution of 2019 CES Survey Data



Source: Canadian Election Study

Figure 10. Income Distribution of 2016 Census Data



Source: Statistics Canada

From Figure 9 and Figure 10, I think people with annual income level of CAD 45,000 to $89,999 are over sampled, while other four groups of different annual income level are under sampled, relatively. But the income distributions of both data follow a pretty same distributed pattern.

Figure 11. Marital Distribution of 2019 CES Survey Data

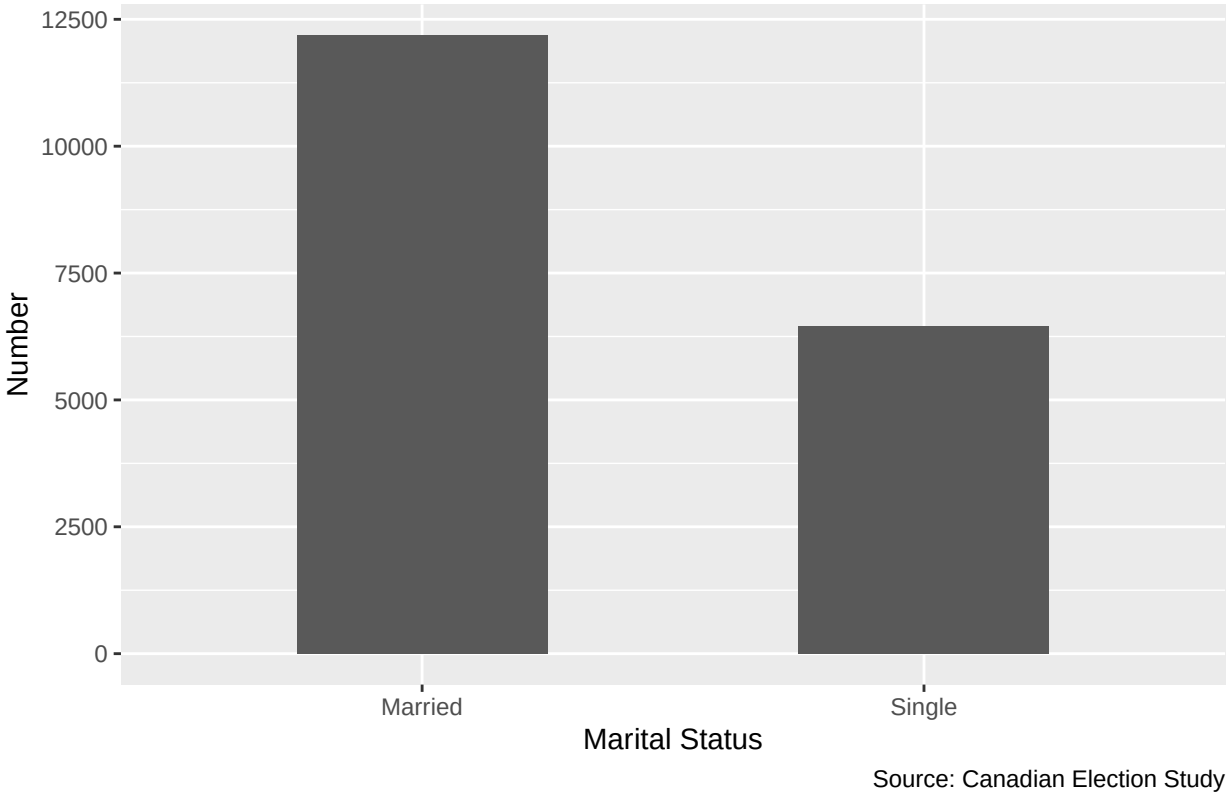
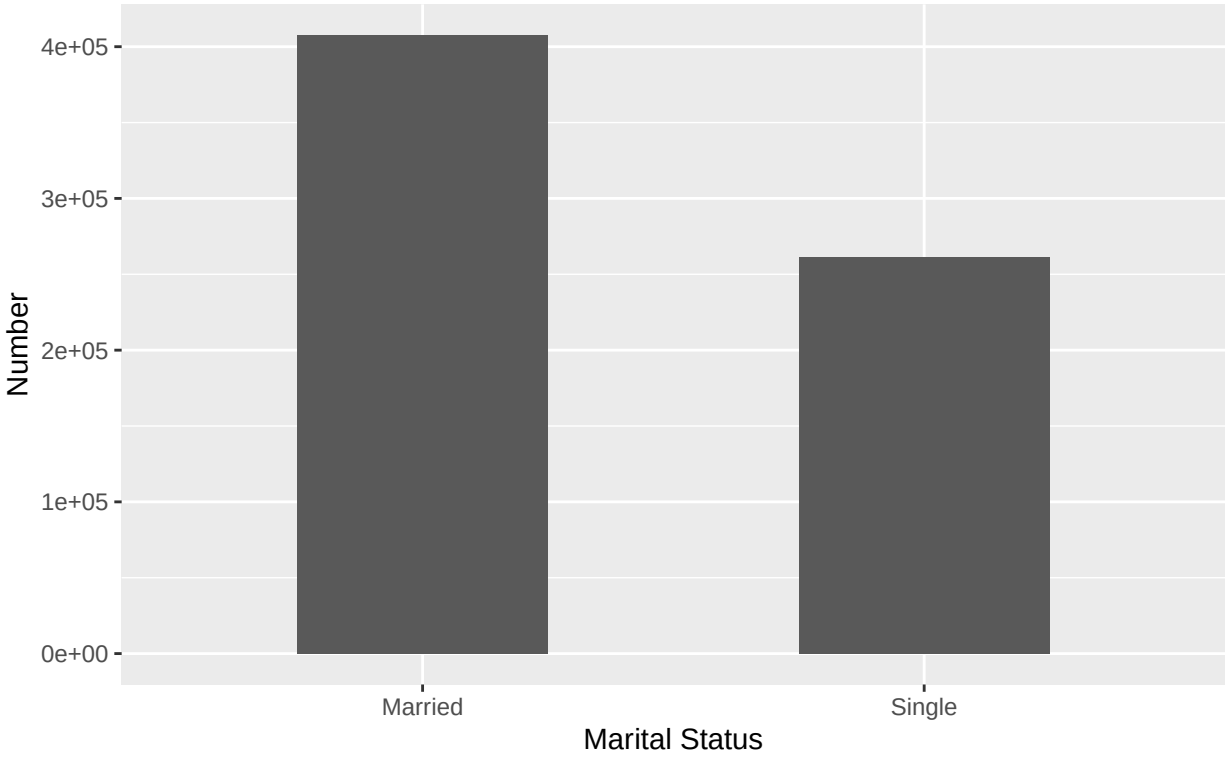


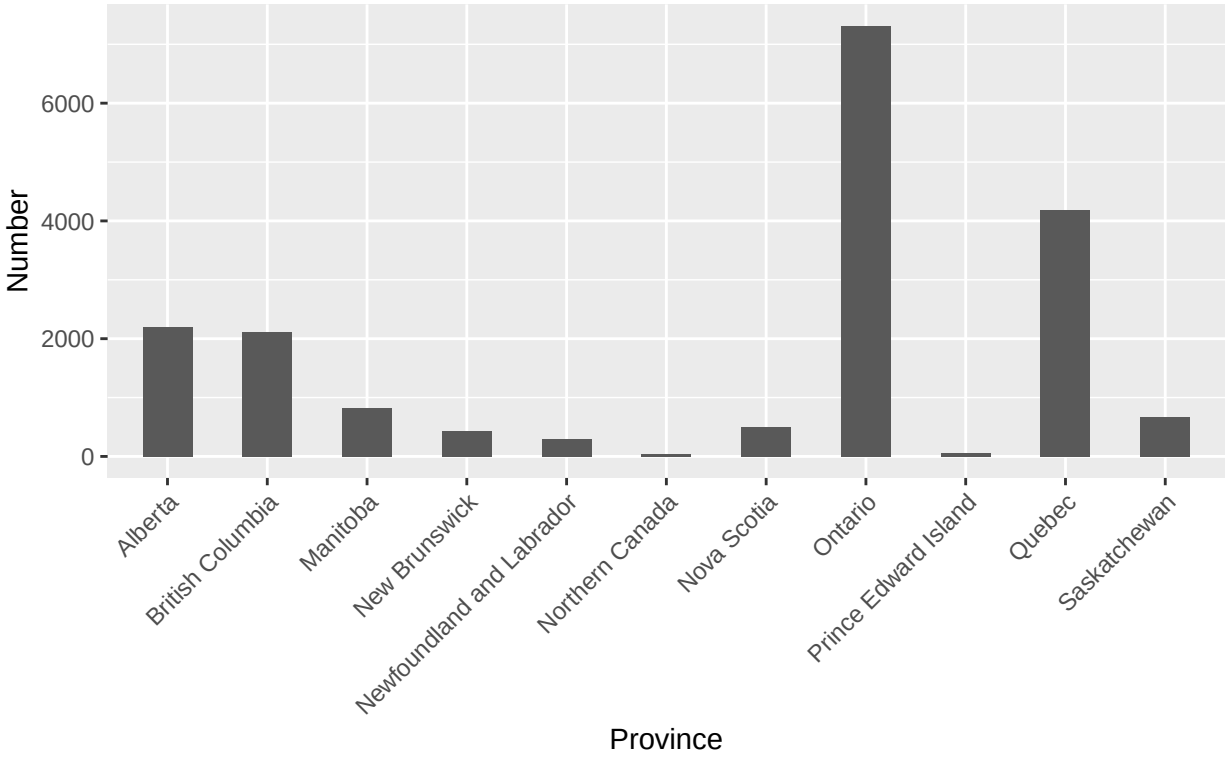
Figure 12. Marital Distribution of 2016 Census Data



Source: Statistics Canada

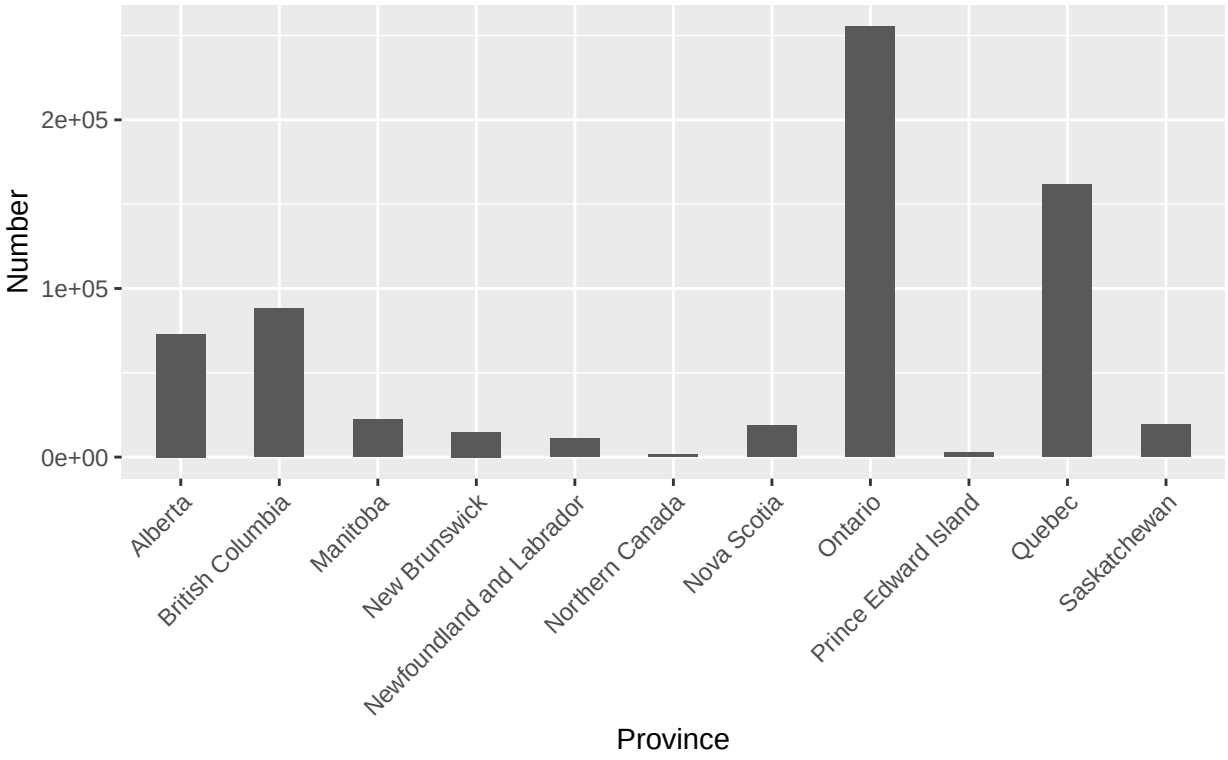
From Figure 11 and Figure 12, it is showed that the distributions of whether people are currently married are almost identical in survey data and census data. The number of people that are currently married is twice as the number of people who are single by the reference date.

Figure 13. Province Distribution of 2019 CES Survey Data



Source: Canadian Election Study

Figure 14. Province Distribution of 2016 Census Data



Source: Statistics Canada

From Figure 13 and Figure 14, it is showed that the distributions of which province people are currently living are almost identical in survey data and census data. The rank of provinces in terms of population is Ontario, Quebec, British Columbia and Alberta. The proportion of population in other province and territory is really small.

3 Model

After discussing the differences between distributions of each pair of explanatory variables in survey data and census data, I found it appropriate to build a logistics regression model to predict the probability that a voter would like to vote for Liberal Party in 2019 Canadian federal election, with explanatory variables of age, gender, education, working status, income, marital status and province. I used the 2019 CES survey data to build this model. However, as we all know, a single level of logistics regression model may not best fit the survey data, because there are obvious and huge differences between individuals, in terms of the 7 demographic variables; besides, some individuals also have something in common. Therefore, building a multilevel logistics regression model can solve this problem.

First, I need to choose a region or cell to divide individuals into several groups and individuals in the same group have something in common. I chose province as the cell of this multilevel model for 2 reasons. The one reason is that the province is a sort of natural group, which is not man-made. The second reason is that we usually hear that some certain provinces are the traditional stronghold of a political party. Grouping by province can hold the traditions consistent.

I built a multilevel logistics regression model as follows:

$$\frac{p_{ij}}{1-p_{ij}} = exp(\beta_{0j} + \beta_{1j}x_{1ij} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \epsilon_{ij})$$

with a random intercept and one random coefficient.

The j represents group, i.e., province j, where j can value 1 to 11, as there are 11 provinces or territories in total. Then we get 11 different logistics regression model with different intercept term β_0 and coefficient β_1 for 11 provinces. The i represents the i'th observations in group j. The fraction $\frac{p_{ij}}{1-p_{ij}}$ represents the odds and x_{2-6} represents gender, age group, education, working status, income and marital status respectively. And the ϵ represents error term.

I used R package lme4 (Bates et al. (2015)) to build the MR model and summarize the model outputs.

Given the R outputs of modeling, one of the multilevel logistics regression models are showed as follows.

(1) model of Alberta

$$\frac{p_i}{1-p_i} = exp(-1.3397905+0.13980249Male_i+0.1205704Age_{30-39}+0.1325729Age_{40-49}+0.1432926Age_{50-59}+0.3260491$$

(2) model of British Columbia

$$\frac{p_i}{1-p_i} = exp(-0.5582058-0.03699328Male_i+0.1205704Age_{30-39}+0.1325729Age_{40-49}+0.1432926Age_{50-59}+0.3260491$$

(3) model of Manitoba

$$\frac{p_i}{1-p_i} = exp(-0.4765806-0.05545706Male_i+0.1205704Age_{30-39}+0.1325729Age_{40-49}+0.1432926Age_{50-59}+0.3260491$$

(4) model of New Brunswick

$$\frac{p_i}{1-p_i} = exp(-0.1352459-0.13266754Male_i+0.1205704Age_{30-39}+0.1325729Age_{40-49}+0.1432926Age_{50-59}+0.3260491$$

(5) model of Newfoundland and Labrador

$$\frac{p_i}{1-p_i} = exp(0.2752678 - 0.22552643 Male_i + 0.1205704 Age_{30-39} + 0.1325729 Age_{40-49} + 0.1432926 Age_{50-59} + 0.3260491 A$$

(6) model of Northern Canada

$$\frac{p_i}{1-p_i} = exp(-0.1717145 - 0.12441829 Male_i + 0.1205704 Age_{30-39} + 0.1325729 Age_{40-49} + 0.1432926 Age_{50-59} + 0.326049$$

(7) model of Nova Scotia

$$\frac{p_i}{1-p_i} = exp(0.1865177 - 0.20545102 Male_i + 0.1205704 Age_{30-39} + 0.1325729 Age_{40-49} + 0.1432926 Age_{50-59} + 0.3260491 A$$

(8) model of Ontario

$$\frac{p_i}{1-p_i} = exp(-0.1498536 - 0.12936326 Male_i + 0.1205704 Age_{30-39} + 0.1325729 Age_{40-49} + 0.1432926 Age_{50-59} + 0.326049$$

(9) model of Prince Edward Island

$$\frac{p_i}{1-p_i} = exp(0.1459720 - 0.19627950 Male_i + 0.1205704 Age_{30-39} + 0.1325729 Age_{40-49} + 0.1432926 Age_{50-59} + 0.3260491 A$$

(10) model of Quebec

$$\frac{p_i}{1-p_i} = exp(-0.3117680 - 0.09273796 Male_i + 0.1205704 Age_{30-39} + 0.1325729 Age_{40-49} + 0.1432926 Age_{50-59} + 0.326049$$

(11) model of Saskatchewan

$$\frac{p_i}{1-p_i} = exp(-1.5435324 + 0.18588925 Male_i + 0.1205704 Age_{30-39} + 0.1325729 Age_{40-49} + 0.1432926 Age_{50-59} + 0.326049$$

4 Results

After building the multilevel regression model, next step is post stratification. By using R package car (Fox and Weisberg (2019)), I applied the model to predict the proportions that each individual in census data may vote Liberal Party in 2019 Canadian federal election. I assumed the census data represents the overall Canadian qualified voters, and based on that assumption, the overall proportions that voters would like to vote Liberal equals the overall vote share or polling of Liberal in 2019 election, if every qualified voters participate in voting.

The table below shows the estimated vote share of each province level for Liberal party, if every qualified voters have voted in 2019 Canadian Federal election.

Table 1. Est. vote share per province |
province | vote share |
:——————— | :——: |
Alberta | 16.70% |
British Columbia | 28.66% |
Manitoba | 29.06% |
New Brunswick | 35.27% |
Newfoundland and Labrador | 44.04% |
Northern Canada | 34.08% |

Nova Scotia | 42.86% |
Ontario | 36.59% |
Prince Edward Island | 41.27% |
Quebec | 32.71% |
Saskatchewan | 13.81% |

From the Table 1, we can see that Newfoundland and Labrador, Nova Scotia and Prince Edward Island are the top 3 in terms of vote share for Liberal party, which are over 40%. The performance in Ontario, New Brunswick, Northern Canada and Quebec are not bad, which are between 30% and 40%. However, Liberal would win small proportion of popular vote in Saskatchewan and Alberta, which are both below 20%. Especially in Saskatchewan, the vote share of Liberal would be as low as 13.81%.

Then we calculate the overall popular vote share of Liberal through introducing the weighted average of population of each province to population of Canada. The formula is:

$$\hat{Y}_{PS} = \frac{\sum_j N_j \hat{y_j}}{\sum_j N_j}$$

The

$$\hat{Y}_{PS}$$

represents the expected overall winning rate of popular vote for Liberal. The

$$N_j$$

represents the weight of number of voters in province j to aggregate level of voters in Canada. The

$$\hat{y_j}$$

represents the expected winning rate of popular vote for Liberal in province j. And the j indicates which province it is.

Using R to sum the vote share of each province in Table 1, we can conclude that the predicted overall vote share for Liberal Party is about 31.80%, if the turnout rate is 100%, which means every qualified voters had voted in 2019 Canadian Federal Election.

5 Discussion

5.1 Discussion over Results

Through the analysis above, the technique I used, which is the multilevel logistics regression model with post-stratification, indicates two results. The first one is that Liberal might have won 31.80% of popular vote if every qualified voters had voted in 2019 Canadian Federal Election. The second is that Liberal might win a lot in Newfoundland and Labrador, Nova Scotia, Prince Edward Island and Ontario, under the same circumstance, which is 44.04%, 42.86%, 41.27% and 36.59% respectively.

As for the overall vote share of Liberal, Liberal actually won 33.10% of popular vote in the federal election last year, 1.3% higher than the estimated rate. That indicates that Liberal might win less if "everyone" have participated. That is to say, many people who missed voting last year are considered not in favor of Liberal Party.

As for the vote share of Liberal in province level, I expect that Ontario is the stronghold of Liberal Party, because Liberal might win a high rate of popular vote if "everyone" voted, and also Ontario has the largest population and the most federal election districts. And I expect that Alberta and Saskatchewan are the strongholds of Liberal's competitors, because the Table 1 shows low vote share of Liberal in these two provinces, which can be proved by the historical data of election results in these two provinces. The actual results of popular vote for Liberal in each province are showed in the following table.

Table 2. 2019 vote share per province |
province | vote share | Gains & losses |
:———————— | :———: | :————: |
Alberta | 13.10% | - 3.60% |
British Columbia | 26.10% | - 2.56% |
Manitoba | 26.20% | - 2.80% |
New Brunswick | 37.60% | + 2.33% |
Newfoundland and Labrador | 44.70% | + 0.66% |
Northern Canada | 34.08% | + 1.22% |
Nova Scotia | 41.10% | - 1.46% |
Ontario | 41.40% | + 4.81% |
Prince Edward Island | 43.60% | + 2.33% |
Quebec | 34.20% | + 1.49% |
Saskatchewan | 11.60% | - 2.21% |

The "+" symbol in column "Gains & losses" represents Liberal won more in the corresponding province in the actual 2019 election. And the "-" symbol is the opposite. Comparing the vote share in table 2 with table 1, I find that if the turnout rate is 100%, Liberal might win more popular vote in 5 provinces, Alberta, BC, Manitoba, Nova Scotia and Saskatchewan. Except for Nova Scotia, other 4 provinces are the traditional strongholds of its competitor, Conservative. If "everyone" had voted, Liberal might win less in other 6 provinces or territories. Especially in its No.1 traditional stronghold, Ontario, Liberal might lose 4.81% more with 100% turnout in the election. In other word, there is a higher participation rate of voters who are in favor of Liberal in its stronghold, compared to with other parties' supporters, which makes sense.

5.2 Weaknesses and Next Steps

There exist many weaknesses of this paper. First, it's the problem of data. 2019 CES survey data is fine to be less representative in a way. However, the census data might not be as representative as I expect. The aim of MRP model is trying to reduce the biases that less representative survey data makes with the step of post-stratification. That means we need a sufficiently representative data to represent the target population. As the census data I used is also sampled from the 2016 Canadian census data by Statistics Canada, it will increase the biases and reduce the representativeness anyway, no matter how fittable the sampling method they used is. Moreover, the census data I used were obtained through the 2016 Canadian census program held by Statistics Canada. The data can be considered out of time, even though I have made an assumption of no big difference about data of population from 2016 to 2019. The importance of data freshness can not be ignored.

Second, it's about modeling. I used a logistics regression model as the model of individual level. The response variable is a binary categorical variable, with value 1 and 0, i.e., voting Liberal or not. Other than the two-party mode of US federal election, there are more than seven political parties as candidates in Canadian federal election. Among those parties, 4 parties are competitive, which is reflected by the number of seats that these 4 parties have. So, simply distinguishing the vote intention to 2 categories seems not so reasonable.

Thus, for the future work, I will improve on these two aspects. First, I need to find a more representative dataset for year 2019. The questionnaire of 2021 Canadian census program has been published in July, 2020. Respondents will answer most of the questions based on their own situation in 2019. That will be really suitable to use as post-stratification data. Second, next time I can build a more complicated and appropriate individual-level model for the first step of MRP. For example, I can build a nominal logistics regression model with response variable that contains at least 3 categories. Using this model, I can define response variable as parties that people prefer to vote in 2019 federal election, which contains 8 categories (7 parties and others like independent candidates). Therefore, I can not only estimate the popular vote share of each party in different level, but also the federal election districts that each party might win. For the latter, I can estimate the vote share of each party in each Canadian federal election districts, and based on the rule of winner-take-all, estimate the number of districts that each party might win; in other words, the number of seats in the House of Commons to the Canadian Parliament. That is to say, I can predict which party

might be in power and which kind of government might be formed (majority or minority) after election, if the turnout rate were 100% in 2019 Canadian federal election.

Appendix

Code and data are available at: https://github.com/JessieZ32/2019_CA_election.

References

2016, Statistics Canada. n.d. *Census of Canada, 2016, Individuals File (Public-Use Microdata File). Statistics Canada (Producer). Using Chass (Distributor).* http://webapps6.ucalgary.ca/~landru/census/2001/cnfam01.html%20(accessed%20December%2012,%202020).

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression.* Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Hodgetts, Paul A., and Rohan Alexander. 2020. *CesR: Access the Ces Datasets a Little Easier.*

Larmarange, Joseph. 2020. *Labelled: Manipulating Labelled Data.* http://larmarange.github.io/labelled/.

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

Stephenson, Allison; Rubenson, Laura B; Harell. 2020. "2019 Canadian Election Study - Online Survey." *Harvard Dataverse, V1.* https://doi.org/10.7910/DVN/DUS88V.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.