Predict the 2019 Canadian election by busing MRP and Post-stratification method

Abstract

Due to the cost of time and money, it is often impossible to obtain all samples during surveys. In this case sample surveys will be a better way.If the samples are representative enough, the actual results will be accurately predicted.However, if the sampled data is not representative, whether it can accurately predict the true result is a question which needs to be considered. In this study, two data sets are selected for prediction experiments. We first build a multi-regression model on the collected data set of the 2019 Canadian Election Research (CES) telephone survey data set to understand how different age groups, different genders, different education levels and different regions affect interviewers' political intentions for the election. Since the data set is far from covering all the people, thus we need the 2016 Education Census to modify the predictive model by establishing a post-hierarchical model. After adjusting the response of the 2019 telephone survey data set through multi-regression and post-layering, results which seem to be more representative are obtained.

Introduction

In this project, using the data from CES by using MRP and post-stratification method to predict the 2019 Canadian us election. Since Canadian election is one the biggest event that people pay a lot of attention.

People would vote for different government based on the policies which benefits their life in many ways like social welfare, public health etc.

In this project MRP is a good method for data analysis, data can be found in Canadian Election Survey (CES) There are two predictors should be considered: demographic data and area information. Post-Stratification method need to use the census data to do the analysis, Canadian 2016 Education census is a good data set for this data analysis.

Data

Two data sets are used in this experiment.

The first one is the 2019 Canadian Election Study (CES) phone survey data set searched in the 2019 Canadian federal election survey. The CES survey data sets are publicly available from various repositories across the internet in various data types (Canadian Election Study, 2019; UBC, 2015; ODESI, 2020). This data set includes 4021 samples with 278 variables. These variables include basic information such as the voter's gender, age, education, and province.It also includes some questions about the election such as "Is there a party you are leaning towards". Despite the large sample size, the 4021 samples are far from being representative of the voting population.

In the data set, the proportion of male is 66.3% and the proportion of female is 31%. In terms of age range, 44.3% of the respondents were over 55 years old, 18.1% were 45-54 years old, 17.3% were 35-44 years old, and 14% were 25-34 years old During the period, only 6.4% of the respondents were aged 18-24. This gender and age range distribution is different from the population distribution of ordinary Canadian voters. Therefore, another data set needs to be used for correction to obtain voting results that can predict "everyone".

The other one we use here is Canadian 2016 Education census data set. This data set contains a survey of the gender, age, education and province in the 2016 census.

Since the census data we use here is from 2016,so we make an assumption that the demographic situation of 2016 is similar to that of 2019.

Method

In order to convert the data of the 2019 Canadian federal election into an accurate estimate of "everyone's" voting intention, the rich population information of Canada's 2016 Education Census is used.

Post-stratification is a popular method to correct the known differences between the sample and the target population. The core idea is to partition the population into the cell population and educational attributes according to various combinations, use the sample to estimate the response variable in each cell, and finally aggregate the unit estimates the proportion of the relative population of each cell to the national estimated weight. Use y to represent the result of interest, and the estimation after stratification is defined as:

$$\hat{\mathrm{y}} = \frac{\sum_{j=1}^{J} N_j \hat{y}_j}{\sum_{j=1}^{J} N_j}$$

where

$$\hat{y}_j$$

is the estimate of y in cell j, and

$$N_j$$

is the size of the jth cell in the population. Analogously, we can derive an estimate of y at any sub-population level s by

$$\hat{\mathrm{y}} = \frac{\sum_{j=1}^{J} N_j \hat{y}_j}{\sum_{j=1}^{J} N_j}$$

where

$$J_S$$

is the set of all cells that comprise s. As is readily apparent from the form of the post-stratification estimator, the key is to obtain accurate cell-level estimates and estimates of the cell sizes.

Model

When given a person with demographic information and the region he belongs to,the model we build here is used to estimate whether he will vote for the liberal party. Before we build the model,we need to do some data pre-processing to make the variables in the two data sets match each other. First of all,we load the packages needed.

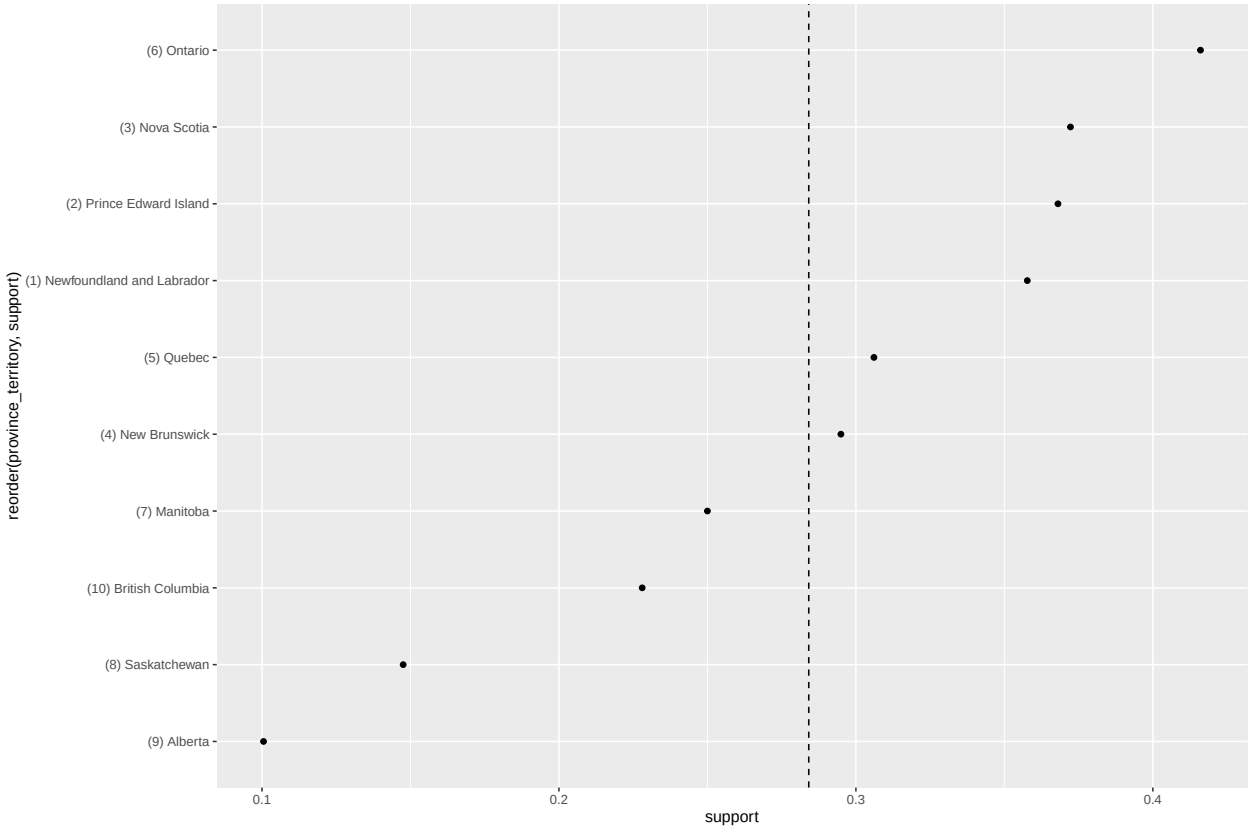
Then,we deal with the CES data set.We load the data and select variables we need and only keep the 'Liberal Party' and set other parties as 'others'.

In order to match the four corresponding variables in the two data sets, we deal with the four variables respectively.In the gender variable, only respondents whose gender options are male or female are considered. In terms of age, there is no option of 18-24 years old in 2016 census data set, so we use "15 years and over" minus "25 to 64" to get the specified group. As for the education variable, delete the options of "Don't know","Refused"and"Skipped".And we map the education variables one to one in the two data sets.Lastly, Only keep options with data in terms of province or territory.

Now, we finish data pre-processing.It's time to build the MRP model.Before we build the model,we take a quick look at the support rate of liberal party in the 2019 telephone interviews which is grouped by province or territory.

```
## # A tibble: 10 x 2
##    province_territory            support
##    <fct>                           <dbl>
##  1 (1) Newfoundland and Labrador   0.358
```

```
##  2 (2) Prince Edward Island        0.368
##  3 (3) Nova Scotia                 0.372
##  4 (4) New Brunswick               0.295
##  5 (5) Quebec                      0.306
##  6 (6) Ontario                     0.416
##  7 (7) Manitoba                    0.25
##  8 (8) Saskatchewan                0.148
##  9 (9) Alberta                     0.101
## 10 (10) British Columbia           0.228
```



First of all,we build the multilevel model which is given by:

$$\begin{aligned}&\Pr\left(Y_i=\text{liberal party}\right)\\=\quad&\text{logit}^{-1}\left(\alpha_0+\right.\\&+a_{j[i]}^{\text{geo}}+a_{j[i]}^{\text{edu}}+a_{j[i]}^{\text{Sex}}+\left.a_{j[i]}^{\text{age}}\right)\end{aligned}$$

```
## glmer(formula = vote.of.Liberal ~ (1 | Sex) + (1 | education) +
##     (1 | Age) + (1 | Geographic.name), data = temp2, family = binomial(link = "logit"))
## coef.est  coef.se
##    -1.05     0.24
##
## Error terms:
##  Groups          Name        Std.Dev.
##  Geographic.name (Intercept) 0.53
##  Age             (Intercept) 0.13
##  education       (Intercept) 0.25
##  Sex             (Intercept) 0.13
##  Residual                    1.00
## ---
```

```
## number of obs: 2658, groups: Geographic.name, 10; Age, 5; education, 5; Sex, 2
## AIC = 3105.1, DIC = 2991.4
## deviance = 3043.2
```

Analyze these coefficients, and then analyze each specific coefficient. The following table shows the weight coefficients of different age groups to liberal political parties. The 1-5 in the table indicate different age range groups.1 represents 15-24, 2 represents 25-24, 3 represents 35-44, 4 represents 45-54 and 5 means 55 and older.It can be concluded that the larger the age range of respondents, the higher chance he will support rate for liberal.

```
##   (Intercept)
## 1 -0.08206202
## 2 -0.08373613
## 3 -0.02116702
## 4  0.05226989
## 5  0.13721799
```

In terms of gender, observing the table below shows that the weight coefficient of women is positive, while that of men is negative(1 means male, 2 means female). Therefore, it can be concluded that women are more inclined to vote liberal.

```
##   (Intercept)
## 1 -0.08581594
## 2  0.08831207
```

Below are the weight coefficients shown by respondents with different education levels for the approval rate of liberal. Among them, 4 are college degrees and below and 5 are college degrees and above. We can conclude from the table that people with high education are more inclined to vote for liberal.

```
##   (Intercept)
## 1  0.02124957
## 2 -0.13718007
## 3 -0.27761707
## 4  0.16815174
## 5  0.23498814
```

Finally, on the CES data set, the weighted results of respondents' support for liberal parties in different regions, 1-10 correspond to the regions on the right.

```
##    (Intercept)                geographic
## 1  -0.99741073                   Alberta
## 2  -0.21996787          British Columbia
## 3  -0.11187446                  Manitoba
## 4   0.07961973             New Brunswick
## 5   0.33235797 Newfoundland and Labrador
## 6   0.42163330               Nova Scotia
## 7   0.60007927                   Ontario
## 8   0.41323982      Prince Edward Island
## 9   0.18032364                    Quebec
## 10 -0.65542379              Saskatchewan
```

The next analysis is based on the 2016 education census data set, and the weights should be adjusted on this data set. We need to weigh the regression effects by the relative population as shown in the census.

The first thing is to calculate 'cpercent', which is the frequency of each region's voting rate divided by the total number of people.Then, we get the support rate of different regions after revision. Since each post-stratifying cell or category is given as a percentage of the total population of a state by the Census. And our mega-poll is not a random sampling of each state, to get state-level outcomes in the proper ratios we need to weigh each cell by their percent of the province or territory population. After that, create a

vector for each cell, the specific operations are shown in the table below.

For the predicted value of each cell, scale it by uses the cpercent mentioned above.

Finally we get the table which if the support rate of different regions after revision.

```
## # A tibble: 10 x 3
##    Geographic.name pred.support geographic
##    <chr>                  <dbl> <chr>
##  1 1                      0.139 Alberta
##  2 10                    15.4   Saskatchewan
##  3 2                      6.22  British Columbia
##  4 3                      0.918 Manitoba
##  5 4                      0.831 New Brunswick
##  6 5                      1.26  Newfoundland and Labrador
##  7 6                      2.96  Nova Scotia
##  8 7                      0.476 Ontario
##  9 8                      0.443 Prince Edward Island
## 10 9                      0.579 Quebec
```

Summary:

In this study, two data sets were selected for prediction experiments. We first build a multi-regression model on the acquired data set 2019 Canadian Election Study (CES) phone survey data set to understand how different age groups, different genders, different education levels, and different regions affect the interviewer's political intentions for the election. However, since this data set is not representative, and the sample is not enough to represent the election intentions of the people across the country. Therefore, the 2016 Education Census was used to modify the prediction model through the establishment of a post-stratification model. After adjusting the response of the 2019 telephone survey data set through multi-regression and post-layering, relatively representative results were obtained. Therefore, it can be considered that modifying the data set through the MRP method can enable non-representative polls to not only predict election results, but also measure public opinion in a wide range of social, economic, and political fields.

Conclusions:

Election forecasts must not only be accurate, but also relevant, timely, and cost-effective. In this experiment, we used extremely unrepresentative and small data to construct a prediction that satisfies all these requirements. Although the data is collected on a proprietary public opinion survey platform, in principle people can collect these unrepresentative samples at a fraction of the cost of traditional survey design. In addition, the forecasts generated by these data are both relevant and timely because they can be updated faster and more regularly than standard election polls. Therefore, one of the main goals and main contributions of this article is to assess the extent to which accurate predictions are generated from unrepresentative samples. Due to the limited basic facts of election predictions, it is difficult to accurately determine the accuracy of our predictions.

Weakness & Next Steps:

The shortcomings are obvious. The census data set we use doesn't contain any suitable variables that can reflect the political intentions of voters. This point will become far-fetched when analyzing the data, and it is not convincing enough to prove that the data prediction modified by this model can represent the voting willingness of "everyone" across the country.

Therefore, if there is a chance to make further progress, I will find a more suitable data set to include some political variables.

Alternatively, the way we use here to deal with missing categories is by simply deleting them. While another way to reconstruct missing categories is by using the variation estimated by the model. Since each

individual's intercept within that group is actually pulled from a normal distribution of mean zero and the listed standard deviation. If we would like to know the total uncertainty about our knowledge in Northwest territories, Yukon and Nunavut, we actually want to sample that distribution and show the overall changes.

References:

[1]Wang W, Rothschild D, Goel S, et al. Forecasting elections with non-representative polls[J]. International Journal of Forecasting, 2015, 31(3): 980-991. [2]Stephenson, L., Harell, A., Rubenson, D., & Loewen, P. (2020, May 01). 2019 Canadian Election Study - Online Survey. Retrieved December 10, 2020, from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FDUS88V [3]Government of Canada, S. (2017, November 27). Education Highlight Tables, 2016 Census. Retrieved December 10, 2020, from https://www12.statcan.gc.ca/census-recensement/2016/dp- pd/hlt-fst/edu-sco/index-eng.cfm