STA304 Final Project

Chris Shen

2020/12/16

The link to the GitHub repo: . https://github.com/shenjin4/STA304finalproject

Abstract

This report introduces the complete process of conducting a post-stratification analysis using a survey data to build the model and correct the result with the census data. The question of interest is the investigate how the result of 2019 Canadian Election would change if the entire qualified population has voted. Survey data from CES through web quesionare are collected and census data are simulated from summary tables. Separate logistic regression models are combined to predict the supporting parties. The results show that the proportion of people who support the Conservative party might increase based on the post-stratification prediction, leading to a reverse in election result. To improve the analysis, more data could be collected with useful features to enhance the prediction accuracy.

Keywords: MRP, Canadian Election, Simulated Census Data

Introduction

Survey results are facing the problem caused by differences between sample population and target population. Multilevel regression with poststratification [3][2] is used to address this issue by adjusting the pool results weighted by charactersitics of the target population. It is commonly used when there is a survey targeting national population. However, the sampling process might be biased since most of the survey are volunteer-based, or some target proportion could not be reached by the design of the sampling method. In this case, target population attributes could be estimated by census data, which is mandatory and hence much more unbiased.

As demonstrated in Wang's paper [4], the survey data from Xbox players are used to predict election results. These two populations seem uncorrelated and not typical survey data, which fully illustrates the advantages of using MRP given the limitation in data collection. The author of the paper points out two main reasons that cause the pool to fail: high non-response rate and cheap but unreliable data collection. Similar to Jack Bailey's idea, instead of analyzing the result of US election, this report conducts the MRP process for 2019 Canadian election.

The 2019 Canadian Federal Election was a close race, the country seemed widely divided in their votes. The majority people in coastal provinces voted for the liberals while many people in the central provinces voted for the conservatives. Such divided result raised a question: would the result be different if there were more people voted? Because Canada is

an immigration country, the population consists different races that comes from different regions, and voters from different cultural backgrounds might have different opinions about the election. Language is the best tool to distinguish voters from each region. To better construct a model that represent the Canadian population according to its demographic distribution, we should build a multilevel regression with poststratification. As a result, a prediction can be made about the results of 2019 Canadian federal election if everyone had voted.

In this report, the Canadian Election Study, 2019, Online Survey data [1] is included as the modeling dataset. The CES data include some demographics of the individuals who provided responds, which could be used to do the post-stratification on population dataset. More importantly, it includes the opinions of Canadians during and after the 2019 federal election. In the Methodology Section, the process of data simulation from census pivot tables and the detailed modeling steps are performed. The output of the analysis will be displayed in the Results Section. Last but not least, the Discussion will include the main conclusions, limitations, and future improvements of this project.

Methodology

In this section, the two main datasets used in this study are introduced and key properties are explained in details. Also, the variables used in modeling are selected in the Model Section.

Data

This study focuses on two datasets: Canadian Election Study (CES) 1 data, and (2016 Canadian Census data)[https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/dt-td/index-eng.cfm]. The CES data has individual-level records with some basic demographic properties along with the survey answers to a series of questions related to 2016 Canadian election. The web survey result dataset is chosen, which has 37,822 records in total. The variables kept are: language learnt as a child, age, gender, and the target variables.

Age and gender variables take intgers. For gender, 1 represents male, 2 represents female and 3 represents others. In this case, language means mother tongue, or language learnt as a child. And it is a categorical varaibles that are classified into 4 groups: “Eng” - English, “Fre” - French, “EngFre” - both English and French, and “Neither” - other languages.

The target variables are prepared based on two questions in the original questionare, namely `pes19_votechoice2019` and `cps19_votechoice`; the questions were framed as “Which party did you vote for?” and “Which party do you think you will vote for?” respectively. The choices are: Liberal Party, Conservative Party, NDP, Bloc Qu, Green Party, People’s Party, Another Party, and Do not know / Prefer not to answer. The response rate of `pes19_votechoice` is relatively low, there are only 8,607 valid answers out of 37,822 total records. In comparison, `cps19_votechoice` has 31,564 valid records with 4,908 people answered “Do not know / Prefer not to answer”. All the invalid records includes “Do not know / Prefer not to answer” are excluded from the analysis since it does not make

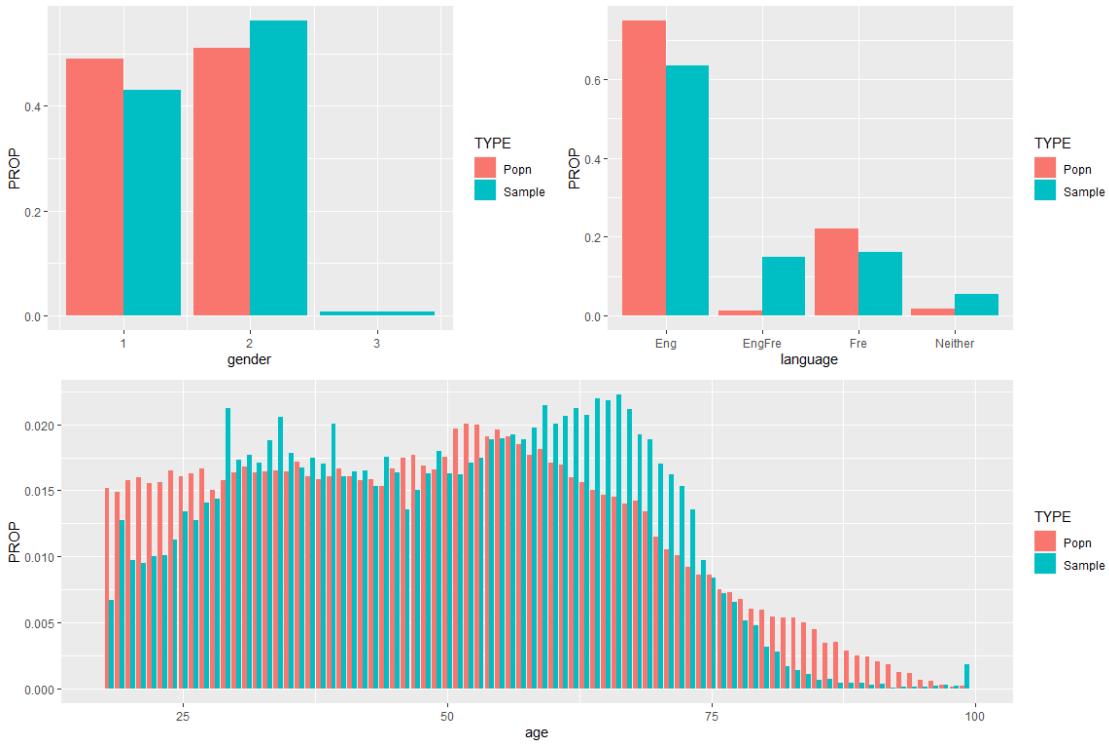
sense to impute the target feature. Also, repsondents who are not Canadian citizens or less than 18 years old are excluded from the study since they are not eligible to vote.

After the data exclusion, the target variable is split into 4 indicator varaibles: `y_liberal`, `y_conservative`, `y_ndp`, and `y_other`, which indicates the supporting party of the respondent - Liberal, Conservative, NDP and all other parties. Originally, the question could be formed as a multinomial classification problem. With these four variables defined, there are four independent 2-class classification problems. The method used is further discussed in the Model Section.

	x
y_liberal	0.3384497
y_conservative	0.3212040
y_ndp	0.1675740
y_other	0.1727723

From the table above, the proportion of support for each party in the sample data is displayed. Liberal and Conservative Party takes a third of samples respectively and NDP and other parties make up the rest.

The second main dataset used is the 2016 census data. From the website, all the high-level pivot tables are available. However, there is no individual level cesus data in consideration of privacy issue. Hence, the post-stratification dataset are simulated based on the proportions derived from different pivot tables. The fields extracted align with the CES dataset - age (from 18 to 99), gender (male or female) and language - mother tongue. Note that every feature is simulated independently from other features; that is, the interaction effect between different variables are not considered in this case.



From the plots above, there are significant differences between sample and population distributions in various aspects. Firstly, in the survey, the gender has three classes: man, woman and other. However, the census data only has two classes - man and woman. However, since “Other” only accounts for 0.77% of the survey data as shown in the table below, the training model will be the original survey data and the post-stratification will be conducted without the type “Other”. This is the similar problem mentioned in the paper during lecture. 5

Moreover, even for “Male” and “Female” categories, there is a gap between the proportions. In the population, the gender distribution is almost even for male and female. But the survey data is a much smaller base and hence more unbalanced - there are significantly more females conducted the survey.

gender	Popn	Sample
1	0.48998	0.4296653
2	0.51002	0.5626427
3	0.00000	0.0076920

Similarly, we look at the language variable. In the population dataset, there are more people who only has one language as their mother tongue, either English or French. However, in the survey responses, a much higher proportion is assigned to people who learnt both English and French during childhood. There are more official language minority in the survey data than in target population.

language	Popn	Sample

Eng	0.74968	0.6355239
EngFre	0.01238	0.1490288
Fre	0.22033	0.1614625
Neither	0.01761	0.0539848

The distribution of age differs in two datasets. In the population, the distribution of age is almost uniform from 18 to 50, and has a small peak at 50 then starts to decrease. On the other hand, the survey data has a less stable distribution, with lower proportion for 18-30 young people, and higher proportion within 55-75 range.

Model

As discussed in the Data Section, four indicator variables are created to represent the resondent's political opinion/preference. A logistic regression is performed for each variable using the explanatory variables of `age`, `language` and `gender` and the interaction of `gender` and `age`. After the probability of each target variable is computed, the overall choice of supporting party will be the largest probability among them. Rather than using a multinomial logistic regression package, computing 4 logistic regression models and combining them may be more robust and gives a similar performance. The formula are shown below:

$$log(p^{lib}/(1-p^{lib})) = \beta_0^{lib} + \beta_1^{lib}x_{age} + \beta_2^{lib}x_{gender} + \beta_3^{lib}x_{lang} + \beta_4^{lib}x_{gender}x_{age} + \epsilon^{lib}$$

$$log(p^{con}/(1-p^{con}))$$
$$= \beta_0^{con} + \beta_1^{con}x_{age} + \beta_2^{con}x_{gender} + \beta_3^{con}x_{lang} + \beta_4^{con}x_{gender}x_{age} + \epsilon^{con}$$

$$log(p^{ndp}/(1-p^{ndp}))$$
$$= \beta_0^{ndp} + \beta_1^{ndp}x_{age} + \beta_2^{ndp}x_{gender} + \beta_3^{ndp}x_{lang} + \beta_4^{ndp}x_{gender}x_{age} + \epsilon^{ndp}$$

$$log(p^{other}/(1-p^{other}))$$
$$= \beta_0^{other} + \beta_1^{other}x_{age} + \beta_2^{other}x_{gender} + \beta_3^{other}x_{lang} + \beta_4^{other}x_{gender}x_{age} + \epsilon^{other}$$

$$party^{pred} = argmax\{p^{lib}, p^{con}, p^{ndp}, p^{other}\}$$

Note that x_{gender} and x_{lang} represents the indicator variables and corresponds to more than one parameter β.

We chose this type of model since each respondent could only choose one of: Liberal, Conservative, NDP or other. If simple logistic regression is used, there might be cases where one response is predicted as two classes or no classes. In order to avoid that, the party with the largest probability is selected as the final output.

Results

The results of the logistic regression models are shown below and overall accuracy of the model is displayed:

term	estimate	std.error	statistic	p.value
(Intercept)	-0.8205497	0.0655689	-12.514306	0.0000000
age	0.0013148	0.0011693	1.124457	0.2608192
as.factor(gender)2	-0.3662516	0.0824553	-4.441821	0.0000089
as.factor(gender)3	-1.2967751	0.4270329	-3.036710	0.0023918
languageEngFre	0.3404312	0.0357181	9.531043	0.0000000
languageFre	-0.0711073	0.0356355	-1.995404	0.0459988
languageNeither	0.2840261	0.0547323	5.189367	0.0000002
age:as.factor(gender)2	0.0088591	0.0015481	5.722643	0.0000000
age:as.factor(gender)3	0.0218371	0.0091656	2.382508	0.0171951
term	estimate	std.error	statistic	p.value
(Intercept)	-0.8359279	0.0658899	-12.6867431	0.0000000
age	0.0098229	0.0011687	8.4053092	0.0000000
as.factor(gender)2	-0.4032441	0.0852957	-4.7276038	0.0000023
as.factor(gender)3	-0.9669468	0.4503012	-2.1473335	0.0317667
languageEngFre	-0.5559484	0.0397098	-14.0002789	0.0000000
languageFre	-1.0196498	0.0414092	-24.6237221	0.0000000
languageNeither	-0.0649451	0.0557030	-1.1659170	0.2436480
age:as.factor(gender)2	0.0016757	0.0015901	1.0538612	0.2919465
age:as.factor(gender)3	0.0064575	0.0097563	0.6618766	0.5080503
term	estimate	std.error	statistic	p.value
(Intercept)	-0.9011216	0.0870400	-10.352963	0.0000000
age	-0.0182047	0.0016676	-10.916625	0.0000000
as.factor(gender)2	0.8221031	0.1040369	7.902031	0.0000000
as.factor(gender)3	2.0584955	0.4056276	5.074840	0.0000004
languageEngFre	-0.1414253	0.0450827	-3.137018	0.0017068
languageFre	-0.6379827	0.0535623	-11.911038	0.0000000
languageNeither	-0.3072204	0.0775185	-3.963187	0.0000740
age:as.factor(gender)2	-0.0085378	0.0021035	-4.058952	0.0000493
age:as.factor(gender)3	-0.0229575	0.0101789	-2.255407	0.0241078
term	estimate	std.error	statistic	p.value
(Intercept)	-1.8172229	0.0545205	-33.331023	0.0000000
age	-0.0037926	0.0010026	-3.782598	0.0001552
languageEngFre	0.4472702	0.0468071	9.555609	0.0000000
languageFre	1.6226329	0.0379912	42.710809	0.0000000
languageNeither	-0.0981062	0.0851403	-1.152289	0.2492025

```
## [1] 0.3875874
```

pred	prop
1	0.2484633
2	0.5206350
3	0.0709143
4	0.1599874

Note that in the model of predicting `y_other`, the variable `gender` and its interaction have p-values greater than 0.1. Hence, the variables related to `gender` is removed from the last model.

By comparing the coefficient of the models, it could be noticed that people who learnt both English and French in their childhood are more likely to choose Liberal party than Conservative and all other parties. Women and the third type of gender are more likely to choose NDP than Liberal and Conservative. Liberal and Conservative supports have a positive relationship with the respondent's age while NDP and other parties have negative relationship.

Overall, the variables selected are quite significant by looking at the p-values. However, since there are too few features collected, the accuracy is only 38.7% comparing to the real party chosen. The predicted classes are shown in the table above - there is a siginificant difference between proportion predicted and actual proportions due to lack of explanatory variables.

Then we applied the same models on the population dataset. And the proportion of support is summarized in the table below.

pred	prop
1	0.08990
2	0.61961
3	0.07853
4	0.21196

Comparing to the prediction result using the CES data, the proportion of supporting the Conservative Party has increased to 61.9% from 52%. The post-stratification prediction of Liberal Party largely decreased to less than 10%.

Discussion

Summary

In this report, the technique of post-stratification has been used to examine the potential outcome of the election if everyone has voted. Firstly, the individual-level survey data is collected with some demographic features that could act as explanatory variables in the model. The census data is simulated from summary pivot tables and it is used as the post-

straification population. Only qualified people are left in the census data. In the modeling stage, four logistic regression models are built for each party and combined to generated the predicted party overall. Although the model accuracy is low due to too few explanatory variables, there are still insights that could be drawn from the results by comparing prior prediction and post-stratification prediction result to view the effect to shifting distributions.

Conclusion

Overall, it could be concluded that the result of the election could have changed if everyone has voted. However, the exact result still need more statistical evidence to find out. We can simply conclude that, the proportion of supporting the Liberal Party might have decreased based on the result of our models. And the proportion of people who support Conservative party might have increased.

Based on the real election result of 2019 (CBC), the Liberal Party won with minority votes, that is, the difference between votes for the Liberal and Conservative parties are really small. In this case, it is reasonable to say that if the number of votes changed as the model predicted, the result might have altered. We could not simply compare the post-stratification prediction with the actual proportion of support as indicated in the table in Data Section.

Weaknesses & Next Steps

One obvious limitation of this model is its low accuracy in predicting preferred party. The reason for that is the lack of sufficient number of independent variables to be included in the model. It could be seen that the variables selected are all statistically significant and are providing some prediction power to some extend. However, we need more demographic features that could better reflect and explain people's political preferences.

In future analysis, some other features could be extracted from the CES survey data as well as corresponding features in the census data. Other potential features that might affect one's supporting party might be: level of education, income level, household size and province.

Another limitation occurs in the data preparation phase. Since there is no individual-level cesus data, only picot tables with summary counts are used to generate the post-stratification population. In the simulation process, no interaction effects between variables are considered. That is, each variables are simulated based on their own proportions and not dependent on any other variables. The interaction might also affect the distribution and hence predicted results.

To address the second issue, one solution is to find a way to obtain individual census data without revealing some personal information. The other measure is to consider the interaction effect in the pivot tables. However, this is a very time-consuming process since with increasing number of explanatory variables and increasing levels in each variable, the number of interactions to consider increases exponentially.

Last but not least, the election was in 2019; however, the census data used in the analysis is collected in 2016. During the three years, the distribution of data might have changed thus cause some error in the post-stratidfication modeling.

Overall, the performance of the analysis in this report could be improved by collecting more useful data as well as collecting data with higher quality.

Reference

[1] Hodgetts, P. A., & Alexander, R. (2020, September 3). cesR: An R package for the Canadian Election Study. https://doi.org/10.31235/osf.io/a29h8

[2] Gelman, Andrew; Lax, Jeffrey; Phillips, Justin; Gabry, Jonah; Trangucci, Robert (28 August 2018). "Using Multilevel Regression and Poststratification to Estimate Dynamic Public Opinion" (PDF): 1–3.

[3] Marnie Downes, Lyle C Gurrin, Dallas R English, Jane Pirkis, Dianne Currier, Matthew J Spittal, John B Carlin, Multilevel Regression and Poststratification: A Modeling Approach to Estimating Population Quantities From Highly Selected Survey Samples, American Journal of Epidemiology, Volume 187, Issue 8, August 2018, Pages 1780–1790, https://doi.org/10.1093/aje/kwy070

[4] Wang, Wei; Rothschild, David; Goel, Sharad; Gelman, Andrew (2015). "Forecasting elections with non-representative polls" (PDF). International Journal of Forecasting. 31 (3): 980–991. doi:10.1016/j.ijforecast.2014.06.001.

[5] Kennedy, Lauren; Khannay, Katharine; Simpsonz, Daniel; Gelman, Andrew (2020). "Using sex and gender in survey adjustment" (PDF). arXiv:2009.14401