Towards a Estimation of Supporting Rate for Liberal Party

Zhixing Hong

12/15/2020

Abstract

The purpose of this paper is to build and train a logistic regression model to calculate the 2019 Canadian election result. We also simulate the supporting rate for the Liberal Party, when assuming all the qualified Canadians all voted during the campaign period. Based on the survey data obtained from the Canadian Election Study (online survey) and the census data from 2017 General Social Study data, I applied the post stratification method to calculate the supporting rate for liberal party in 2019 Election. With 6 different categorical variables in the logistic model, we get a supporting rate close to the real supporting rate, indicating that the model is applicable.

Keywords: 2019 Canadian Election, Canadian Election Study, General Social Survey(2017), Post-Stratification, Logistic Modeling, Logistic Regression, Supervised Learning, Parametric Statistical Learning, Confusion Table

code and data supporting this analysis is available at: https://github.com/neverknowhen/2019-Canadian-Election-Study

Introduction

Not only play an important role in helping the journalists and citizens understand the meaning of the campaign and the election, but election polls also overhang for their ability to predict the supporting rate for different parties and their winning rate. As the election survey cannot cover the whole population, a statistical method is required for stimulating the result for the entire population. The post-stratification is useful in a dual-frame survey, as we need to connect the election survey with the census data. The dual frame survey was first studied by Hartley (1962, 1974) based on post-stratified samples under the following classic setting. The same setting was also used by Fuller and Burmeister (1972), Skinner and Rao (1996), Lohr and Rao (2000, 2006), and Rao and Wu (2010b), among others. [Wu & Thompson, 2020]

During the 2019 Canadian Election, Liberal Party did win the final seat. However, the supporting rate for the Liberal Party was lower than the Conservative Party. Nevertheless, compared to the 2015 Canadian election, the supporting rate for the Liberal Party also decreased. The reasons behind this might be complicated as involving multiple aspects. The importance of predicting the supporting rate for the Liberal Party was shown. Analyzing the survey data from the campaign period and connecting it with census data could help to estimate the supporting rate for the Liberal Party. Therefore, this study is focusing on estimating the supporting rate for the Liberal Party during the 2019 Canada Election, assuming all the eligible Canadian participated.

There are two different datasets used in the paper to build the logistic regression model. The trained logistic regression model is then used for calculating the supporting rate of the Liberal Party in the 2019 Canadian Election. In the Methodology section (Section 2), the details about the data, the logistic regression model and post-stratification are discussed. In the result section (Section 3), I discussed the supporting rate of the

2019 Liberal Party obtained by the model. The inference of the data and potential improvement for the model is mention in the Discussion Section(Section 4.)

Methodology

Data

General Information

There are two different datasets are used in the paper.

The survey dataset is a dataset obtained by Canadian Election Study (Short as CES) during the election campaign period time. 37822 respondents took the survey between September 13th to October 21st, 2019. The daily sample size is approximately 800 to 850 people; while the lase five days, there is a huge increase in the daily sample size, around 1600 respondent on average, per day. The survey is distributed online, and due to the design of the questionnaire, respondents are able to answer certain questions based on the choice for specific questions. There are 620 variables in the raw dataset, including some basic information about the respondent, the respondent's opinion and thought about political issues and the respondent's vote decision. In order to make the prediction of the support rate for the Liberal Party when all eligible voters had been voted, I select some of the variables and clean those variables.

For the data used to represent the census, data from the 2017 General Social Study (Short as GSS) is reformatted and cleaned. The 2017 GSS is a sample survey with a cross-sectional design. The target population is all non-institutionalized persons 15 years of age and older, living in the 10 provinces of Canada. The GSS was conducted by telephone surveys to collect the response from people all over Canada. There are 460 questions included in the survey, which could be classified into 14 different categories. That includes the entry component, family origins, conjugal history, children of respondents, and so on. The variables that I select are similar to the variables selected from the survey dataset.

Data Cleaning and Data Preparation

The variables selected are: `province`, `gender`, `age`, `education`, `household_size`, `importance of religion`, `marital status`, and one more variable `vote_combine` from the sample data is also included. All the variables, except the `vote_choice`, are used as the predictor for the logistic regression model.

As the variable `province`, in the sample and census data indicates the living place of the respondents, I left the variable unchanged. In both datasets, I select the respondents at 18 years old or older, since the qualified voters in the Canadian Election must older or at 18 years old. After selecting the qualified voters, I create `age group` to replace `age`. This variable divides people into age groups that have similar characteristics and are more significant for the model. For each small age group, the difference between the upper bound and the lower bound is 5 years. In the CES data, `gender` is included, while `sex` is used in the GSS data. As suggested by Kennedy that there is a difference between the definition of `sex` and `gender`, ignoring the difference is algorithmic injustice. [Kennedy, Khanna, Simpson & Gelman, 2020] Therefore, followed the sociological perspective, I combine the variable level `a woman` and `Other (e.g. Trans, non-binary, two-spirit, genderqueer)` together in the survey data. Then I renamed the predictor as `sex`. For the variable `education`, it originally had 12 levels in the survey data set and 8 levels in the census data. To match this categorical variable, I combined several groups. The final `education` takes 6 levels. The `religion` is also an important feature for people in making voting choices during the election. Due to the limitation of the datasets, I chose the `religion_importance` as the predictor variable. `Marital status` in the survey data is kept the same as given. I reformatted that predictor in the `census data`, which could be paired with the survey data.

As the voting process has different stages during the 2019 Canadian Election, the vote choice in the survey data is not contained in the same variable. As there are people who voted in advance, people who willing and likely to vote during the voting period and also there are people who expressed that they 'will not vote or not willing to vote'. Therefore, to have a variable that combining all different people's voting choices is crucial. Hence, `vote_combine` combined all the voting choices for all the respondents who took the survey. As the goal of the paper is to simulate the real supporting rate for Liberal PArty, if every eligible voter voted during the 2019 election, I add one more variable `vote_liberal`. This is a binary variable that demonstrates whether the people voted for the Liberal Party during the Advance voting, or the respondent had the intention to vote for the Liberal Party.

Here is the overview of the two datasets. The detailed information for the two datasets is in Appendix A.

Table 1: Cleaned Census Data (from 2017 GSS)

province	sex	age_group	education	religion	marital	feeling_life
Quebec	2	(48,53]	High school diploma or certificate	Somewhat important	Never Married	Very satisfied
Manitoba	1	(48,53]	University certificate or diploma below the bachelor's level	Don't know/ Prefer not to answer	Married	Very satisfied
Ontario	2	(63,68]	Bachelor's degree	Very important	Married	Very satisfied
Alberta	2	(78,83]	High school diploma or certificate	Not important at all	Married	Very satisfied
Quebec	1	(23,28]	University certificate or diploma below the bachelor's level	Not important at all	Living with a partner	Very satisfied
Quebec	2	(58,63]	High school diploma or certificate	Very important	Married	Very satisfied

Table 2: Cleaned Survey Data (from 2019 CES)

province	sex	age_group	education	religion	marital	vote_combine
Quebec	2	(28,33]	University certificate, diploma or degree above	Don't know/ Prefer not to answer	Don't know/ Prefer not to answer	Green Party
Ontario	2	[18,23]	University certificate or diploma below the bachelor's level	Somewhat important	Never Married	Liberal Party
Ontario	1	[18,23]	University certificate or diploma below the bachelor's level	Somewhat important	Never Married	Conservative Party
Ontario	2	[18,23]	University certificate or diploma below the bachelor's level	Not important at all	Never Married	Liberal Party
Ontario	2	[18,23]	University certificate or diploma below the bachelor's level	Not very important	Don't know/ Prefer not to answer	Liberal Party
Ontario	1	[18,23]	High school diploma or certificate	Somewhat important	Never Married	Liberal Party

Strength and Limitation for the Survey and Data

There are several strengths for the survey data and census. First, the GSS survey took stratification as its sampling method. Based on the province information, the data is divided into strata, providing an unbiased sample. For the CES data, the dataset provided users with several efficient ways to remove the non-response problem. As the CES was conducted by an online survey, there is the time recorded for the respondent as they opened the survey and finished the survey. Based on the time given, I am notified by the inattentive respondents and speeders.

For the GSS, many missing values in the data were imputed as "Not Stated" or "Don't Know". While the assumption under which these imputations were conducted was probably reasonable, it may be able to add more useful and important information to the census data. Moreover, the sample size for this survey is relatively small comparing to the survey dataset. Therefore we have to use `weight_person` to estimate the number of people who fall in this category, which `weight_person` was calculated based on the census data on 2016 Canada Census. What is also worth mentioning is that CES was conducted online. Therefore, this makes a bias in the sample. All the respondents were defaulted to be the ones who had the access to the internet and who were able to finish the survey through the computer.

Model

Model Specifies

A logistic quasi-binomial regression model is employed to predict the supporting rate for the Liberal Party in the 2019 Canadian Election when all eligible Canadian voted.

As `survey data` is obtained from the survey conducted by CES, the sampling method of the survey has been taken into consideration. This is beneficial to reduce the errors to avoid inaccurate results when raining the logistic regression model. The finite population is specified as the total population that has been sampled, which is the number of Canadians who are 18 years old or older. [Wu & Thompson, 2020] Since all the predictors in the model are categorical, the quasi-binomial model could provide more accurate results compared to the binomial model. [Sinha, Das & Mukhoti, 2008]

After finalized the designed method, 6 different variables used in the model, `age_group`, `province`, `education`, `sex`, `marital_status` and `religion`. These 6 variables represent the voter's social and economic status, which is a miniature for the respondents' thoughts on the political issues and federal government. As all the variables included in the logistic model are categorical, `family = quasibinomial` is used in the logistic model training process. The logistic regression model used can be expressed as follows:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 gender_{male} + \sum_{i=2}^{13} \beta_i province_i + \sum_{j=14}^{29} \beta_j age\_group_j +$$
$$\sum_{l=30}^{34} \beta_l edcation_l + \sum_{k=35}^{39} \beta_k religion_k + \sum_{q=40}^{45} \beta_q marital\_status + \epsilon$$

Where p represents the proportion of voters who would like to vote for Liberal Party. β_0 represents the intercept of the logistic regression model. Additionally, $\beta_m (m \in [1, 45])$ represents the slope of each different variable, or each levels of the variables. For example, β_8 is the coefficient for the people from Ontario. The coefficients in the logistic regression model are not able to show the relationship between the variable and probability vote for Liberal Party directly, only if after the transformation. By applying the transformation on the coefficients, the general relationship is decreasing the log odds, increasing the real transformed coefficients. `R` software is used to build the logistic model and compute the model result.

Additional Information

As the aim of this paper to estimate the supporting rate for Liberal Party when all eligible Canadian have voted, the accuracy of the model is worth validating. I split the survey data into two, train and test datasets, with a ratio of 3:1. The training data is used to train the logistic quasi-binomial model. Using the training dataset, the coefficients before each variable is calculated using the parametric method. The test dataset is used to test the accuracy of the model. Based on the predictor values of each observation in the test dataset, the model will give the calculated result of each observation whether they would vote for the Liberal party or not. By comparing the calculated result with the original vote decision, the accuracy of the model is obtained. The accuracy of the model is conducive to deciding whether improvement with the logistic quasi-binomial model is required. [James, Witten, Hastie, & Tibshirani, 2017]

Results

Here, the β_0 to β_k represents the 'estimated' column of the table above, they are just coefficients of each variable, and according to the result of table3, it shows that most variables were negatively related to supporting Donald Trump in the 2020 election.

The detailed information for the logistic quasi-binomial regression model can be view in Appendix B.

As the coefficients of the logistic model is the log-odds, I transferred the estimated value of each β into the normal standard. Followed that, the new estimates values are used together with the census data to calculate the supporting rate for Liberal Party during 2019 Canadian Election, when all qualified Canadian participated in the voting. The obtained supporting rate for the Liberal Party = 0.363. The 95% confidence interval is $[0.3628, 0.3632]$. Here, to be more specified, 0 represents the people who would not like to vote for Liberal Party and 1 represents to those who support the Liberal Party. As 0.364 is more closer to 0, that the rate for supporting the Liberal Party is not very high.

Liberal_Party_Supporting_Rate
0.3636257

The estimated result, `prediction_result = 0.363` is a bit high than the real supporting rate during the 2019 Election. The real supporting rate for Liberal party is 34.3%. All the results are based off the post-stratification analysis of the proportion of voter leans Liberal Party. The logistic quasi-binomial model works with the survey method, with 6 categorical variables, `province`, `age_group`, `sex`, `education`, `marital`, `religious`.

Discussion

Summary

During the election period, companies from different fields and journalists would try to estimate the supporting rate for each party. In this paper, I build one logistic regression model with 6 categorical predictors, based on the CES survey data. After the logistic regression model is trained by the survey data, the census data is used. Together with the post stratification method, I estimate the supporting rate for the Liberal Party, assuming all the qualified Canadians voted during the 2019 Election. The estimated supporting rate for the Liberal Party = 0.363 .

Conclusions

To estimate the supporting rate for the Liberal Party when all eligible Canadian voted, I chose 6 predictors from a different perspective. At the beginning of the variable choosing process, there is one more predictor `household_size`. However, due to the subtle relationship between the response variable, this predictor is removed from the model. The other variable, `the satisfaction with the federal government` is also one intriguing variable. As the name of the variable indicates, this variable has a strong relationship with the response variable `vote_liberal`. The accuracy of predicting the test dataset is increasing significantly after adding this variable to the logistic regression model. However, as there is no matching variable in the GSS dataset, this variable could only be removed. Therefore, the final model is constructed by `province`, `age_group`, `sex`, `education`, `religion` and `marital`.

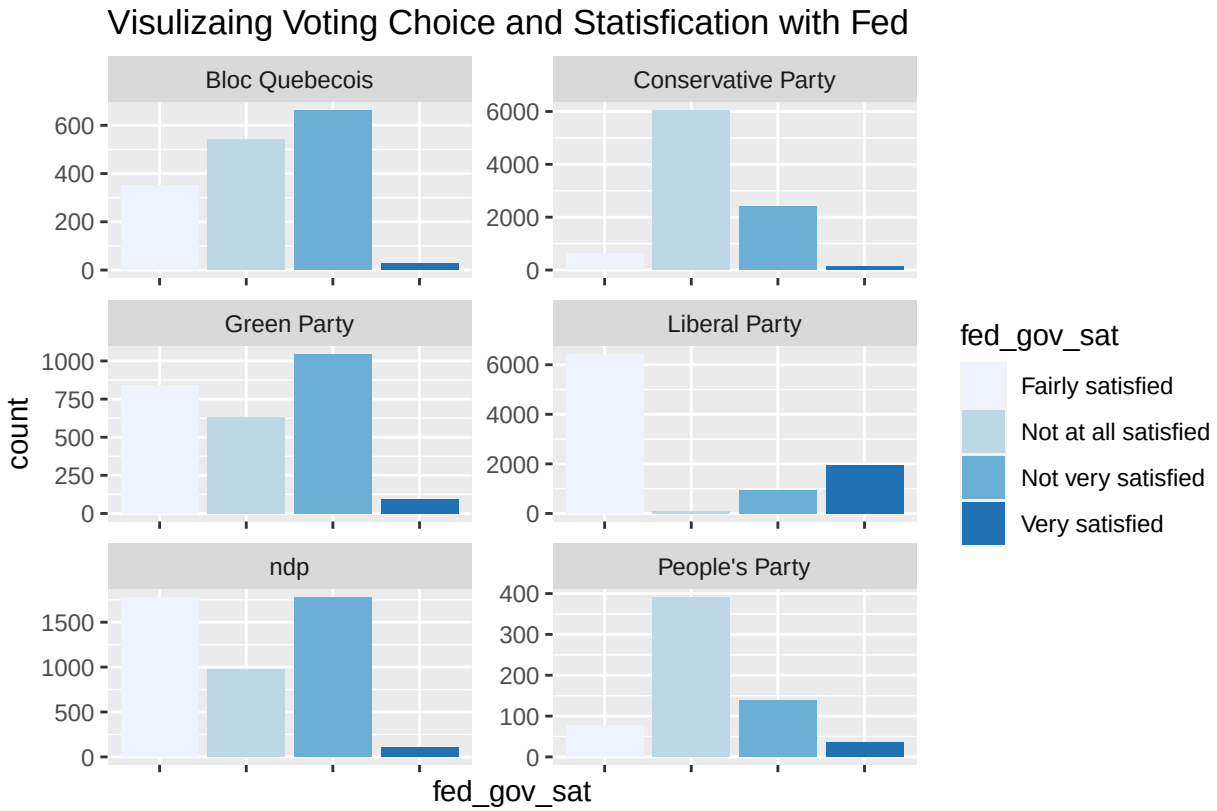


Figure 1

The supporting rate for the Liberal Party was actually lower than that for the Conservative Party during the 2019 Canada Election. The real supporting rate for Liberal party is at 0.331, while that for the Conservative Party is 0.343. Based on our estimation, the 95% confidence interval is $[0.3628, 0.3632]$. All these estimated value is significantly higher than the real value. However, it is important to recall that winning party during the 2019 Canada Election was actually the Liberal Party. Therefore, it is reasonable to argue that the model can not only be used for estimating the supporting rate, but also is applicable for predicting the winning party (different from U.S election, the number of parties involved leading to the winning probability should not compare with 0.5).

Weakness and Next Steps

There are some limitations we found during the research. Firstly, it is the problem with the census data. Though GSS data has a weight variable with allowed me use as the weight during the post stratification process. The weight is still not accurate, as the weight is calculated based on the 2016 Canada Census data.

There is three year gap between the survey data and census data. This adds errors to the model and influced the estimated value. Secondly, specificity of the logistic model is not as high as ideal. The model tends to mark more observations as not vote for Liberal, while those people's original response was vote for the Liberal Party. Though the overall accuracy for the model is above 70%, the specificity need to be improved.

Table 4: Result of Confusion Matrix

	x
Sensitivity	0.7369284
Specificity	0.6031746
Pos Pred Value	0.9136354
Neg Pred Value	0.2869836
Precision	0.9136354
Recall	0.7369284
F1	0.8158229
Prevalence	0.8506693
Detection Rate	0.6268823
Detection Prevalence	0.6861405
Balanced Accuracy	0.6700515

It is hoped that in the future work, the problem of the census data could be resolved. As more accurate the census data, more precise the model would estimate the supporting and winning rate for each party. It is also hoped that one can extend the model globally based on more global data. With more data used for training the model, the bias will be reduced.

Reference

- Data source: 2017 General Social Survey (GSS): Families Cycle 31, provided by Statistics Canada under the terms of the Data Liberation
- Data source: Stephenson, Laura B., Allison Harell, Daniel Rubenson and Peter John Loewen. The 2019 Canadian Election Study – Online Collection. [dataset]
- Wu, C., & Thompson, M. E. (2020). Sampling Theory and Practice (ICSA Book Series in Statistics) (1st ed. 2020 ed.). Springer.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics) (1st ed. 2013, Corr. 7th printing 2017 ed.). Springer.
- Sinha, B. K., Das, K. K., & Mukhoti, S. K. (2008). On Some Aspects of Unbiased Estimation of Parameters in Quasi-Binomial Distributions. Communications in Statistics - Theory and Methods, 37(19), 3023-3028. doi:10.1080/03610920802162706
- Kennedy, L., Khanna, K., Simpson, D., & Gelman, A. (2020). Using sex and gender in survey adjustment. doi:arXiv:2009.14401
- Safiya Umoja Noble. Algorithms of Oppression: How Search Engines Reinforce Racism. NYU Press, 2018.
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- T. Lumley (2020) "survey: analysis of complex survey samples". R package version 4.0.
- Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. URL http://www.jstatsoft.org/v21/i12/.
- David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.2. https://CRAN.R-project.org/package=broom
- John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/
- Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.
- Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret

Appendix

Appendix A

The visualization for the survey and census data is presented below. The bar plots are used to present the frequency of each level for each variables.

Figure 2: In this plot, the frequency of `education`, `religion`, `marital status` in the CES data is shown.

Figure 2.A: Education in Survey Data

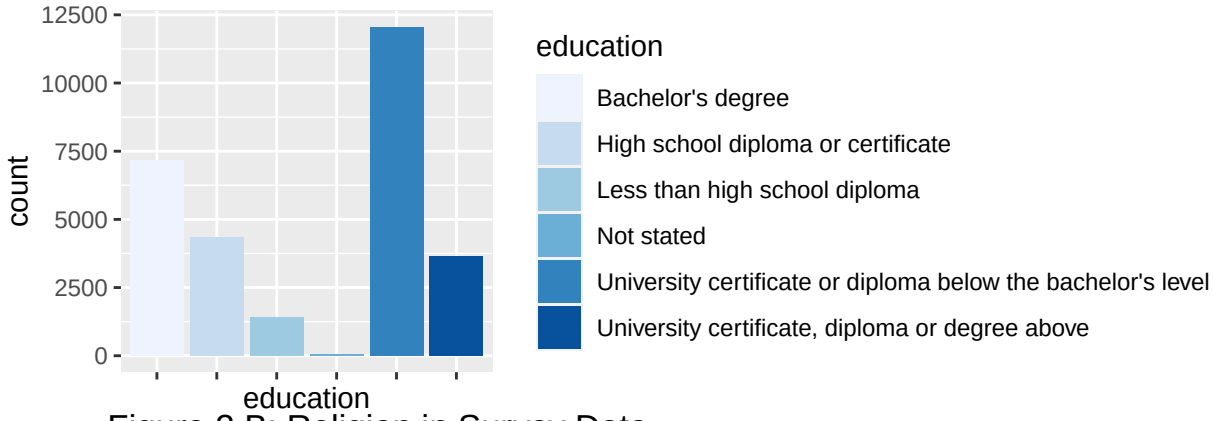


Figure 2.B: Religion in Survey Data

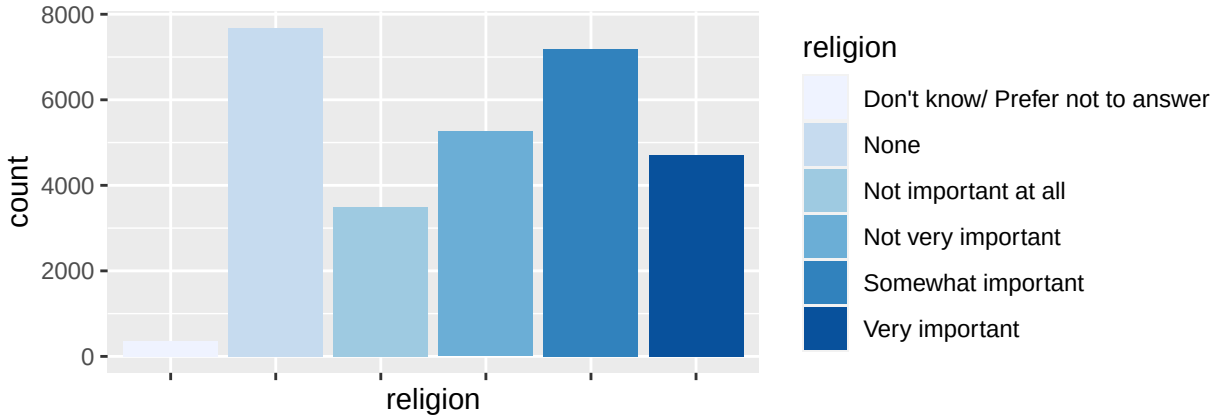


Figure 2.C: Marital Status in Survey Data

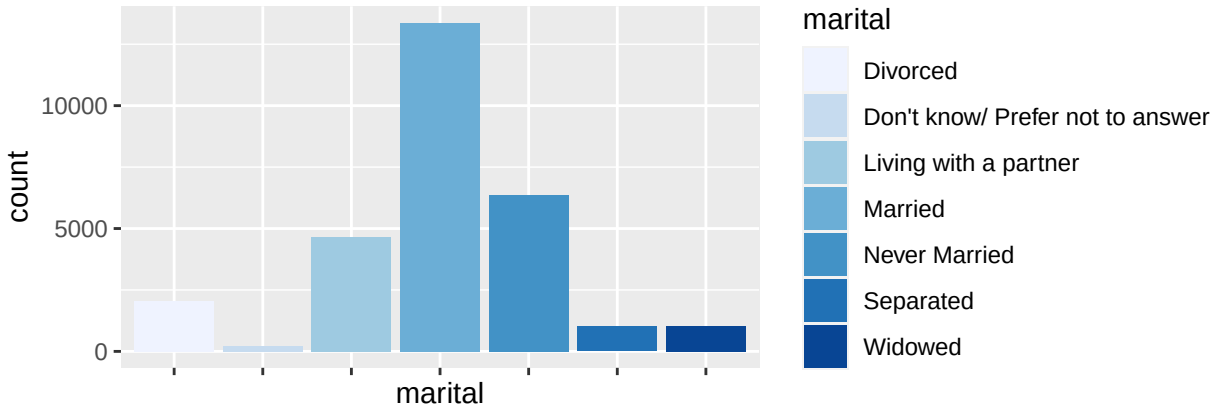
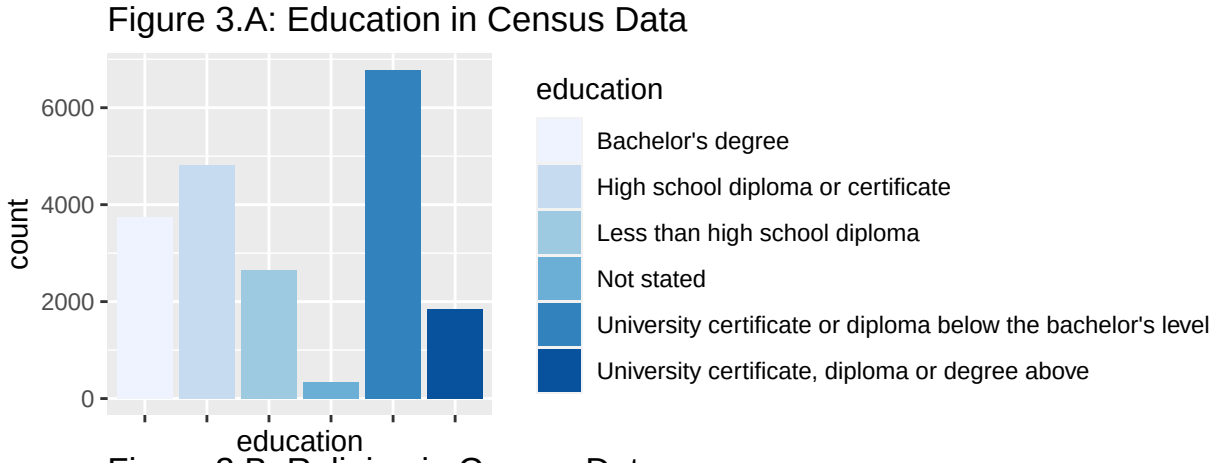
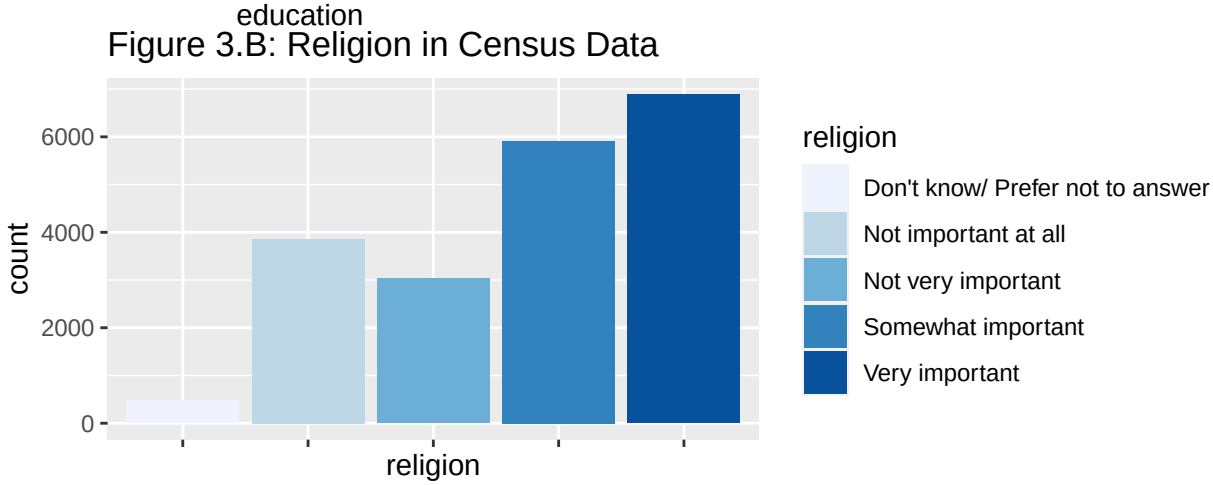
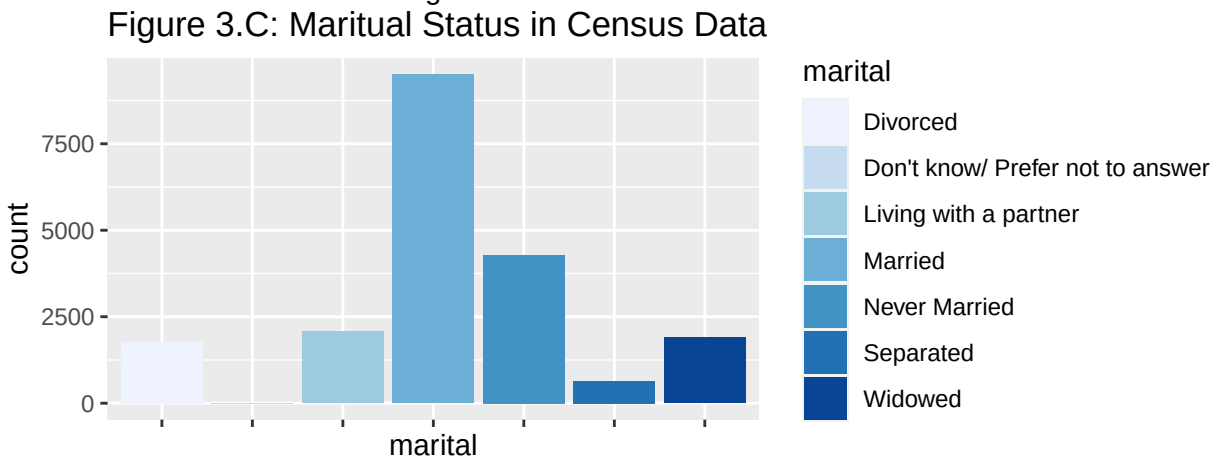


Figure 3: In this plot, the frequency of `education`, `religion`, `marital status` in the GSS data is shown.







Appedix B

The detailed model information is presented as below. The estimates for each variables are the log-odds from the logistic regression model.

Table 5: Logistic Regression Model Result

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.4240442	0.1869750	2.2679186	0.0233441
provinceBritish Columbia	-1.3635537	0.0612905	-22.2473741	0.0000000
provinceManitoba	-1.0094545	0.0791376	-12.7556882	0.0000000
provinceNew Brunswick	-1.6424613	0.1115538	-14.7234840	0.0000000
provinceNewfoundland and Labrador	-1.8080177	0.1393121	-12.9781810	0.0000000
provinceNorthwest Territories	-1.6839668	0.5634449	-2.9886982	0.0028049
provinceNova Scotia	-1.9332684	0.1143287	-16.9097319	0.0000000
provinceNunavut	-1.3370094	0.6339657	-2.1089616	0.0349594
provinceOntario	-1.4266690	0.0484741	-29.4315662	0.0000000
provincePrince Edward Island	-1.9704837	0.2682109	-7.3467703	0.0000000
provinceQuebec	-2.1018257	0.0585223	-35.9149772	0.0000000
provinceSaskatchewan	-0.4214891	0.0848939	-4.9648938	0.0000007
provinceYukon	-1.5976027	0.4450971	-3.5893355	0.0003323
age_group(28,33]	0.1664697	0.0880290	1.8910777	0.0586274
age_group(33,38]	0.1941732	0.0879607	2.2074994	0.0272897
age_group(38,43]	0.3799732	0.0890507	4.2669332	0.0000199
age_group(43,48]	0.3581515	0.0910753	3.9324781	0.0000843
age_group(48,53]	0.4841172	0.0882297	5.4870123	0.0000000
age_group(53,58]	0.4203864	0.0868784	4.8387884	0.0000013
age_group(58,63]	0.4071730	0.0869356	4.6836169	0.0000028
age_group(63,68]	0.3062420	0.0874183	3.5031816	0.0004607
age_group(68,73]	0.3878252	0.0915225	4.2374857	0.0000227
age_group(73,78]	0.5462563	0.1083123	5.0433427	0.0000005
age_group(78,83]	0.5397884	0.1490363	3.6218579	0.0002932
age_group(83,88]	0.6674949	0.2675611	2.4947384	0.0126125
age_group(88,93]	0.2152597	0.3942256	0.5460319	0.5850497
age_group(93,98]	0.5099675	0.4957678	1.0286417	0.3036597
age_group(98,Inf]	0.4212463	0.3540745	1.1897109	0.2341732
age_group[18,23]	-0.2406898	0.1040114	-2.3140719	0.0206732
sex	-0.4142605	0.0326079	-12.7043069	0.0000000
educationHigh school diploma or certificate	0.4240401	0.0513643	8.2555457	0.0000000

	Estimate	Std. Error	t value	Pr(>|t|)
educationLess than high school diploma	0.2843388	0.0777472	3.6572209	0.0002556
educationNot stated	0.0041192	0.3659302	0.0112569	0.9910186
educationUniversity certificate or diploma below the bachelor's level	0.2378318	0.0403765	5.8903548	0.0000000
educationUniversity certificate, diploma or degree above	- 0.2181668	0.0566283	- 3.8526078	0.0001172
religionNone	- 0.2416302	0.1478417	- 1.6343846	0.1021928
religionNot important at all	- 0.0739478	0.1517915	- 0.4871674	0.6261447
religionNot very important	0.1138452	0.1490546	0.7637818	0.4450057
religionSomewhat important	0.2892703	0.1474538	1.9617687	0.0498023
religionVery important	0.5686498	0.1488576	3.8200922	0.0001338
maritalDon't know/ Prefer not to answer	0.0487682	0.1941975	0.2511267	0.8017186
maritalLiving with a partner	0.1399050	0.0736630	1.8992580	0.0575440
maritalMarried	0.4387014	0.0623448	7.0366999	0.0000000
maritalNever Married	- 0.1999508	0.0725349	- 2.7566145	0.0058452
maritalSeparated	- 0.0954132	0.1057937	- 0.9018797	0.3671309
maritalWidowed	0.2377217	0.0983800	2.4163617	0.0156848

Appedix C

Here is the variance inflation factor(short for VIF) of the variables that are included in the logistic regression model. From the variance inflation factor table, it is clear to notice that all the variables included in the model does not have a strong correlation, as the VIF are below the threshold 5. Therefore, the model statisfies the model assumption.

Table 6: Variance Inflaction Factor for Predictors

	GVIF	Df	GVIF^(1/(2*Df))
province	1.152745	12	1.005940
age_group	1.582105	16	1.014439
sex	1.077604	1	1.038077
education	1.108980	5	1.010398
religion	1.122178	5	1.011594
marital	1.473495	6	1.032830