Prediction of voting result for the 2019 Canadian federal election

Can the Conservatives stand out

Junjie Lin & Yihang Luo (Group number 28)

December 1, 2022

Introduction

A country's federal election not only affects the future and development of the country but also has global relevance. Take the United States as an example, the struggle between American partisan politics will play an important role for direction of foreign policy[5]. A significant activity during the election is campaign. Although campaigns cannot be the major factor that influence election outcomes, voters utilize it as a basis for voting[11]. In addition, we can imagine that each party has a different governing philosophy that leads to various domestic policy making. Similarly, Canadian federal election is crucial for Canada as a democracy. "Elections are the one instrument available to eligible citizens and one that operates with a credible accounting system[6]." Citizens can vote for own interests, but they can also be influenced by personal conditions and surrounding factors[7]. Thus, our goal is to create a logistic regression model to predict whether voters will vote for the Conservatives in the next Canadian federal election.

There are several reasons why we choose Conservatives as the dependent variable in the model. We find that the Conservatives is a party with a long history. Even if the Progressive Conservatives were to die off for a while, the Conservatives Party has a huge influence in most elections to get a huge number of vote[8]. Thus, we choose logistic regression since it is an appropriate statistical model to model a binary response variable. To make the model more convincing, we select several independent variables that have effects on voting according to professional papers. Firstly, there are 4 interesting variables of personal attributes. Age is the first variable we are interested in because people of different ages have more or less life experience that will affect their intentions of voting. The journal[9] claims that several parties appeal to women voters in various. Similarly, the local vote is essential for each party[8], so sex and province are regarded as important variables. Moreover, education is one element that can support voters in decide[10]. Then, family income and household size 2 more surrounding factors as predictors in the model since parental education and income will affect how individuals vote[12&13]. Thus, we assume that family income and household size can influence a person's judgment about choosing a political party in Canadian federal election.

All variables come from both 2017 GSS census data and the survey data. The 2017 GSS census data includes all non-institutionalized persons 15 years of age and older, living in the 10 provinces of Canada with 81 variables and 20602 observations[14], The survey data is from the the 2019 federal election campaign and a post-election follow-up survey with 273 variables and 4021 observations[15]. We plan to use a post-stratification estimation after making a logistic regression model. It is a technique used in sampling surveys to improve the effect of variables with large population ratios on estimates of survey and census data. As a matter of fact, a good prediction model is important because it can help the candidates understand the reasons behind voters behaviors, then they can "spend their time and energy more wisely"[16].

Data

The 2017 GSS, conducted from February 2nd to November 30th 2017, is a sample survey with cross sectional design. Its target population includes all non-institutionalized persons 15 years of age and older, living in the 10 provinces of Canada. It is provided by Statistics Canada and maintained by Computing in the Humanities and Social Sciences (CHASS) at the University of Toronto[14]. There are 81 variables with 20602 observations.

The CES for 2019 included a dual-mode, two-wave data collection with a rolling cross-section during the 2019 federal election campaign and a post-election follow-up survey[15], which can be assessed by anyone. There are 273 variables with 4021 observations.

Before using these data to do the analysis, we need to clean up the data first. Since we will utilize a logistic regression model with post-stratification to predict whether the Conservatives will win the next Canadian federal election, we will clean and organize both sets of data for six important variables: age, sex, province, family income, and household size. For the census data, we need to round up the age and plus 2 for each observation to synchronize the age with the 2019 survey. Then, we find that several group names of categorical variable education are too long to be easily understood and read by the audience. In order to solve this problem, the appropriate shortening was made. The specific categories are shown in Figure 1. The other 4 variables have applicable names and values in the census data.

However, we also need to change the name and value of variables in the survey data to correspond to the existing variables in the census data for post-stratification estimation. The variable q2 records the birth year of the 2019 participants, so we can get age by subtracting their year of birth from 2019. To get sex information in the survey data, we just need to exclude those observations that answer other, refused, or do not know because the value of sex in the census data only has female and male. The variable q4 corresponds to province in the census data, whereas there are several invalid and refusal answers and provinces that do not appear in the census. We should delete these observations. Similarly, only observations with values between 1 and 6 should be left for the variable q71. The variable q69 shows the income of a family with specific numbers, but the family income in the census data is categorical. We need to change this numerical variable to a categorical variable. Then, variable q61 record the education of each observation with a wider range of value. The only thing we should accomplish is narrowing the range of its values to make it the same as education in the census data. Especially, we should create a variable to record whether each observation will vote Conservative or not through q11 and q12. The method was introduced by Perry in 1960[19], where the leaning question for undecided voters could mostly decide their final choices. The value 1 represents that the observation is certain to vote Conservative or has the intention to vote Conservative. Otherwise, the value will be 0 except for refusing to answer or not voting. Finally, we will delete unrelated variables for both sets of data. After that, removing NA observations is the last step.

Here is a resource for grabbing the CES2019 data: https://awstringer1.github.io/sta238-book/section-short-tutorial-on-pulling-data-for-assignment-1.html#section-canadian-election-study

In six important variables, age and household size are numerical variables. Sex, province, education, and family income are categorical variables. The meaning of each variable can be known directly from the name. Just to emphasize again, sex of observations in both set of data is either male of female since only these two categories are recorded in the census data. Similarly, a limited number of provinces are included in the category of provinces. There are Alberta, British Columbia, Manitoba, New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario, Prince Edward Island, Quebec, and Saskatchewan. More information of age, household size, education, and family income will be introduced in tables and figures below.

Table 1: Numerical Summaries of Survey Data

Variable	Min	Median	Mean	Max
Age	18	51	50.730	100
Household Size	1	2	2.568	6

Table 2: Numerical Summaries of Census Data

Variable	Min	Median	Mean	Max
Age	15	54	52.150	80
Household Size	1	2	2.348	6

Tables 1 and 2 are numerical summaries of the survey data and the census data respectively. Range and center for household size in both sets of data are similar. In contrast, the range of age in the survey data is wider. Maximum age is 20 years older than in the census data.

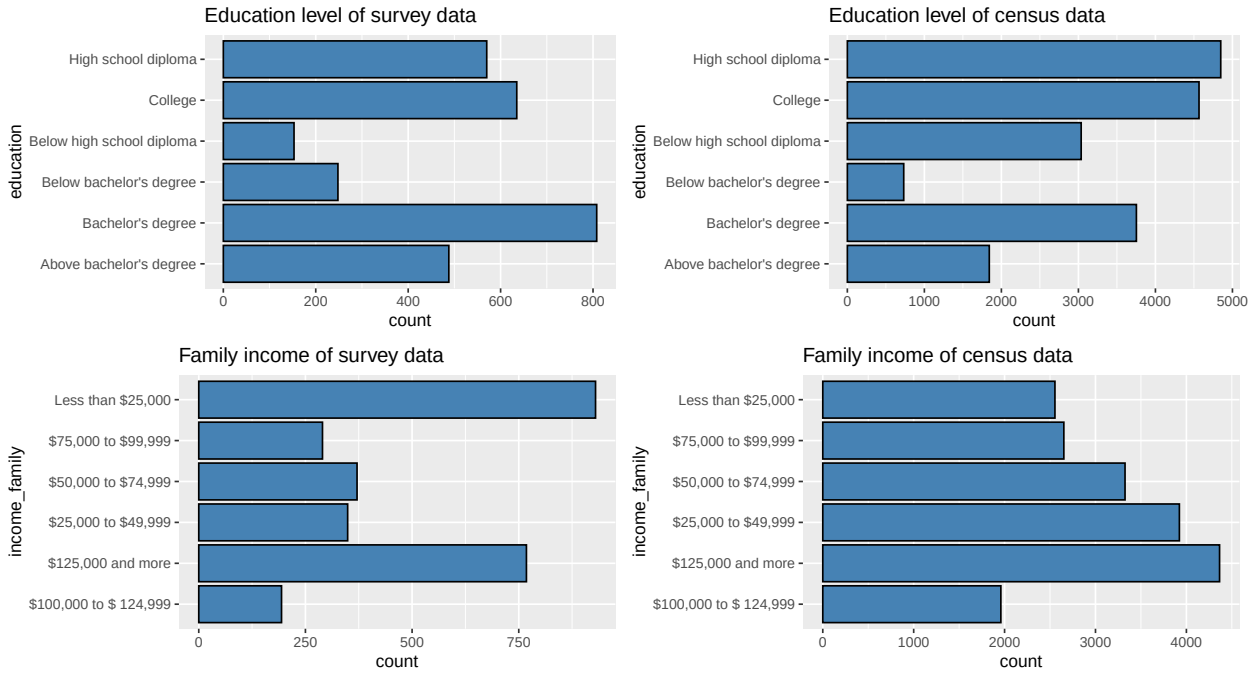


Figure 1: Histogram of numeric variables age and household size of survey data and census data

Figures in the first row are histograms of education level. There are six categories from low to high: below high school diploma, high school diploma, college, below bachelor's degree, bachelor's degree, and above bachelor's degree. We can discover that there are more highly educated observations in survey data. Figures in the second row are histograms of family income. The survey data collect a lot of families with incomes below $25,000.Family income in the census data is almost equally distributed in every category, but the proportion of high-income families is significantly larger.

As a quick summary of data, the six important variables will be the key to the methods we are about to introduce. Please remember them and their characteristics.

Methods

Logistic Regression Model

The Conservatives is a party with a long history. Even with the temporary demise of the Progressive Conservative Party, the Conservatives have enormous influence in most elections and can garner a large number of votes[8]. Using the logistic regression model, we can see the proportion of voters who will vote for the Conservative Party, since the predicted outcome "if voters will vote for the the Conservative Party or not" is a binary response variable. We should first check the following assumptions:

- outcome is binary
- linearity in the logit for continuous variables
- absence of multicollinearity
- lack of strongly influential outliers

Suppose we are building a simple logistic regression model and we are trying to use age, sex, province, education, family income and household size to predict whether more people vote for the Conservative Party. The model we are trying to estimate is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{sex} + \beta_3 x_{province} + \beta_4 x_{education} + \beta_5 x_{incomefamily} + \beta_6 x_{hh_{size}}$$

Where

- p represents the proportion of voters who will vote for Conservative Party.
- β_0 represents the intercept of the model, and is the log of odds of voting for Conservative Party when the individual is at a certain age, sex, province, education, family income and household size.
- β_1 represents the slope of age in the model: for everyone one unit increase in age, a β_1 is expected to increase log odds of voting for the Conservative Party.
- β_2 represents the average difference in log odds of voting for the Conservative Party between Male and Female with certain age, province, education, family income and household size.
- β_3 represents the average difference in log odds of voting for the Conservative Party between distinct provinces with certain age, sex, education, family income and household size.
- β_4 represents the average difference in log odds of voting for the Conservative Party between different level of education with certain age, sex, province, family income and household size.
- β_5 represents the average difference in log odds of voting for the Conservative Party between various family income with certain age, sex, province, education and household size.
- β_6 represents the slope of hh_size in the model: for everyone one unit increase in household size, a β_6 is expected to increase log odds of voting for the Conservative Party.

Outcome binary

The outcome response, the variable vote_Conservatives we choose for our model, is binary, because we choose only the proportion of people who will or will not vote or lean to vote for the Conservative Party.

Linearity of logit with continuous variables

We use Box-Tidwell estimation to check this assumption by testing whether the logit transform is a linear function of the predictor, effectively testing whether this addition makes the prediction better by adding the original predictor's nonlinear transformation as an interaction term. We also use the Hosmer and Lemeshow test, which compares the observed and expected values within each group and sums the groups to assess

whether the logistic regression model is well calibrated. A large p-value indicates a good fit and a small p-value indicates a poor fit.

Box-Tidwell has: H_0 : linearity between continuous variable and log-odds vs. H_A : non-linearity between continuous variable and log-odds. The two continuous variables in the predictors are age and household size. Let $\alpha = 0.05$.

Hosmer and Lemeshow test has: H_0 : the observed proportions of $Y = 1$ in these groups to be similar to their within-group average predicted probabilities. H_A : the observed proportions of $Y = 1$ in these groups to be different from their within-group average predicted probabilities. Let $\alpha = 0.05$. [17]

Table 3: Summary of the Box-Tidwell estimation

Variable	p_value
age	0.07588908
household size	0.39011970

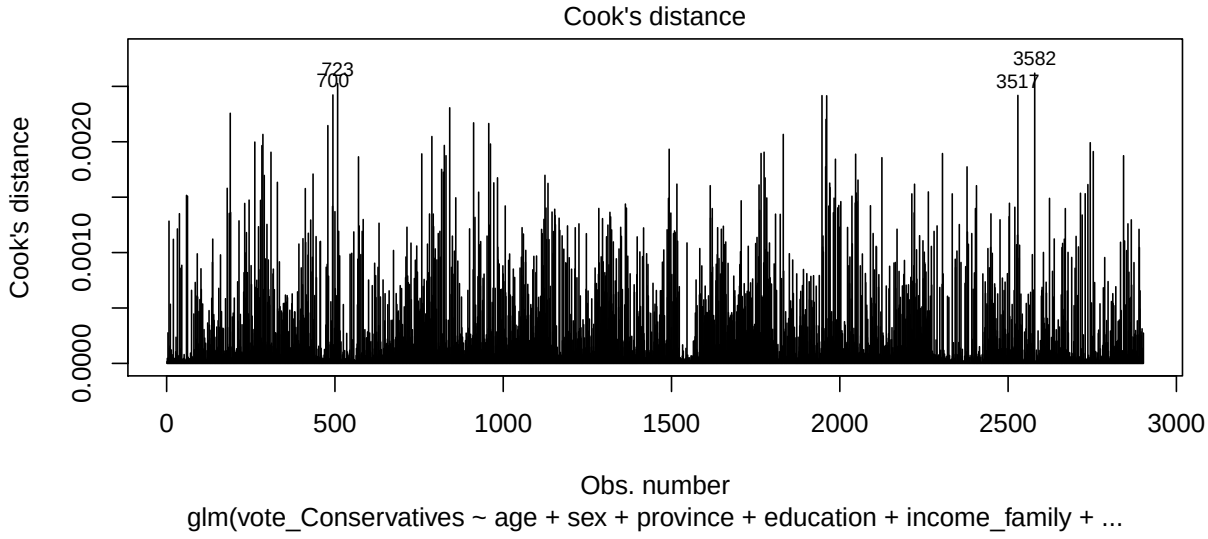
Table 4: Summary of the Hosmer and Lemeshow test

Variable	p_value
age	1
household size	1

[18] Since from table 3, p-value of both age and household size is larger than α, we can FTR H0 and the linearity assumption seems to be satisfied. From table 4, p-value of both age and household size is larger than α, so we do not reject the null hypothesis of a perfect fit, and we conclude that the fit is adequate.

Checking influential values

Influential values are individual extreme data points that can change the quality of a logistic regression model, including extreme outliers. We use r_j for each j observation to find out outliers, and Cook's distance method or plot to find out influential points.



Since data points with an absolute standardized residuals above 3 represent possible outliers, as well as the calculation shows, we can conclude that there is no extreme outliers or influential points existing in our model from the graph.

Multicollinearity

Multicollinearity corresponds to the case where the data contain highly correlated predictor variables. A VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity.

Table 5: VIF value of each predictor

Variable	GVIF	DF
Age	1.212971	1
Sex	1.032163	1
Province	1.113209	9
Education	1.165383	5
Family income	1.241857	5
Household size	1.260718	1

From the result of VIF value we get above in table 5 with no value exceeds 5 or 10, so there is no problematic amount of collinearity, we do not need to remove any variable.

AIC and BIC

Since we are not adding additional predictors to the model, no adjusted R^2 should be taken into account. The AIC is essentially a statistic that balances the goodness of fit of the model with a penalty term reflecting how complex the model is. The BIC is the final criterion and actually has nothing to do with the idea of statistical information. As with the AIC, smaller values of BIC indicate the better model. The automated selection will output a model only based on goodness by computer calculation, which is a systematic stepwise selection that takes into account the conditional nature of regression and can change the relationship with the removal or addition of predictors.

We use AIC, BIC and automated selection to check the model. Besides, we decide to choose two predictors from all the variables to see if they make the model better, or we would remove them. We suppose that Canada is not as closely linked within regions and family income may not play a great role for voting the Conservative Party, so we choose family income and province. We also use step for the automated model decision.

Table 6: AIC and BIC of each model

Model	AIC	BIC
Model 1	3298.805	3436.188
Model 2	3517.762	3601.386
Model 3	3304.593	3412.110
Auto_Model	3298.805	3436.188

From the above AIC and BIC result in table 6, we see that for the variable province or family income in the model, AIC increases while BIC decreases, thus we could hardly tell they should be removed or not. So we choose to believe the automated selection model, which is the only model checking method in this project where AIC and BIC both decreases when removing the variable household size it advises. The possible

reason is that there is a relatively weak link between the policies of each party and the number of family members.

So we attain our new model:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{sex} + \beta_3 x_{province} + \beta_4 x_{education} + \beta_5 x_{income_{family}}$$

Post-Stratification

Post-stratification estimation is a technique used in sampling surveys to improve estimate efficiency. If appropriate census-style information exists about the variables of interest for the analysis, then we can use these population totals to adjust the sampling weights and improve the estimates.

To more accurately estimate the proportion of voters who will vote Conservative, a post-stratification analysis would be a good option. I build cells based on different age, sex, province, education, family income and household size. By using the model above, I will estimate the proportion of voters in each strata, The estimates for each stratum are then weighted by the population size of that stratum, and these values are summed and divided by the overall population size. As a result, we can overcome the inconsistencies in the characteristics of participants such as age and household size(see Appendix Figure 2), as well as more effectively increases the effect of variables with large population ratios on estimates.

$$\hat{y}^{PS} = \frac{\sum N_j \widehat{y}_j}{\sum N_j}$$

Where

- N_j represents the population total for the strata j.
- $\widehat{y}_j$ represent the sample mean for the strata j.

So logistic regression and post-stratification are the two methods we perform to analyze our predictor outcome. The model simulated by logistic regression passes the assumptions it requires well. After determining the feasibility of the two models, we visualize them in mathematical equations to facilitate the next step in the computation.

All analysis for this report was programmed using `R version 4.0.2`.

Results

Table 7: Summary of the logistic regression model

term	estimate	std.error	statistic	p.value
Intercept	-1.224082	0.287889	-4.252	0.000021
age	0.014552	0.002598	5.600	0.000000
sexMale	0.540835	0.089264	6.059	0.000000
provinceBritish Columbia	-1.696637	0.180466	-9.401	0.000000
provinceManitoba	-0.935620	0.215138	-4.349	0.000014
provinceNew Brunswick	-1.411236	0.236642	-5.964	0.000000
provinceNewfoundland and Labrador	-1.877156	0.248861	-7.543	0.000000
provinceNova Scotia	-1.759067	0.248410	-7.081	0.000000
provinceOntario	-1.689402	0.179650	-9.404	0.000000
provincePrince Edward Island	-1.807293	0.252337	-7.162	0.000000
provinceQuebec	-2.373477	0.192758	-12.313	0.000000
provinceSaskatchewan	-0.608794	0.214607	-2.837	0.004557
educationBachelor's degree	0.552529	0.146326	3.776	0.000159
educationBelow bachelor's degree	0.629687	0.190115	3.312	0.000926
educationBelow high school diploma	1.251807	0.218503	5.729	0.000000
educationCollege	1.022127	0.150878	6.775	0.000000
educationHigh school diploma	1.139547	0.154005	7.399	0.000000
income_family$125,000 and more	0.546711	0.189399	2.887	0.003895
income_family$25,000 to $49,999	-0.044269	0.214563	-0.206	0.836540
income_family$50,000 to $74,999	0.083160	0.209000	0.398	0.690708
income_family$75,000 to $99,999	0.093358	0.220301	0.424	0.671731
income_familyLess than $25,000	0.210022	0.188080	1.117	0.264139

Predictors are significant when the p-value is less than 0.05. From the table 7 summary above, obviously that age, gender, province and education are significant predictors. However, we do not intend to remove them because our goal is to predict the proportion of voting for the Conservative Party.

Then we can get the formula for our logistic regression model:

$$\log(\frac{\hat{p}}{1-\hat{p}}) = -1.656126 + 0.017282X_{age} - 12.837403X_{province} + 0.543546X_{Male} + 1.000184X_{income_{family}}$$

And the result of logistic regression and post-stratification estimation:

Table 8: Result of logistic regression and post-stratification estimation

logistic_predict	liberal_predict
0.3304618	0.3465934

So we know that the proportion of voters who will vote for Conservative Party we predict is 0.346593387536034 from table 8. I think this result is quite accurate in terms of the real campaign results for 2019[4], and the error produced or difference with the true result probably due to the fact that at the beginning we didn't include people who refused to participate in the survey overall while few of them chose the Conservative Party when they voted at the end. Moreover, it is a little more closer to the true 2019 Conservative Party voting result compared to that of doing only logistic regression, since post-stratification can decrease the heterogeneity in participant characteristics.

Conclusions

We chose the Conservative Party as our research subject of this project as it has a huge influence in most elections and gets a relatively large number of votes[8]. Logistic regression was used as a very suitable statistical model to simulate binary response variables. We found several variables of interest to me in selecting the independent variables that have an impact on voting: age, where people of different ages have different life experiences that may influence their voting intentions; gender, where female voters are attracted to several different parties[9]; since, where local voting is crucial for each party[8], education, which can support voters in making their decisions[10], as well as household income and family size, parents' education and income may influence how individuals vote[12&13].

All variables were obtained from the 2017 GSS census data as well as survey data (14&15). Then, we use the post-stratified estimation method, which can reduce the heterogeneity of voters characteristics. Looking at the real campaign results for 2019, the result it produces is relatively accurate (4), and it is much closer to the true 2019 Conservative Party vote than the results from a logistic regression only.

The sample vote proportion of logistic regression is 0.3304618 and the result through post-stratification is 0.346593387536034. In our projections, voters who voted for the Conservative Party would account for close to 40% of the total vote, which is already very high for a partisan campaign, so it would be very much expected to be the new party in office in 2019. Of course, the final result was that the Liberals won that year's election, but the Conservatives did hold a big advantage in the results, which was not far from our estimates.

Certainly there are many potential problems with our model. During the procedure of cleaning data, we exclude observations who refused to say which party to vote for. It can be a bias of the model. In addition, when we create the first logistic regression model, we may add more possible predictors to get a better final model by VIF, AIC and BIC. Or we can overfit the model because we cannot know which predictor actually works at the current level. Whether the predictors should be divided into several categorical options is also a consideration. Here we choose not to divide since there are too many of them from province, education and family income. While there may produces bias. Finally, solely trust the automated model based on goodness may remove variables that is very important to our model, and we cannot say that household size is definitely not a key section of our model.

Since the ultimate question we explore is actually which party will be most likely to win the campaign from the factors likely to be involved, while our predictions are relatively accurate, making predictions only for the Conservative Party may limit our research. In the future, we can make predictions for the Liberals and other parties in the same year to get a more comprehensive and accurate analysis. From the case that post-stratification works well we can directly use post-stratification estimation for the corresponding prediction.

Bibliography(APA format)

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)

2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.

3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)

4. Elections Canada. Home / Resource Centre / Reports Elections Canada's Official Reports / 43rd General Election: Official Voting Results (raw data) [https://www.elections.ca/content.aspx?section=res&dir=rep/off/43gedata&document=summary&lang=e]

5. Schwartz, T. A. (03 2009). "Henry, ... Winning an Election Is Terribly Important": Partisan Politics in the History of U.S. Foreign Relations. Diplomatic History, 33(2), 173–190. [https://doi.org/10.1111/j.1467-7709.2008.00759.x]

6. Kanji, M., Bilodeau, A., & Scotto, T. J. (Eds.). (2012). The canadian election studies : Assessing four decades of influence. UBC Press.

7. Rolfe, M. (2012). Voter Turnout: A Social Theory of Political Participation. Cambridge University Press.

8. Koop, R., & Bittner, A. (2013). Parties, elections, and the future of Canadian politics. UBC Press.

9. Sanders, A., Gains, F., & Annesley, C. (2021). What's on offer: how do parties appeal to women voters in election manifestos? Journal of Elections, Public Opinion and Parties, 31(4), 508–527. [https://doi.org/10.1080/17457289.2021.1968411]

10. Simon Feess (Author), 2001, Does education influence voter turnout?, Munich, GRIN Verlag [https://www.grin.com/document/101356]

11. Nickerson, D. W., & Rogers, T. (2020). Campaigns influence election outcomes less than you think. Science, 369(6508), 1181–1182. [https://doi:10.1126/science.abb2437]

12. Polacko, M. (2020). Party Positions, Income Inequality, and Voter Turnout in Canada, 1984-2015. American Behavioral Scientist, 64(9), 1324–1347. [https://doi.org/10.1177/0002764220941238]

13. Gidengil, E., Wass, H., & Valaste, M. (2016). Political Socialization and Voting: The Parent-Child Link in Turnout. Political Research Quarterly, 69(2), 373–383. [http://www.jstor.org/stable/44018017]

14. SDA @ CHASS / GSS data [https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/analysis]

15. 2019 Canadian Election Study [http://www.ces-eec.ca/]

16. Newman, B. I., & Sheth, J. N. (1985). A Model of Primary Voter Behavior. Journal of Consumer Research, 12(2), 178–187. [http://www.jstor.org/stable/254350]

17. Hosmer, David W., and Stanley Lemeshow. 2000. Applied Logistic Regression. Second edtion. New York: John Wiley & Sons.

18. Lele, Subhash R., Jonah L. Keim, and Peter Solymos. 2019. ResourceSelection: Resource Selection (Probability) Functions for Use-Availability Data. [https://CRAN.R-project.org/package=ResourceSelection.]

19. Roper, E., Perry, P., & Field, M. D. (1960). Election Polling Trends, 1960. American Behavioral Scientist, 4(2), 3–5. [http://doi:10.1177/000276426000400202]

Appendix

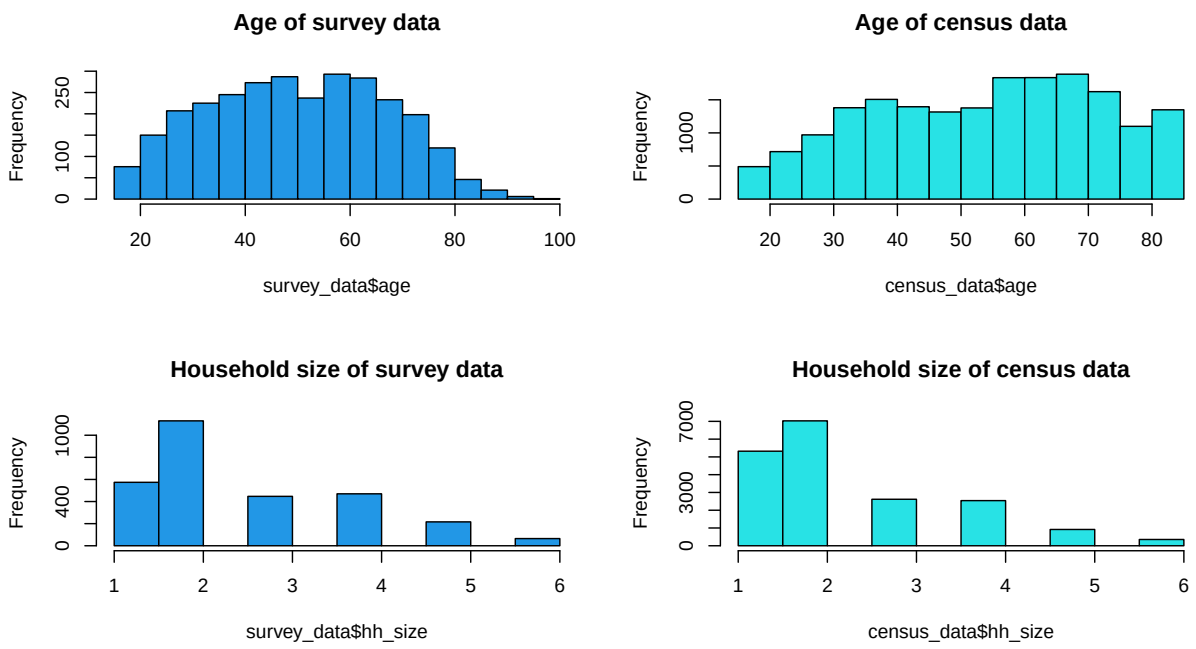


Figure 2: Histogram of variables of survey data and census data