Would the 2019 Canadian Federal Election have been different if ‘everyone’ had voted? A Post-stratification Analysis on the Election Results

Mingze Xu (1004839605)

December 21, 2020

Abstract

We fit a multilevel, multivariable logistic regression model for the mean of a binary response variable conditional on poststratification cells. This approach combines the modeling approach often used in small-area estimation with the population information used in poststratification. Qe applied this method to the 2019 Canadian Federal Election, poststratified by age, sex and the usual demographic variables. We envision the most essential usage of this method to be not forecasting election results but estimating public opinion on a variety of issues at the local level.

Keywords: Canadian Federal Election 2019, poststratification, national polls, survey sampling

Introduction

The 2019 Canadian federal election took place on October 21, 2019, to elect members of the House of Commons to the 43rd Canadian Parliament. The Liberal Party won 157 seats to form a minority government and lost the majority they had won previously in the 2015 election. The Conservatives got the popular vote over the Liberals. The Conservative Party won 121 seats and remained the Official Opposition Party. The Bloc Québécois won 32 seats to get their official party status back and became the third party. The New Democratic Party won 24 seats. After the election, Prime Minister Justin Trudeau, who led the Liberal Party in the election, rejected a coalition and subsequently his new cabinet was sworn in on November 20, 2019. The federal election was a nationwide election in which abudant eligible voters voted. However, since not every eligible voter had voted, the primary research question of this study is to examine whether the 2019 Canadian Federal Election results would have been different if 'everyone' (every eligible voter in Canada) had voted.

We used the 2019 Canadian Election Study - Online Survey as our survey dataset and the Canadian General Social Survey as our census dataset. We used a multilevel logistic regression model with post-stratification to predict the probability of voting for the Liberal Party. We wanted to see if the results (Liberals forming a minority government) would be overturned if everyone had voted in the election.

Methods

Data

```
## TO CITE THIS SURVEY FILE: Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John
##            https://doi.org/10.7910/DVN/DUS88V, Harvard Dataverse, V1
## LINK: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DUS88V
```

```
## TO CITE THIS SURVEY FILE:
##
## - Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, '2019 Canadian El
##
## - Stephenson, Laura, Allison Harrel, Daniel Rubenson and Peter Loewen. Forthcoming. 'Measuring Prefe
##
## LINK: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DUS88V
```

The 2019 Canadian Election Study - Online Survey was conducted to record the attitudes and various characteristics of Canadians during and after the 2019 election. It follows the tradition of Canadian Election Studies that began in 1965. This dataset contains data in this particular online survey.

Canada's General Social Survey (GSS) program was established in 1985 and designed as a collection of independent, annual, voluntary, cross-sectional surveys. Each survey covering a single topic in-depth. The global objectives of the program were, and continue to be, to collect data on social trends in order to monitor dynamics in the living conditions and well-being of Canadians, and to provide useful information on specific social policy problems.

GSS data has served as evidence behind quintessential government programs and policies focused at improving the health and well-being of Canadians. It regularly conducts comprehensive research projects on a variety of important topics and has become a important training tool for quantitative methods in post-secondary institutions across Canada. The GSS is an vital foundational social survey within our country's national statistical and data collection system.

The most recent GSS program surveys have focussed on the following themes: life at work and home; families; caregiving and care receiving; giving, volunteering and participating; victimization; social identity and time use.

Each of these themes have in the past been explored in detail roughly every five to seven years. Other than the core topic, space is reserved in each cycle for new information that targets emerging, policy-related problems. As well, each survey includes socio-economic information including age, sex, education, religion, immigrant status, place of birth, Indigenous group, population group/visible minority status, and more. More recent surveys also include questions on disability and veteran status. Normal collection of cross-sectional data enables for trend analysis, and for the validation and development of new ideas.

Since its release, the GSS has validated new methods and worked closely with parties of interest in the government, academia, and social sectors. After 2013, valuable data have been collected using a combination of self-completed web questionnaires and phone interviews. This new trend is continuing with increased drive as the GSS spearheads the way in re-defining itself within Statistics Canada's revamp. New methods to data collection, processing and release of information are being researched to assure the GSS stays up-to-date and of use in the future.

Methodology

```
## TO CITE THIS SURVEY FILE:
##
## - Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, '2019 Canadian El
##
## - Stephenson, Laura, Allison Harrel, Daniel Rubenson and Peter Loewen. Forthcoming. 'Measuring Prefe
##
## LINK: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DUS88V
```

```
## # A tibble: 1 x 1
##   alp_predict
##         <dbl>
## 1       0.286
```

We fit a multivariable logisitc regresison model on the survey dataset (The 2019 Canadian Election Study - Online Survey) to predict whether a person would vote for the liberal party or not.

In statistics, the logistic model (or logit model) is used to model the probability of a particular class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model multiple classes of events such as determining whether an image contains a monkey, donkey, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1, with a sum of these probabilities equal to one.

Logistic regression model is a statistical model that in its basic form uses a logistic function to model a binary response variable, although many more complex extensions can be modelled. In regression analysis, logistic regression is estimating the parameters of a logistic model (a form of binary linear regression). Mathematically, a binary logistic model has a response variable with two possible values, such as pass or fail and is represented by a isngle indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds of the event of interest happening) for the value labeled "1" is a linear combination of one or more independent variables (these are called the "predictors"); the independent variables can each be a binary variable (two groups, coded by an indicator variable) or a continuous variable (any value on the real line). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling. The function that converts log-odds to probability is the logistic function. The unit of measurement for the log-odds scale is called a logit which represents the log of the odds. Similar models with a different sigmoid function instead of the logistic function can also be used, such as the probit model which uses a standard normal link. The defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively affects the odds of the given outcome at a fixed rate, with each predictor variable having its own parameter; for a binary dependent variable this generalizes to the odds ratio.

In a binary logistic regression model, the dependent variable has two levels. Outputs with more than two values are modeled by multinomial logistic regression and, if the multiple categories are ordered, by ordinal logistic regression. The logistic regression model in and of itself plainly models probability of output in terms of input; it does not perform statistical classification, but it can be used to make a classifier, for example by choosing a cutoff value and classifying inputs with probability greater or less than the cutoff as one class, below the cutoff as the other; this is prevalent way to make a binary classifier.

Poststratification involves tweaking the sampling weights so that they sum to the population sizes within each poststratum. This usually results in declining bias because of nonresponse and underrepresented groups in the target population. Poststratification also has the tenancy to result in smaller variance estimates. We used the GSS as our poststratification dataset to poststratify the proportion of voters voting for the Liberals.

Results

Table of baseline characteristics is below:

```
##
##                  level  Overall
##   n                     37,531
##   age (mean (SD))        48.76 (16.59)
##   sex (%)        Female  21980 (58.6)
##                  Male    15551 (41.4)
```

Model summary of the logistic regression model is below.

Observations	31335 (6196 missing obs. deleted)
Dependent variable	vote_party
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(2)$	21.27
Pseudo-R^2 (Cragg-Uhler)	0.00
Pseudo-R^2 (McFadden)	0.00
AIC	37377.28
BIC	37402.34

	Est.	S.E.	z val.	p
(Intercept)	-1.10	0.04	-27.29	0.00
age	0.00	0.00	4.11	0.00
sexMale	0.04	0.03	1.40	0.16

Standard errors: MLE

The poststratification estimate of the proportion of Canadians voting for the Liberals is 0.286. That is, roughly 29% would have voted for the Liberals if they had voted.

Discussion

We used the 2019 Canadian Election Study - Online Survey as our survey dataset and the Canadian General Social Survey as our census dataset. We used a multilevel logistic regression model with post-stratification to predict the probability of voting for the Liberal Party. We wanted to see if the results (Liberals forming a minority government) would be overturned if everyone had voted in the election.

Conclusions

The poststratification estimate of the proportion of Canadians voting for the Liberals is 0.286. That is, roughly 29% would have voted for the Liberals if they had voted.

Weaknesses and Next Steps:

There are some disadvantages of using logistic regression modeling and they include: 1) if the number of observations is smaller than the number of predictors, logistic Regression should not be employed, otherwise, it may cause overfitting; 2) it defines linear boundaries; 3) a big limitation of logistic Regression is the assuming of linearity between the response variable and the predictor variables, it can only be used to predict discrete functions. So, the response variable of logistic regression is bound to the discrete number space; 4) non-linear problems cannot be solved with logistic regression because it has a linear decision space; 5) perfectly linear data is very hard to find in real-world situations; 6) logistic regression requires independence in the predictors (no multicollinearity); 7) it is challenging to obtain complex predictor associations using logistic regression and stronger algorithms such as neural networks can easily outdo logistic regression in many settings, and 8) in normal linear regression response and predictor variables are related linearly, but logistic regression requires response variables are linearly related to the log odds and predicted by a linear combination of the predictors.

References

Datasets

Press, C., Finance, Y., & Newsweek. (2020, October 30). New: Second Nationscape Data Set Release. Retrieved November 03, 2020, from https://www.voterstudygroup.org/publication/nationscape-data-set

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

Stephenson, Laura, Allison Harrel, Daniel Rubenson and Peter Loewen. Forthcoming. 'Measuring Preferences and Behaviour in the 2019 Canadian Election Study,' Canadian Journal of Political Science.

LINK: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DUS88V

Team, M. (n.d.). U.S. CENSUS DATA FOR SOCIAL, ECONOMIC, AND HEALTH RESEARCH. Retrieved November 03, 2020, from https://usa.ipums.org/usa/index.shtml

Literature

GeeksForGeeks, A., AmiyaRanjanRout, & GeeksForGeeks, T. (2020, September 02). Advantages and Disadvantages of Logistic Regression. Retrieved December 21, 2020, from https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

General Social Survey (GSS). (2019, February 20). Retrieved December 21, 2020, from https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2019001-eng.htm

Logistic regression. (2020, December 18). Retrieved December 21, 2020, from https://en.wikipedia.org/wiki/Logistic_regression