2019 Canadian Federal Election Poststratification Analysis

Yida Wang

December 21, 2020

Abstract

We analyzed the 2019 Canadian Election Study data, web version and used it as our survey dataset. We specifically analyzed a post-processed subset of this dataset. We also used the 2017 Canadian Social Survey (GSS) on the Family as our poststratification/census dataset. We built a multivariable linear regression model to model the political spectrum score. We found that on average, Canadians are neither too leftwing or too rightwing on the political spectrum. Therefore, we conclude that the liberals won the minority government in the 2019 Canadian Federal Election just by chance.

Keywords: Canada, Justin Trudeau, 2019 Canadian Federal Election, Poststratification

Introduction

Liberal majority government formed and spearheaded by current Prime Minister Justin Trudeau was the result of the 2015 Canadian Federal Election. The Conservative Party became the Official Opposition (with Stephen Harper announcing his resignation as party head) and the New Democrats (NDP) became the third most powerful party. While members of the Bloc Québécois and the Greens were elected by the people to the House of Commons, both failed to achieve the required number of MPs for official party identification. Bloc leader Gilles Duceppe announced his resignation shortly after the election, and was succeeded by Parti Québécois MNA Martine Ouellet. After d a leadership review, Ouellet announced she would step down as Bloc leader on June 11, 2018,and was succeeded by Yves-François Blanchet on January 17, 2019.

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
```

Survey dataset

The 2019 Canadian Election Study: voting and elections are probably the most essential elements of democracy. Since 1965, the understanding of electoral democracy in Canada has been greatly improved by the Canadian Election Study, a big-scale survey of Canadian citizens conducted each election year.

During its long history, the CES has been a enriched source of data on Canadians' political behaviour and attitudes, measuring preferences on vital political issue like free trade with the United States, social-related spending and Quebec's position in Canada; political players, like parties, party heads and the government; and behavioral concerns, such as female's place at home, support for immigrants, and attitudes toward non-sexual genders; as well as political preferences and participation. These data provide an superior snapshot and record of Canadian society and political environment.

The `decon` dataset is a subset of the 2019 CES online survey under the name "decon" (demographics and economics) that provides a tool for teaching survey datasets and their analysis. This dataset is fantastic since it includes many demographic, socio-demographic and socio-economic attributes/factors/variables of the overall Canada's population. This dataset is helpful for data analysis because it does not have a large number of variables including adminstrative-type variables and only includes key variables like income, sex, and more.

Postratification census dataset

Tis dataset concerns users of the Public Use Microdata File (PUMF) of the 2017 General Social Survey (GSS) on the Family. Its goals are to give context and background information, to familiarize data users with the content of the survey, and to describe protocols and concepts related to data integrity, statistical estimation, data collection, post-processing and mathematical and statistical survey methodology.

The 2017 GSS, carried out from February 2nd to November 30th 2017, is a sample survey with a typical national-level cross-sectional design. The target population includes all non-institutionalized persons over 15 years old living in the 10 provinces of our country. The survey uses a new sampling frame, created back in 2013, that collates telephone numbers (landline and mobile) with Statistics Canada's Address Register, and grabs data via phone. Data are not immune to sampling and non-sampling errors as usual.

The GSS program, created in 1985, does phone and web surveys across the ten provinces. The GSS is recognized for its routine grabbing of cross-sectional data that allows for trend analysis, and its ability to validate and devise new methods that address ongoing or emerging problems.

The two primary end goals of the General Social Survey are:

1) To collect data on social trends in order to track changes in the standards of living and well-being of our citizens over the time horizon

2) To yield information on particular social policy problems of ongoing or emerging attention. To satisfy these objectives, data grabbed by the GSS consists of two parts: core content and classification variables. Core content is meant to measure and track changes in the country related to standards of living and well-being, and to supplement data to address particular policy problems. Classification variables (like age, sex, ethnicity and household income) help decompose population groups for usage in the core data analyses.

The central role family plays in people's lives is not debatable. Family now, however, must go through through changing conjugal, family, and work tracks. While our understanding of families in our nation has deepened greatly over the past few years, the future of families is a question of interest as we see that families are getting more diverse. How many families are there in the country? What are their characteristics and socio-demographic statuses? What do families at different stages of person life look like? How prevalent are step or single-parent families? The GSS on families has the answers to these and many other questions so that's why we used it as our poststratification dataset.

Methodology

In statistical sciences, linear regression modeling is a linear method to modeling the relationship between a response variable and one or more explanatory variables. The situation of one explanatory variable is called simple linear regression modeling; for more than one predictors, the framework is called multiple linear regression modeling. This terminology is distinct from multivariate linear regression, where multiple correlated response variables are modeled, rather than a single constant response variable.

In linear regression modeling, the relationships are modeled using linear predictor functions in which unknown but constant model parameters are estimated from the data. Such models are referred as linear models. The conditional mean of the response conditioned the values of the explanatory variables is modeled to be a linear function of those values. Linear regression stresses the conditional probability distribution of the response conditional on the values of the predictors, instead of on the joint probability distribution of all of these explanatory variables, which relates to the space of multivariate analysis.

Linear regression modeling was the first type of regression analysis to be explored and proved rigorously, and to be used extensively in real-world applications. The reason is because models which depend in a linear fashion on their unknown constant parameters are simpler and more convenient to fit than models which are not linearly related to their parameters and because the statistical properties of the derived estimators are easier to specify and compute.

Linear regression has many real-world applications. Most applications fall into one of the below two general types:

Should the objective be prediction, forecasting, or error reduction, linear regression modeling can be employed to fit a predictive model to an observed data set of data values of the response and explanatory variables. After developing this type of model, if extra values of the explanatory variables are gathered without an associating response value, the fitted model can be used to make a prediction of the response value.

Should the objective be explain the variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be used to measure the strength of the relationship between the response and the explanatory variables, and specifically to find out if some explanatory variables might have no linear relationship with the response at all, or to find out which subsets of explanatory variables may contain repetitive or redundant information about the response.

Linear regression models are usually fitted using the least squares approach, but they may also be fitted in other approaches, such as by minimizing the "absence of fit" in some other measure (as with least absolute deviations regression), or by minimizing a penalized version of the least squares objective function as in ridge regression, LASSO or elastic-net. A generalized version of the linear regression model could be used to model non-continous response variables such as count data, binary data, multi-level data or survival data. The technique is called generalized linear regression modeling.

We first fit a simple linear regression model using sex to predict the political spectrum score which is our response variable then we estimate the average score using poststratification. The higher the score, the more "right".

Results

Table of demographic variables:

```
##
##           level  Overall
##   n               37,531
##   sex (%) Female  21980 (58.6)
##           Male    15551 (41.4)
```

The average political score estimated by poststratification is 5.19.

```
## # A tibble: 1 x 1
##   alp_predict
##         <dbl>
## 1        5.19
```

Model summary:

Males on average have a higher score (higher by 0.59226) than females. So males are more likely to vote for rightwing parties than females.

Table 1: Fitting linear model: lr ~ sex

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.922	0.01912	257.4	0
sexMale	0.5923	0.0286	20.71	1.33e-94

Discussion

We used a simple linear regression model, powerful in its own way, to estimate the average political score. We found that males have a statistically higher score on average than females, thus they are more likely to vote for rightwing parties. Because this is an univariate analysis, we did not account for confounders. Next steps include using multivariable linear regression modeling to account for more predictors.

The next steps also include checking the four linear regression model assumptions.

1) Linearity
2) Constant variance
3) Uncorrelated errors
4) Normal errors

Since we are fitting a linear regresion model, we assume that the relationship truly is linear, and that the errors, or residuals, are simply random variations around the true regression linear.

Next, We assume that the variation in the response variabl does not increase or decrease as the value of the predictor increases. This is the assumption of equal variance or homoscedasticity.

More over, we assume that the observations are independent of each other. Correlation between sequential observations, or auto-correlation, can be a problem with time series data. That is, the data have a natural time-ordering.

So how do we check regression assumptions? We can look at the residuals vs. fitted values plot to assess the linearity assumption. If points on the plot scatter randomly around the horizontal line at y = 0, the the linearity assumption is met.

We can look at the scale-location plot and if there are not fanning-out patterns, then the homoscedasticity assumption is satisfied. Any sort of irregular pattern is an indication of violation of this assumption.

If there are no clusters of points on the residuals vs. fitted values plot or the scale-location plot, the uncorrelated errors assumption is met.

Finally, to assess the normality assumption, we would look at the Normal quantile quantile plots of the standardized residuals. If the points lie closely to the 45-degree line, then the normality assumption is met. As long as there are no extreme deviations from the 45-degree line, the normality assumption can assumed to be satisfied since with a large sample the Central Limit Theorem would force the parameter estimates and the residuals to be normally distributed, allowing robust statistical inference against non-normality.

References

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

Linear regression. (2020, December 17). Retrieved from https://en.wikipedia.org/wiki/Linear_regression#:~:text=In statistics, linear regression is,as dependent and independent variables).

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

2019 Canadian federal election. (2020, December 17). Retrieved December 22, 2020, from https://en.wikipedia.org/wiki/2019_Canadian_federal_election