Post-Stratification to Predict Election Results

STA304 - Assignment 2

Alan Yue

November 24, 2022

Introduction

The Canadian Election Study (CES) is an independent survey conducted each election year on voting issues and preferences [1]. The Canadian General Social Survey (GSS) is comprised of a series of annual surveys conducted by Statistics Canada on different social aspects of life [2]. The question becomes how we can leverage the larger sample of the latter survey to get more accurate estimates of some variable of interest in the former. In particular, we want to forecast the party that receives the most votes in the 2025 Canadian federal election.

The effort to predict election results has long been a staple of political science and study design research and application, not least of which because it's helpful to know in advance who will be given the power to govern us. Election polling can reveal how demographic information is correlated with voting preferences, giving us insight as to the attitudes certain groups of people have towards political issues. This can help us and politicians to focus on how to solve these issues in a way that appeal to the public. It is thus of great benefit if we are able to obtain equally accurate estimates of voting preferences while conducting surveys with smaller samples, saving time, money, and organizational effort.

We will explore a technique for doing this called post-stratification, which is a method for scaling estimates broken down by group by the prevalence of that group in the larger population. We will be using the 2019 CES phone survey [3] and the 2017 GSS family survey data [4] in our analysis. In the following sections, we will take a closer look at this data before building our predictive model and explaining the post-stratification process in further detail.

Because the 2019 Canadian federal election had the popular vote go to the Conservative party over the other two major parties (albeit by a slim margin) [5] Liberal and NDP, we hypothesize that our model will predict the Conservative party to win the most votes in 2025.

Data

The CES data we're using was collected through telephone interviews using a form of random digit dialing [6]. A minimum of six initial attempts were made to elicit a response before moving on from a generated number. Respondents had to be 18 or older and a Canadian citizen to be included in the study. The GSS data was also collected through telephone using a combined frame of telephone numbers and Statistics Canada's Address Register [7]. Provinces were stratified by region, within which a simple random sample without replacement was taken.

Because we intend to post-stratify with the ultimate end goal of forecasting the party that will receive the most votes, we want to consider as many variables as possible that both the CES and GSS measure. This allows us maximum freedom in picking our model. There were a handful of these measurements such as employment status, household income, and marital status for which there was a significant proportion of missing data. We chose to ignore these because we want to make predictions based on all the variables in our model. We thus only keep variables for which data exists in both datasets and that requires us to remove a minimum number of observations with missing data.

We then created shared categories between variables across the datasets so that we are able to post-stratify the sample data using the same categories as the GSS data after fitting the models. For this we had to use the CES study documentation [8] because much of the data were recorded as numbers that corresponded to a particular category rather than the category itself. Finally, we created binary indicator variables that tell us if a sampled individual intended to vote Liberal, Conservative, or NDP in the 2019 Canadian general election.

We end up with the following variables from the CES data we will explore as predictors in our model.

Age: We treat age as a continuous variable here despite the fact that the CES survey records age as a discrete whole number so as to not lose accuracy from binning. The minimum age in the GSS data is 15, but because this survey was conducted in 2017, all respondents will be of voting age by 2025.

Sex: Male or Female. The CES survey records gender and does allow respondents to choose an "other" option, but because only 1 person chose this option in the survey and because the GSS records a binary sex, we will pick it as our variable.

Educational level:

"< High School" for highest education below high school completion

"High School" for high school completion or equivalent but no higher

"< Bachelor's" for trade certificates, community college, CEGEP, etc.

"Bachelor's" for completion of a Bachelor's degree but no higher

"> Bachelor's" for anything higher than a Bachelor's Degree e.g. Master's, doctorate, etc.

Religion: Religious, Non-religious/Atheist, or Agnostic/unsure

Province: 1 of 10 Canadian provinces, excluding territories. Both the CES and GSS data have no recorded observations from the territories.

Table 1: Province and Voting Preference Statistics

Province	Count	Prop Liberal	Prop Conservative	Prop NDP
Quebec	782	0.2391304	0.1099744	0.0664962
British Columbia	783	0.1954023	0.2298851	0.1519796
Ontario	796	0.3178392	0.2311558	0.1180905
Alberta	280	0.0892857	0.5500000	0.0714286
Manitoba	255	0.2078431	0.3803922	0.1019608
Saskatchewan	259	0.1081081	0.4478764	0.1119691
Newfoundland and Labrador	195	0.2820513	0.1692308	0.1538462
Prince Edward Island	196	0.2857143	0.2040816	0.0357143
Nova Scotia	194	0.2731959	0.2113402	0.1185567
New Brunswick	197	0.2335025	0.2487310	0.0253807

In Table 1, we have the proportion of those from the sample intending to vote Liberal, Conservative, and NDP broken down by province. We see that proportions vary across provinces, suggesting that we have a group level correlation on voting preference that we may need to account for.

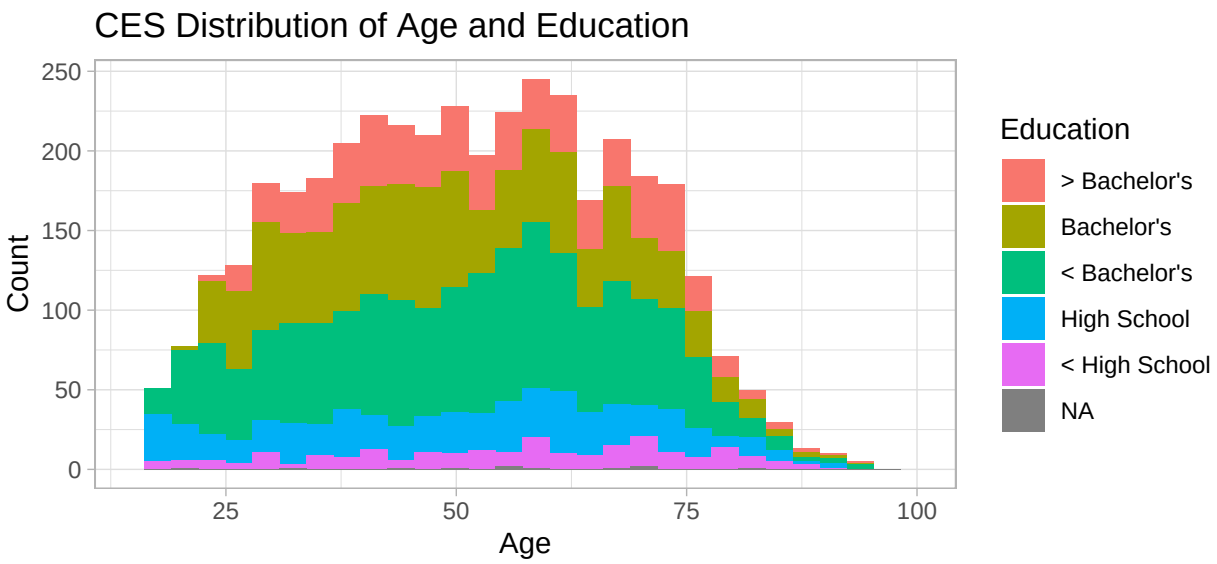

Figure 1: CES Histogram

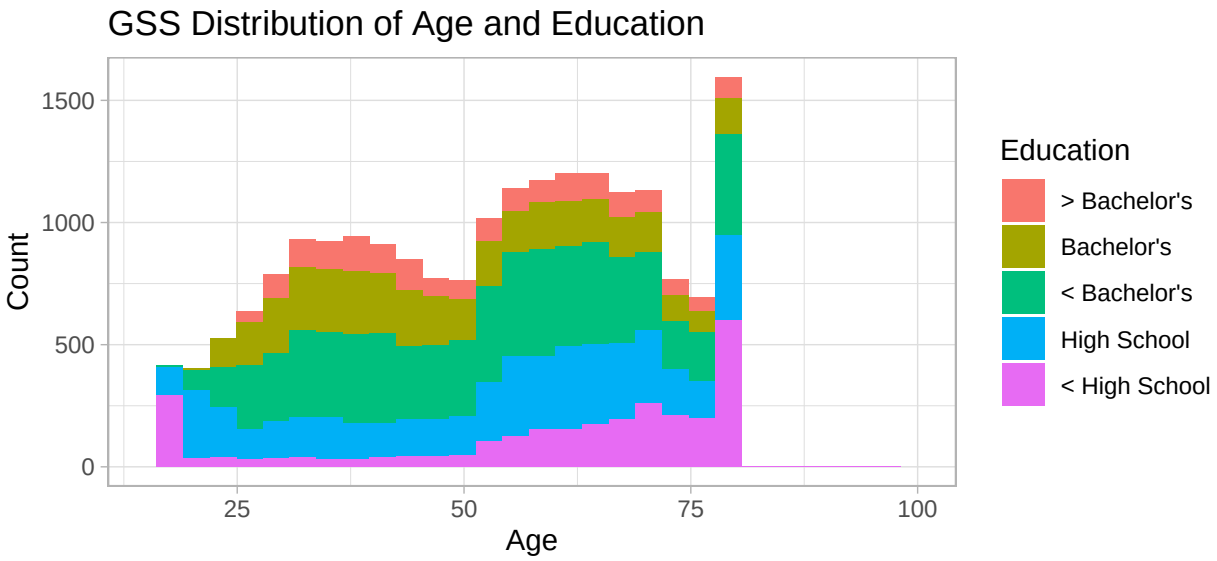

Figure 2: GSS Histogram

The reason for the large rightmost bin in Figure 2 is because the GSS caps their age measurements at 80. The GSS data appears to have a greater proportion of older people with a lower educational level, as seen by the larger area of the blue and purple regions in the second histogram. If we were to not use post-stratification in our methods section, we may undervalue the voting preference of this group of individuals and end up with more inaccurate estimates of total voting proportions.

Methods

Model Specifics

Because we're trying to model whether an individual intends to vote for a party, we will be using logistic regression, as this outcome is binary for a particular party. Through variable selection, we found that each of the covariates we compiled in the previous section had at least one category present as a significantly significant fixed effect for one of the parties, so we decide to include the full set in our model. This is to maximize predictive power, as well as to ensure our model is consistent across parties. This helps with interpretability because we essentially only need one model equation. That equation is:

$$log(\frac{p_{iv}}{1-p_{iv}}) = \beta_0 + \beta_1 x_{i,age} + \beta_{i,sex} + \beta_{i,education} + \beta_{i,religion} + r_j$$

where p_{iv} is the probability of person i voting for party v,

β_0 is the fixed part of the intercept,

β_1 is the slope for age,

$x_{i,age}$ is the age in years of the ith person,

$\beta_{i,sex}$ is a value depending on the sex of the ith person,

$\beta_{i,education}$ is a value depending on the educational level of the ith person,

$\beta_{i,religion}$ is a value depending on the religious affiliation of the ith person,

and $r_j \sim N(0, \sigma_r^2)$ is the random effect of province j. We model this as a random effect because the provinces can be seen as groups that those with different individual characteristics are picked from. It's easy to conceive of some provinces having people more disposed to vote a certain way, shifting up or down the log odds.

It's important to note that all terms on the right-hand side of the equation, except for $x_{i,age}$ is also dependent on the party v. We are essentially fitting three different models for the three parties under consideration, but taking the same approach in doing so. We drop the v in the subscripts for these terms for notational simplicity, but it should be remembered that $\beta_{i,sex}$ is different depending on whether we're predicting for the Liberal or Conservative parties. The categorical variables are denoted with one term each also for notational simplicity instead of having a coefficient and dummy variable for every category.

Because we have 5 covariates for each of 3 fitted models, we will not show diagnostic plots for model assumptions for the sake of simplicity, but we assume the left-hand side of the equation or log odds has a linear relationship with the covariates on the right-hand side of the equation. We also assume a lack of strongly influential outliers and lack of multicollinearity (strong correlation amongst the fixed effect variables).

Post-Stratification

After fixing our models, we have a way to predict the voting patterns of an individual based on their age, sex, educational level, religious affiliation, and province. It seems reasonable that these personal characteristics would be correlated with voting preference, but it's possible the proportions of people in different category combinations in the sample do not reflect the actual population. To get a more accurate representation of these proportions, we can leverage the higher number of observations of the GSS data.

First we count the number of people who fall in each combination of demographic categories in the GSS data. Because we have 10 provinces, 2 sexes, 5 educational levels, and 3 religious categories, this corresponds to 300 counts, which is reduced to 266 after the removal of missing data. Because age is not a category but a continuous variable in our model, we associate each combination with the mean age of people who fall in that combination of categories. Next we use our three fitted models to predict the probability someone in each category combination will vote for the Liberal, Conservative, and NDP parties, remembering to convert

the output of the model from log odds to probability. Finally we compute the following estimate for the proportion of people voting for each party v:

$$\hat{p}_v = \frac{\sum_{k=1}^{K} N_k \hat{p}_{v,k}}{\sum_{k=1}^{K} N_k}$$

where K is the number of category combinations,

N_k is the number of people in the GSS data in category combination k,

and $\hat{p}_{v,k}$ is the predicted probability of someone in category k voting for party v.

The formula can be interpreted as follows: the numerator is the estimated number of people across all category combinations that intend to vote for party v. Dividing this by the total number of people gives us an estimate for the proportion of total people voting for party v.

Results

Below are tables of the terms for each of the three models. The base educational level is less than completion of high school and the base level for religion is non-religious/atheist.

Table 2: Liberal Model

Term	Estimate	Std. Error	Statistic	p value
Intercept	-2.0676862	0.2668561	-7.7483203	0.0000000
Age	0.0075835	0.0024469	3.0992716	0.0019400
Female	0.0450621	0.0789207	0.5709795	0.5680136
High School	0.1888561	0.2112345	0.8940590	0.3712903
< Bachelor's	0.1460918	0.1922756	0.7598039	0.4473718
Bachelor's	0.6112470	0.1935866	3.1574860	0.0015914
> Bachelor's	0.7371685	0.2007485	3.6721002	0.0002406
Religious	-0.0057092	0.0871623	-0.0655003	0.9477757
Agnostic	-0.2269209	0.2912707	-0.7790720	0.4359373
SD(Province)	0.4196771	NA	NA	NA

Table 3: Conservative Model

Term	Estimate	Std. Error	Statistic	p value
Intercept	-1.4498797	0.2978660	-4.8675575	0.0000011
Age	0.0057673	0.0024421	2.3616287	0.0181949
Female	-0.6705006	0.0833198	-8.0473095	0.0000000
High School	0.1200950	0.1863978	0.6442940	0.5193848
< Bachelor's	0.0318069	0.1697551	0.1873692	0.8513712
Bachelor's	-0.2358216	0.1771725	-1.3310284	0.1831797
> Bachelor's	-0.5407003	0.1946577	-2.7776982	0.0054745
Religious	0.7211272	0.0926978	7.7793370	0.0000000
Agnostic	-0.2797087	0.3348798	-0.8352509	0.4035765
SD(Province)	0.6570325	NA	NA	NA

Table 4: NDP Model

Term	Estimate	Std. Error	Statistic	p value
Intercept	-1.1545762	0.3799994	-3.0383632	0.0023787
Age	-0.0290956	0.0035337	-8.2337529	0.0000000
Female	0.4521324	0.1100344	4.1090093	0.0000397
High School	0.3373353	0.3270500	1.0314489	0.3023304
< Bachelor's	0.3578951	0.3049899	1.1734652	0.2406093
Bachelor's	0.2006232	0.3115975	0.6438537	0.5196703
> Bachelor's	0.4959857	0.3211201	1.5445489	0.1224554
Religious	-0.5655935	0.1161013	-4.8715534	0.0000011
Agnostic	0.2622933	0.3043119	0.8619225	0.3887302
SD(Province)	0.5327825	NA	NA	NA

Notice that it's a bit difficult to interpret the estimates directly because the models output log odds, but we can see that as age increases, the predicted probability of voting NDP decreases. Females are less likely to vote conservative, as are those who have a Bachelor's degree or higher, whereas those who are religious seem more likely. As educational level increases, people seem more likely to vote both Liberal and NDP.

After running the above post-stratification procedure, we get:

$\hat{p}_{lib} = 0.2248$, the estimated total proportion voting Liberal

$\hat{p}_{con} = 0.2685$, the estimated total proportion voting Conservative

$\hat{p}_{ndp} = 0.0869$, the estimated total proportion voting NDP

Our model predicts 22.48% of people to vote Liberal, 26.85% to vote Conservative, and 8.69% to vote NDP.

Conclusions

To recap what we've done so far, we wanted to forecast the party that receives the most votes in the 2025 Canadian federal election with the hypothesis that our model predicts the Conservative party to win the most votes. We fitted three logistic regression models for the three major Canadian parties and used them to predict the probability someone in a demographic category combination will vote for a particular party. Finally we post-stratified these category combinations to get more appropriate estimates using demographic data from a larger sample survey.

Our model predicted that the Conservative party would get the most votes of all three major parties in the 2019 Canadian federal election, which did indeed happen [5]. Thus using our model and post-stratification we predict the Conservative party to win the most votes in the 2025 election. The viability of these methods would let us make equally accurate estimates with smaller surveys if we have a larger dataset we can use for post-stratification.

The estimates of our terms in Tables 2-4 aren't too surprising, as they follow popularly held beliefs about demographic voting tendencies. Females and those with higher education tend to have left-leaning voting preferences, while those who are older and more religious tend to have right-leaning voting preferences.

An obvious weakness of our model is that it is based on old data. 2019 voter preferences may be quite outdated by the time 2025 rolls around and the GSS data is even older. Our model did accurately predict the winner of the 2019 election based solely on pre-election data, but in trying to extrapolate this result to 2025, we could instead just take the winner of the last election as our prediction. Additionally, because the GSS dataset was also gathered through a phone survey and not a proper census, it's difficult to say whether the proportions of people in each category combination are more accurate just because there's more data there.

Because we required the larger GSS dataset to perform post-stratification, the methods outlined in this paper are not applicable in the absence of voluminous demographic or other categorical data that coincides with the covariates of interest in the smaller study being conducted. We also shouldn't scale our estimates to the proportions of category combinations in the larger dataset if that data was collected in a flawed way that introduced some kind of bias into play. In that case, the larger volume of data would be of little benefit.

Future studies could attempt a similar procedure outlined here with more current data leading up to an election, i.e. in 2025 conducting a smaller survey using post-stratification with 2025 census information to predict the result of the election. This would give a more practical, current, and likely more accurate application of the procedure. Future studies also shouldn't be restricted to just election prediction. Post-stratification is a general procedure with rich potential to be applied in a variety of domain areas.

Bibliography

1. Canadian Election Study. (2019). Retrieved Nov. 24, 2022 from http://www.ces-eec.ca/

2. "General Social Survey: An Overview." (Feb. 20, 2019). *Statistics Canada.* Retrieved Nov. 24, 2022 from https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2019001-eng.htm

3. Laura B Stephenson et al. "2019 Canadian Election Study - Phone Survey." *Harvard Dataverse,* V2. 2020. DOI: https://doi.org/10.7910/DVN/8RHLG1

4. "General Social Survey - Family (GSS)." *Statistics Canada.* Retrieved Nov. 24, 2022 from https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=335816

5. Peter Zimonjic. "Liberals take losses but win enough in Quebec and Ontario to form minority government." (Oct. 21, 2019). *CBC News.* Retrieved Nov. 24, 2022 from https://www.cbc.ca/news/politics/federal-election-results-2019-cbc-leaders-1.5329485

6. Laura B Stephenson et al. "2019 Canadian Election Study - Phone Survey Technical Report." *Consortium on Electoral Democracy*, 2020. DOI: https://doi.org/10.7910/DVN/8RHLG1/1PBGR3

7. "General Social Survey, Cycle 31: Families, Public Use Microdata File Documentation and User's Guide." *Diversity and Sociocultural Statistics,* 2020.

8. Laura B Stephenson et al. "Canadian Election Study, 2019, Phone Survey Study Documentation." (Aug. 11 2020).