Predicting 2019 Canadian Federal Election Results If All Eligible Individuals Voted

Maxwell Garrett

Decmeber 7th 2020

Github URL

https://github.com/maxwell-garrett/final-assessment

Abstract

Data from the General Social Survey (2017) (Statistics Canada 2017) and Canadian Election Study (2019) Phone (Stephenson et al. 2020) datasets are used to build logistic regression models with post-stratification of the popular vote results for Canadian political parties. The popular vote for the 2019 federal Canadian election is predicted and compared with the actual recorded vote ("43rd General Election: Official Voting Results (Raw Data)" 2019).

Keywords

Keywords: Post-stratification, logistic regression, Canadian election, voter turnout

Introduction

It is important for politicians to know how they can best represent the interests of their constituents. Canadian election polling data provides important information on voting preferences of Canadians on important issues (Stephenson et al. 2020). It can be difficult to obtain large and representative samples of citizens' voting preferences. Often due to time and resource constraints, the sampling procedure results in certain subpopulations being under- or over-represented, with respect to the target population. In addition, commonly used election survey methods, such as telephone surveys, face nonresponse and undercoverage issues (Lipps, Pekari, and Roberts 2015). Therefore, methods for standardizing the sampled data so that it is representative of the broader population are important in order to produce reliable results.

In this report, we use the technique of post-stratification with logistic regression models for predicting the 2019 Canadian election outcome. This technique is used to make the sampled survey data representative of the population. We will use Statistics Canada's General Social Survey administered in 2017 as census data to use in post-stratification weighting procedures. For example, if a province was underrepresented in the survey dataset relative to the target population then the weighting procedure will adjust for this. As well, we will use Canadian Election Study's 2019 phone dataset as survey data to understand the voting intentions of individuals in different subgroups of the population.

The two datasets mentioned above will be used to build counterfactual models predicting the potential Canadian election results if all eligible voters in Canada had voted in the 2019 election. We will be specifically modeling the popular vote. We are interested in investigating if the popular vote would significantly change in the hypothetical scenario where all eligible voters participated in the election. In the methodology section, we describe the data, and the models of the election intentions. We then show our resulting election predictions based on the models fit in the results section. Finally, we discuss what our results indicate about the importance of increasing voter turnout.

Methodology

In this section, we will first explain features of the datasets that we are using and then we will explain the popular vote proportion models we have developed.

Data

The first dataset we discuss is the General Social Survey (2017) dataset. We will be using this dataset to calculate representative population proportions to use in post-stratification. We selected sex, place_birth_canada, and province as co-variates of interest. This dataset's target population is all individuals above the age of 15 (inclusive) in Canada excluding individuals in the Yukon, Northwest Territories, and Nunavut (Beaupré 2020). As well, individuals in full-time institutions are excluded (Beaupré 2020). The dataset has been filtered to remove individuals below the age of 18 because in Canada an individual must be at least 18 years old to vote.

Table 1: Characteristic table of selected variables from GSS data.

characteristic	percent
Female	54.50
Born in Canada	79.75
Alberta	8.38
British Columbia	12.28
Manitoba	5.79
New Brunswick	6.47
Newfoundland and Labrador	5.33
Nova Scotia	6.98
Ontario	27.17
Prince Edward Island	3.42
Quebec	18.60
Saskatchewan	5.58

The variable sex, an individual's biological sex, was selected for analysis as there is no missing values for this variable and this variable has similar categories in the CES dataset allowing for straightforward post-stratification. Table 1 tells us that approximately 55% of the sample is female, the rest male. The variable place_birth_canada, indicating if an individual was born in Canada, was selected as we thought that the experience of an individual immigrating to Canada may have an impact on their voting preference. As well, this variable has very few missing values which means we keep almost all of the information from the dataset. Some missing values were replaced if the respondent had indicated the macroregion which they were born in. In this case, if the macroregion did not include Canada then they were marked as 'Born outside of Canada'. Approximately 80% of the respondents indicate that they are born in Canada according to Table 1. The last variable selected is province which is the current province that the respondent resides in. We selected this as we thought that location where an individual lives may play into their voting preference. As well, this variable is missing zero values and is similar to the province categories provided in the CES dataset. A drawback of this variable is that it unfortunately does not include territories while the CES dataset does, this means that we do not include territories in our analysis. In Table 1, we see the proportion of individuals reporting to reside in each province.

The second dataset we discuss is the Canadian Election Study (2019) phone survey (Stephenson et al. 2020). We will be using this dataset to form our model on voting preference. This survey was performed over phone through random sampling of Canadian phone numbers (Stephenson et al. 2020). The phone calls were performed the day after the election (Stephenson et al. 2020). The variables selected from this dataset were sex, place_birth_canada, province, and voting_pref. This dataset only includes individuals eighteen years old or older (Stephenson et al. 2020).

Table 2: Characteristic table of selected predictor variables from CES data.

characteristic	percent
Female	42.52
Born in Canada	85.20
Alberta	7.21
British Columbia	20.54
Manitoba	6.77
New Brunswick	4.84
Newfoundland and Labrador	4.62
Nova Scotia	4.93
Ontario	20.70
Prince Edward Island	4.87
Quebec	18.73
Saskatchewan	6.77

The variable sex was created based on the variable q3 (gender) in the dataset. This variable was reduced to sex as the method for determining gender was not sufficient, because it only involved the interviewer deducing the respondent's gender based on voice pitch (Stephenson et al. 2020). This variable was chosen as it is easily modified to fit with the values found for sex in the GSS dataset. In Table 2 we see that approximately 43% of the respondents were identified to be female. The variable place_birth_canada was created based on the variable q64, the country of birth, by reducing the variable to be binary, indicating only if an individual was born in Canada or not. This was done to correspond with the GSS dataset and allow easier computations later on. It was found that 85% of the respondents were born in Canada according to Table 2. The province variable was created based on the variable q4, the province where they reside. Respondents who indicated living in Nunavut, Northwest Territories, and Yukon were removed as the census dataset does not include those regions. In Table 2, we see the percent of the total respondents surveyed that reside in each province. The variable voting_pref indicates the reported voting preference of the respondent. It was created based on q11 and q12 which describe the voting preference of the respondent, q12 data was used if q11 data was missing. The variable has popular party names indicated and then all other responses indicated as other.

Model

We are looking to predict the popular vote for the 2019 Canadian election had all potential voters in Canada voted. This will be modeled using multiple logistic regression models, one for each party and one for spoiled ballots. For each model, we have a binary response variable created which indicates 1 if the respondent plans to vote for that option and 0 otherwise.

Our logistic regression models will model the log-odds of voting for a party. For example, one model will be modeling the log-odds of voting for the Liberal Party of Canada. There will be a model for voting Liberal, Conservative, NDP, BQ, Green Party, People's Party, other, and spoiling the ballot. The model formula is the following with all covariates being indicator variables taking on the values 0 or 1:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_{11} x_{11} + \epsilon$$

where x_1 is if sex equals male, x_2 is if place_birth_Canada equals 'Born outside of Canada', and x_3 to x_{11} are the provinces of Canada excluding Nunavut, Yukon, and Northwest Territories. The province of Alberta, sex equaling female, and place_birth_canada equaling 'Born in Canada' are set as the reference group. As well, p represents the percent of votes for a party conditional on the covariate values of sex, birth place, and province. The intercept (β_0) represents the estimated log-odds of voting for a particular party for an individual in the reference group (i.e. an individual who is female, born in Canada, and residing in Alberta). Thus, the exponentiated intercept (e^{β_0}) represents the estimated odds of voting for a particular party for an individual in the reference group.

The regression coefficients (β_1 to β_{11}) represent the expected change in the log odds of voting for a particular party (depending on the model) for a unit increase in the corresponding predictor variable, holding the other predictor variables constant at a certain value. Therefore, exponentiating the coefficients represents the expected odds ratio. Additionally, since all of our predictor variables are categorical, a one unit increase means comparing an observation in the reference category to an observation in a different category (e.g. comparing individuals residing in Alberta to individuals residing in Ontario). For example, e^{β_1} represents the expected odds of a male individual ($x_1 = 1$) voting for a particular party, over the odds of voting for a particular party for females ($x_1 = 0$), keeping the other covariates constant (for example, comparing only individuals residing in Ontario, and born outside of Canada).

The model was selected as the predictor variables used are available in both datasets and they contain low percentage of missing values. As well, it is logically plausible that the predictor variables are associated with voting preference as they subset the population into large groups that share some common characteristics.

Post-stratification

We use the census dataset (GSS) to find the number of respondents in each bin of the population. There are forty bins created total, each bin based on sex, birth place, and province. For each model, we predict the proportion of a given bin that would vote for the party or submit a spoilled ballot. We multiply this proportion by the number of respondents in the bin and divide by the total number of respondents in GSS. We then add up all the proportion estimates for bins, and this is our estimate for the proportion of the population voting for a political party. As well, we use the standard error for each bin estimate from the model to calculate a 95% confidence interval of the proportion of voters for a given party (or submitted a spoiled ballot).

Results

Table 3: Predicted percent of popular vote using logistic regression models.

Voting Preference	Percent of The Popular Vote	Lower Limit	Upper Limit
Liberal	35.72	30.16	41.27
Conservative	31.99	26.73	37.25
NDP	15.34	11.14	19.54
BQ	3.58	2.66	4.51
Green Party	10.54	7.04	14.03
People's Party	1.56	0.23	2.89
Other	1.19	-0.04	2.42
Spoil	3.58	2.66	4.51

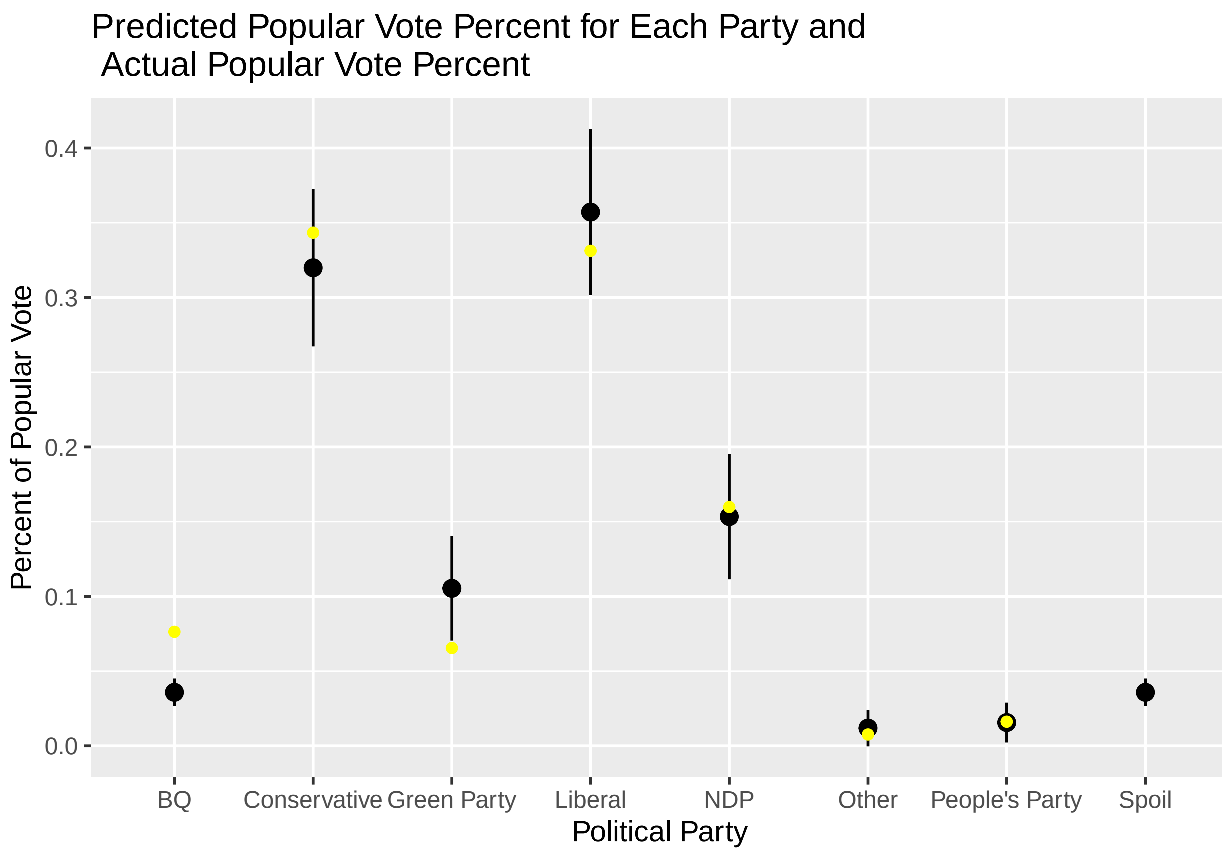
In Table 3, we have the predicted percent of the popular vote that each party would receive had everyone voted according to our models. As well, we have the lower and upper bounds of the 95% confidence intervals around these predicted popular votes. For the liberal party, we estimate the percent of the popular vote to be 35.72%. For the Conservative party, we estimate the percent of the popular vote to be 31.99%. For the NDP, we estimate the percent of the popular vote to be 15.34%. For the BQ, we estimate the percent of the popular vote to be 3.58%. For the Green Party, we estimate the percent of the popular vote to be 10.54%. For the People's Party, we estimate the percent of the popular vote to be 1.56%. We estimate that 1.19% of people vote for any other party option not listed. Finally, we estimate that 3.58% of voters spoil their ballot.

Table 4: Actual 2019 election results provided by Stats Canada.

Party Voted For	Percent of The Popular Vote
Liberal	33.12
Conservative	34.34
NDP	15.98
BQ	7.63
Green Party	6.55
People's Party	1.62
Other	0.76
Spoil	NA

Table 4 provides the actual observed popular vote results from the 2019 Canadian Election ("43rd General Election: Official Voting Results (Raw Data)" 2019). We lack information on the number of spoiled ballots in this dataset.

Figure 1: Plot of predicted popular vote confidence intervals, and the actual popular vote collected in the 2019 election.



In figure 1, we have the predicted popular vote percentages plotted in black. As well, we have the 95% confidence interval around the predicted value marked with the black line. The actual popular vote percentages from the election results are marked in yellow points. We have no data for spoiled ballots therefore there is no yellow point for that category. We see that the proportion of votes for the Conservative, Liberal, NDP, Other, and People's Party are quite similar to our models predictions. The observed results for the parties mentioned above fall within the 95% confidence interval. However, we notice that BQ received a higher proportion of votes than what was expected if all eligible voters participated in the election. As well, the Green Party was expected to receive a higher proportion of overall votes than what was actually observed.

Discussion

Summary

In summary, we created eight logistic regression models predicting the proportion of voters for each party for each population group. We then used the technique of post-stratification to estimate the proportion of voters for each party across Canada. Finally, we visually compared the predicted election result confidence intervals with the actual election results received in 2019 in a plot.

Conclusions

In our results, we see the scenario where every eligible citizen in Canada participates in the election excluding residents of Nunavut, Northwest Territories, and Yukon. In figure 1, we compare predicted results with the oserved results from the 2019 federal election. There is a noticeable difference in the predicted results from the actual results for the Bloc Quebecois Party and the Green Party. For the BQ Party, we predict less proportional support had everyone voted while the Green Party we predict more proportional support had everyone voted. Other parties seem to have quite similar results. However, based on the discrepancy in the observed versus predicted results for the BQ Party and Green Party, we conclude that there is an impact from having every eligible citizen vote in the election.

As well, based on these results we would predict a Liberal win of the popular vote, as the predicted popular vote is higher than all others. This differs from the actual popular vote winner, which was the Conservative Party. This supports our conclusion that there is a noticeable impact of having all eligible voters vote.

In the 2019 election, there was a voter turnout of 67% of Canadians ("Appendix – Report on the 43rd General Election of October 21, 2019" 2019). In our models, we are assuming a turnout of 100%. This research area is of importance as if turnout is low there may be sub-populations underrepresented by the parties elected. As we want representative democracies, we want voter turnout as close to 100% as possible.

Weaknesses & Next Steps

The first important weakness to consider is that our census dataset (GSS) did not include the territories of Canada, Nunavut, Northwest Territories and Yukon (Beaupré 2020). As well, the GSS dataset does not include institutionalized individuals (Beaupré 2020). Both these groups include many people who are eligible voters. By excluding these groups in the data, our models are not able to fully represent the target population of all eligible voters. A future step would involve finding a dataset that includes observations from these territories This would greatly expand the ability of the models to predict the voting outcomes for the entire country.

The second important weakness to consider is that we are looking at the popular vote, but this does not necessarily provide information on the actual election outcome due to the voting system. The voting system involves each riding in Canada electing a party representative ("FAQs - General Questions" n.d.). This means that our popular vote predictions do not reflect who would actually be elected in the hypothetical scenario where all eligible voters participated in the election. A future step would involve finding a dataset that includes voting preferences with riding information on each respondent. This would allow us to accurately predict election outcomes for each riding and therefore more accurately predicting the election result.

The third important weakness is the handling of sex and gender in both datasets. The CES dataset has the phone operator record gender, but only ask for the gender of an individual when unsure (Stephenson et al. 2020). This is not an accurate way to record sex or gender as voice is not a reliable indicator of this. An improvement to this problem would be finding a dataset that collects gender through asking directly in the interview. This way an accurate answer by the respondent is given.

A fourth weakness is the limited number of covariates considered in the models. More covariates would potentially explain more variation in the data. Future work would include incorporating additional covariates into the analysis, and dealing with missing values using imputation techniques (White, Royston, and Wood 2011).

Additionally, further statistical work on modeling could include using a multinomial regression for this multiclass problem instead of running multiple individual binary logistic regressions (Agresti 2003). This technique has the advantage of potentially reducing standard errors, and therefore allowing for more precise estimates (Agresti 2003).

Appendix

Liberal Party Model

term	estimate	std.error	statistic	p.value
(Intercept)	-1.84	0.20	-9.17	0.00
sexMale	-0.23	0.08	-3.00	0.00
place_birth_canadaBorn outside Canada	0.63	0.11	5.97	0.00
provinceBritish Columbia	0.82	0.21	3.87	0.00
provinceManitoba	0.98	0.24	4.00	0.00
provinceNew Brunswick	1.25	0.26	4.83	0.00
provinceNewfoundland and Labrador	1.61	0.25	6.32	0.00
provinceNova Scotia	1.57	0.25	6.24	0.00
provinceOntario	1.53	0.21	7.38	0.00
provincePrince Edward Island	1.66	0.25	6.59	0.00
provinceQuebec	1.35	0.21	6.44	0.00
provinceSaskatchewan	0.35	0.26	1.33	0.18

Conservative Party Model

term	estimate	std.error	statistic	p.value
(Intercept)	0.52	0.15	3.43	0.00
sexMale	0.58	0.08	7.06	0.00
place_birth_canadaBorn outside Canada	-0.25	0.12	-2.12	0.03
provinceBritish Columbia	-1.67	0.17	-10.00	0.00
provinceManitoba	-0.95	0.20	-4.80	0.00
provinceNew Brunswick	-1.38	0.22	-6.16	0.00
provinceNewfoundland and Labrador	-1.78	0.23	-7.74	0.00
provinceNova Scotia	-1.80	0.23	-7.81	0.00
provinceOntario	-1.67	0.17	-10.07	0.00
provincePrince Edward Island	-1.78	0.23	-7.79	0.00
provinceQuebec	-2.35	0.18	-13.18	0.00
provinceSaskatchewan	-0.53	0.20	-2.63	0.01

NDP Model

term	estimate	std.error	statistic	p.value
(Intercept)	-1.79	0.22	-8.31	0.00
sexMale	-0.47	0.10	-4.72	0.00
place_birth_canadaBorn outside Canada	-0.07	0.14	-0.47	0.64
provinceBritish Columbia	0.74	0.23	3.23	0.00
provinceManitoba	0.29	0.28	1.01	0.31
provinceNew Brunswick	-1.34	0.50	-2.66	0.01
provinceNewfoundland and Labrador	0.84	0.29	2.95	0.00
provinceNova Scotia	0.39	0.30	1.30	0.19
provinceOntario	0.41	0.23	1.73	0.08
provincePrince Edward Island	-0.67	0.39	-1.72	0.09
provinceQuebec	0.09	0.24	0.37	0.71

term	estimate	std.error	statistic	p.value
provinceSaskatchewan	0.37	0.28	1.31	0.19

BQ Party Model

term	estimate	std.error	statistic	p.value
(Intercept)	-21.48	0.15	-145.53	0.00
sexMale	0.08	0.21	0.39	0.70
place_birth_canadaBorn outside Canada	-1.61	0.47	-3.41	0.00
provinceBritish Columbia	0.09	0.08	1.12	0.26
provinceManitoba	-0.02	0.10	-0.16	0.87
provinceNew Brunswick	-0.08	0.11	-0.70	0.48
provinceNewfoundland and Labrador	-0.08	0.11	-0.75	0.45
provinceNova Scotia	-0.09	0.11	-0.80	0.43
provinceOntario	0.06	0.08	0.79	0.43
provincePrince Edward Island	-0.07	0.11	-0.67	0.50
provinceQuebec	20.16	0.13	160.16	0.00
provinceSaskatchewan	-0.10	0.10	-0.97	0.33

Green Party Model

term	estimate	std.error	statistic	p.value
(Intercept)	-3.18	0.36	-8.75	0.00
sexMale	-0.16	0.11	-1.38	0.17
place_birth_canadaBorn outside Canada	-0.52	0.18	-2.89	0.00
provinceBritish Columbia	1.81	0.37	4.84	0.00
provinceManitoba	0.92	0.44	2.12	0.03
provinceNew Brunswick	2.07	0.41	5.06	0.00
provinceNewfoundland and Labrador	0.55	0.50	1.10	0.27
provinceNova Scotia	1.40	0.43	3.25	0.00
provinceOntario	1.04	0.38	2.71	0.01
provincePrince Edward Island	1.76	0.42	4.23	0.00
provinceQuebec	1.03	0.39	2.68	0.01
provinceSaskatchewan	0.02	0.51	0.04	0.97

People's Party Model

term	estimate	std.error	statistic	p.value
(Intercept)	-4.28	0.58	-7.34	0.00
sexMale	0.42	0.29	1.44	0.15
place_birth_canadaBorn outside Canada	-0.27	0.44	-0.62	0.54
provinceBritish Columbia	0.43	0.57	0.75	0.45
provinceManitoba	-1.35	1.12	-1.20	0.23
provinceNew Brunswick	0.09	0.78	0.12	0.91
provinceNewfoundland and Labrador	-14.53	0.51	-28.36	0.00
provinceNova Scotia	-1.00	1.13	-0.89	0.37
provinceOntario	-0.22	0.61	-0.35	0.72

term	estimate	std.error	statistic	p.value
provincePrince Edward Island	0.13	0.78	0.16	0.87
provinceQuebec	-0.01	0.60	-0.01	0.99
provinceSaskatchewan	1.00	0.60	1.66	0.10

Other Parties Model

term	estimate	std.error	statistic	p.value
(Intercept)	-5.59	1.07	-5.22	0.00
sexMale	0.18	0.32	0.57	0.57
place_birth_canadaBorn outside Canada	0.23	0.40	0.57	0.57
provinceBritish Columbia	1.59	1.04	1.53	0.13
provinceManitoba	0.06	1.42	0.04	0.97
provinceNew Brunswick	1.12	1.23	0.91	0.36
provinceNewfoundland and Labrador	0.47	1.42	0.33	0.74
provinceNova Scotia	0.42	1.42	0.30	0.77
provinceOntario	1.02	1.06	0.96	0.34
provincePrince Edward Island	1.84	1.14	1.62	0.11
provinceQuebec	1.00	1.08	0.93	0.35
provinceSaskatchewan	1.20	1.16	1.03	0.30

Spoiled Votes Model

term	estimate	std.error	statistic	p.value
(Intercept)	-21.48	0.15	-145.53	0.00
sexMale	0.08	0.21	0.39	0.70
place_birth_canadaBorn outside Canada	-1.61	0.47	-3.41	0.00
provinceBritish Columbia	0.09	0.08	1.12	0.26
provinceManitoba	-0.02	0.10	-0.16	0.87
provinceNew Brunswick	-0.08	0.11	-0.70	0.48
provinceNewfoundland and Labrador	-0.08	0.11	-0.75	0.45
provinceNova Scotia	-0.09	0.11	-0.80	0.43
provinceOntario	0.06	0.08	0.79	0.43
provincePrince Edward Island	-0.07	0.11	-0.67	0.50
provinceQuebec	20.16	0.13	160.16	0.00
provinceSaskatchewan	-0.10	0.10	-0.97	0.33

References

"43rd General Election: Official Voting Results (Raw Data)." 2019. Elections Canada. https://www.elections.ca/res/rep/off/ovr2019app/51/data_donnees/table_tableau08.csv.

Agresti, Alan. 2003. *Categorical Data Analysis*. Vol. 482. John Wiley & Sons.

Alexander, Rohan, and Sam Caetano. 2020. "GSS Cleaning Code for 2017 Data."

"Appendix – Report on the 43rd General Election of October 21, 2019." 2019. Elections Canada. 2019. https://www.elections.ca/content.aspx?section=res&dir=rep/off/sta_ge43&document=app1&lang=e.

Beaupré, Pascale. 2020. *Cycle 31 : Families Public Use Microdata File Documentation and User's Guide*. Statistics Canada. http://www.chass.utoronto.ca/.

"FAQs - General Questions." n.d. Elections Canada. Accessed 2020. https://www.elections.ca/content.aspx?section=vot&dir=faq&document=faqgen&lang=e#gen1.

Henry, Lionel, and Hadley Wickham. 2020. *Rlang: Functions for Base Types and Core R and 'Tidyverse' Features*. https://CRAN.R-project.org/package=rlang.

Hodgetts, Paul A., and Rohan Alexander. 2020. *CesR: Access the Ces Datasets a Little Easier.*

Lipps, Oliver, Nicolas Pekari, and Caroline Roberts. 2015. "Undercoverage and Nonresponse in a List-Sampled Telephone Election Survey" 9 (2): 71–82.

Lumley, Thomas. 2004. "Analysis of Complex Survey Samples." *Journal of Statistical Software* 9 (1): 1–19.

———. 2010. *Complex Surveys: A Guide to Analysis Using R: A Guide to Analysis Using R*. John Wiley; Sons.

———. 2020. "Survey: Analysis of Complex Survey Samples."

Statistics Canada. 2017. "2017 General Social Survey: Families Cycle 31." http://www.chass.utoronto.ca/.

Stephenson, Laura B, Allison Harell, Daniel Rubenson, and Peter John Loewen. 2020. "2019 Canadian Election Study - Phone Survey." Harvard Dataverse. https://doi.org/10.7910/DVN/8RHLG1.

White, Ian R, Patrick Royston, and Angela M Wood. 2011. "Multiple Imputation Using Chained Equations: Issues and Guidance for Practice." *Statistics in Medicine* 30 (4): 377–99.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

———. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2020. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.