Provincial Voter Turnout to Turn Tides in 2019 Canadian Federal Election?*

Result based on popular votes calculated through multi-level regression with post-stratification and Bayesain Inference.

Samantha Wong

22 December 2020

Abstract

In every election, the voter turnout can heavily impact the outcome, which must be taken into account when forecasting election results. This paper makes use of multi-level regression with post-stratification in finding the probability of the Liberal Party of Canada winning the popular vote in the 2019 Canadian Federal Election from two subsets of the Canadian population, those who have the ability to vote and intend to vote and individuals residing in Canada over the minimum age to vote in the Canadian Federal Election. This paper predicted that 51.21% of actual voters would support the Liberal Party while 50.67% of all Canadians over age of 18 would vote for the Liberal Party. Next steps for analysis include finding the results for the electoral vote, using additional predictors such as income or using a time-series analysis and looking into the representation of immigrants across provinces and their possible votes. **Keywords:** 2019 CA Federal Election; forecasting; multi-level regression with post-stratification; Liberal; Conservative

1 Introduction

In this paper, the data obtained regarding the 2019 Canadian Federal Election is analyzed using statistical software R [R Core Team, 2020] with additional functions and packages: *brm* function from the **brms** package [Bürkner, 2017, Stan Development Team [2020]] and organized the data and results with the **tidybayes** [Kay, 2020] ,**tidyverse** [Allaire et al., 2019], **magrittr** [Bache and Wickham, 2014], **gridExtra** [Auguie, 2017], usmap [Di Lorenzo, 2020], **scales** [Wickham and Seidel, 2020], **haven** [Wickham and Miller, 2020], **broom** [Robinson et al., 2020] [**?**], and **here** [Müller, 2017] packages.

Over the years, political parties have sought to increase the turnout of the federal election in attempt to swing the election to their favor. For the past two elections in 2015 and 2019, the total percentage of the population that voted in the election was just over 77% [**?**] and had shown a significant increase since the previous election in 2011. The most recent US presidential election in 2020 had seen a tremendous increase in voter participation, due in part to social media advertising and motivation from influential public figures. Some popular reasons Canadians choose to not vote in elections are not being interested in politics, busy schedules or a disability, or disapproval of the electoral process. Although 77% still comprises most of the population, an extra 23% of the populations vote or even less can dramatically change the result of the federal election. In addition, many residents in Canada are permanent residents or have some other residency status other than citizenship. According to the 2016 Census, around 21.9% of the whole Canadian population are foreign born immigrants [**?**]. Although they reside in Canada, they are unable to vote in the election. Whether or not these individuals were included in the calculation of the percentage of Canadians who participated in the 2019 Federal Election was not made clear.

*Code and data are available at: [https://github.com/smwong88/CA_2019_Election_Prediction_Comparisons.git].

This paper uses data from the Canadian Election Survey conducted in 2019 and the 2017 iteration of the General Social Survey to perform Multi-level Regression with post-stratification. Two datasets from the Election survey will be made: responses of individuals who are eligible to vote and expressed interest to vote, all responses of individuals regardless of ability or interest to vote. MRP with post-stratification was performed on both of these datasets to compare the estimates on the forecast of the 2019 Federal Election. After performing MRP with post-stratification on both subsets of the Canadian Election Study, the Bayesian model had predicted that the just eligible voters would vote 51.21% in Justin Trudeau's favor, where the full results of the survey predicted 50.67% for the Liberal Party. Considering how Justin Trudeau had actually lost the popular vote in 2019 brings this number into perspective, with the possibility of him actually losing the election had every Canadian voted.

2 Data

Two datasets were used for analysis: the online 2019 Canadian Election Survey [**?**] and the 2019 General Social Survey [**?**].
The model used for this analysis will be a logistic regression with a binomial outcome for the outcome variable. Although multiple main parties had participated in the 2019 Canadian federal election: Liberal, Conservative, Bloc Quebecois, New Democratic and Green, only the two parties with the largest number of support will be considered for the model, leading to a binary outcome. This choice will ultimately lead to only including shown support for the Liberal and Conservative parties. The leader of the Liberal Party is Justin Trudeau and for the Conservative Party is Andrew Scheer.
The predictor variables selected for our multilinear model include: 1) Age 2) Gender 3) Level of Education Attained 4) Province 5) Ethnicity The outcome variable that is regressed over is whether the respondent shows support for the Liberal Party (Justin Trudeau) or the Conservative Party (Andrew Scheer) in the 2019 Canadian federal election. Support for the liberal party is assigned to a dummy variable to create a binary outcome. Intention to vote for the liberal party is denoted at 1 whereas support for the Conservative party is 0.

2.1 Survey Data

To build the MRP post-stratification model, this paper uses data collected from the Canadian Election Survey conducted in 2019 as the sampling data. The Campaign Period Survey was held from September 13th to October 21st, 2019. The sample was built from stratified sampling where the targets where stratified by region. The sample was then balances based on gender and age within each of the regions. Additionally, weights have been places in the gender, age group, and education level of the sample that was based on the 2016 Canadian Census[**?**].
The population of the data is the Canadian population who meet the age requirement to vote in the Canadian Federal Election, however the frame in the CES are the respondents from the market research platform, Qualtrics.
The Canadian Election Survey conducted in 2019 had categorized the question concerning vote choice in the election based on certain eligibility requirements. If individuals had responded that they were likely to vote in the election and they were eligible to vote since they were a citizen of Canada, there response for who they would most likely vote for in the election was recorded under the question "cps2019_votechoice". Other Canadian residents that did not express intention to participate in the 2019 were asked who they would most likely vote for if they did decide to vote. This response was recorded under the question "cps2019_vote_unlikely". Respondents who were either permanent residents or had some other status in Canada other than having citizenship had their choice of party in the election recorded in "cps_vote_lean" and "cps_vote_lean_pr". In this analysis, I will be comparing the results from just using the subset of individuals who expressed interest to vote in the election and were citizens with the subset containing all four groups mentioned. However, the CES had already filtered out individuals who were younger than the age of 18, so I was only able to analyze for individuals who exceeded the minimum age to vote in the Canadian Federal Election.

To match the variable levels with the post-stratification data, those who identified themselves as non-binary in response to the question on gender were exclude from the final dataset used. The GSS only had the variable sex, indicating biological sex with only two possible responses: male of female. After removing all individuals who did not respond as either male or female, the observation count reduced from 37,822 to 37,531. Although this difference in data is minimal, this may still have large effects on the proportions of the cells, thus affecting our predictions for each subgroup. As mentioned previously, all individuals who expressed support for parties other than the Conservative and Liberal party were excluded from the dataset in the interest of fitting a model to a binary outcome. In addition, all observations that had non-response for any of the predictor variables or the response variable were excluded from the final datasets used to successfully create a model in statistical software R [R Core Team, 2020].
After removal of non-responses, the full dataset of the survey had an observation count of 20,351 whereas the just voting dataset had an observation count of 18,831. To cut down on the run time for the Bayesian models for each of these datasets, a random sample was taken of 6,000 observations for both the full data and voting data. More accurate results may have been produced from running the full datasets of each.

Figure 1 shows the number of votes for each party by gender in the dataset for just voters. It appears that on average, more females had responded to the survey tend to support the Liberal party more so than men. This same trend is reflected in Figure 2, which shows the number of votes of each party by gender for the full survey.

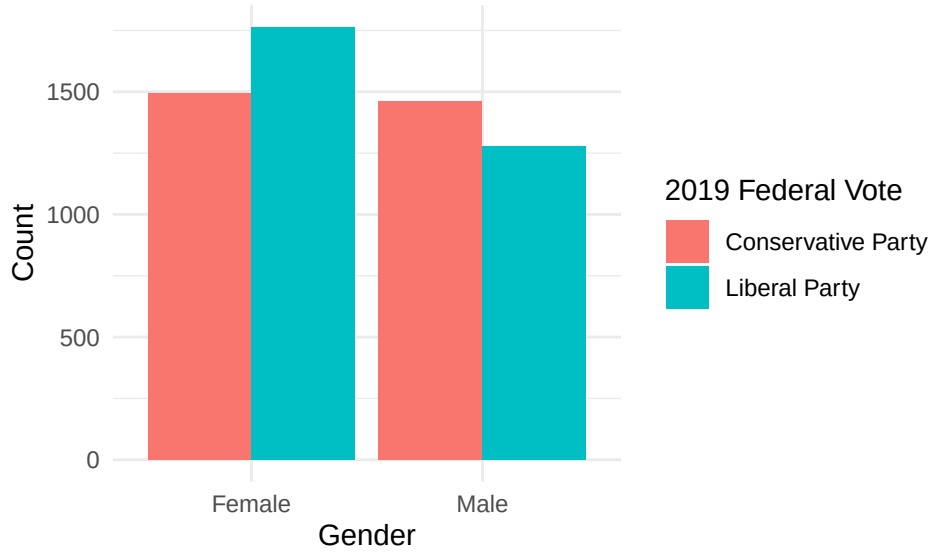

Figure 1: 2019 Presidential Votes by Gender for Just Voters

Figure 3 shows the number of votes for each party by province for just voters. Alberta is significantly more supportive of the Conservative party whereas Ontario and Quebec are more supportive of the the Liberal Party. These three provinces have the largest proportions in the survey, whereas provinces such as Nunavut and Northwest Territories have low representation in the survey. Notable, these provinces have low population counts.

Figure 4 shows the number of votes for each party by province for full survey. Ontario, Alberta, and Quebec have the highest proportions in each of the surveys. Alberta is notably more conservative than almost all other provinces.

Figure 5 shows the number of votes for each party by age group for just voters. Figure 6 repreents the number of votes for each party by age group for the full survey. Similar proportions are shwon for each age group across the two subsets. the 60 plus age groups seems to have the most representation in each of the surveys, whereas the 18 to 29 age group has the least representation.

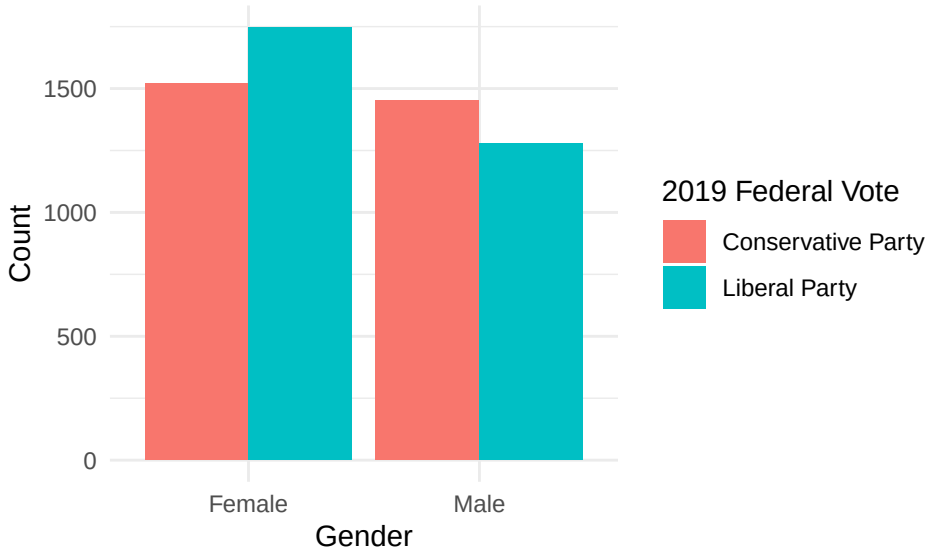

Figure 2: 2019 Presidential Votes by Gender

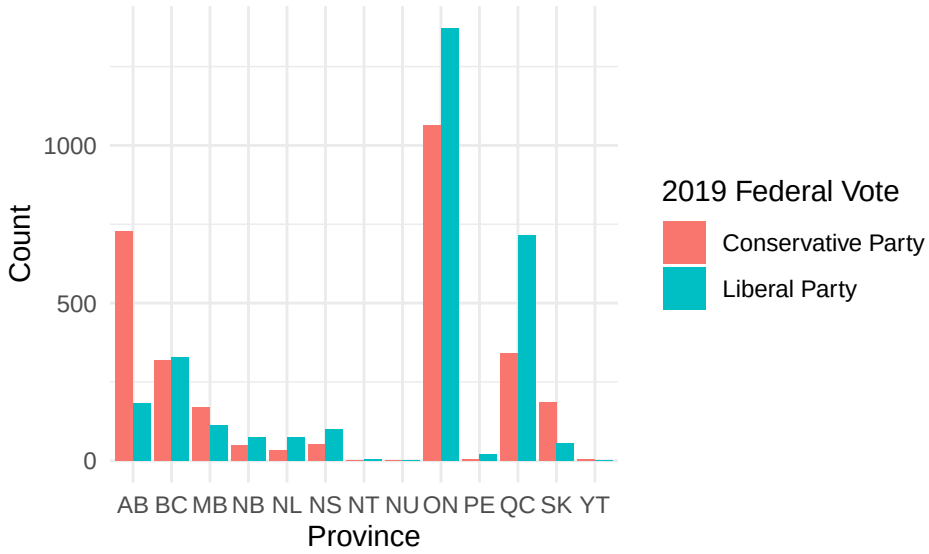

Figure 3: 2019 Presidential Votes by Province

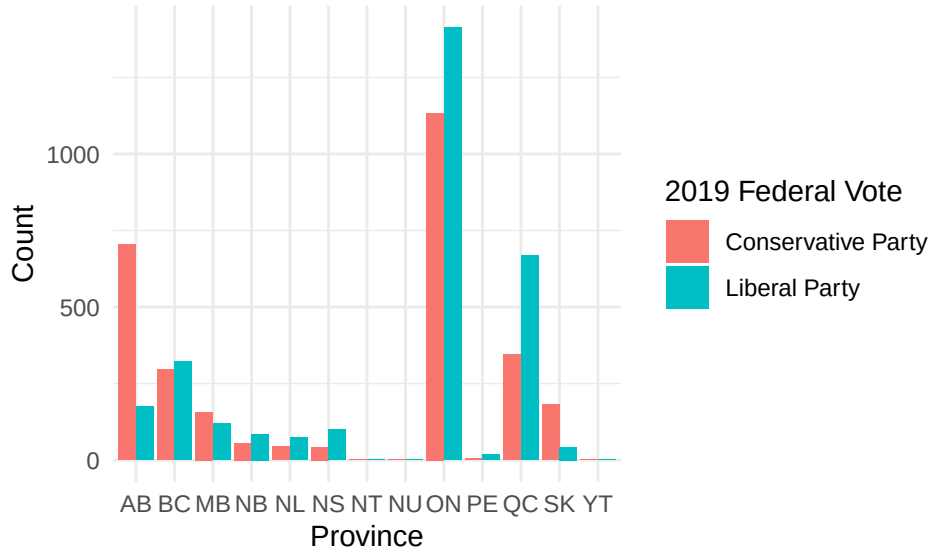

Figure 4: 2019 Presidential Votes by Province for Full Survey

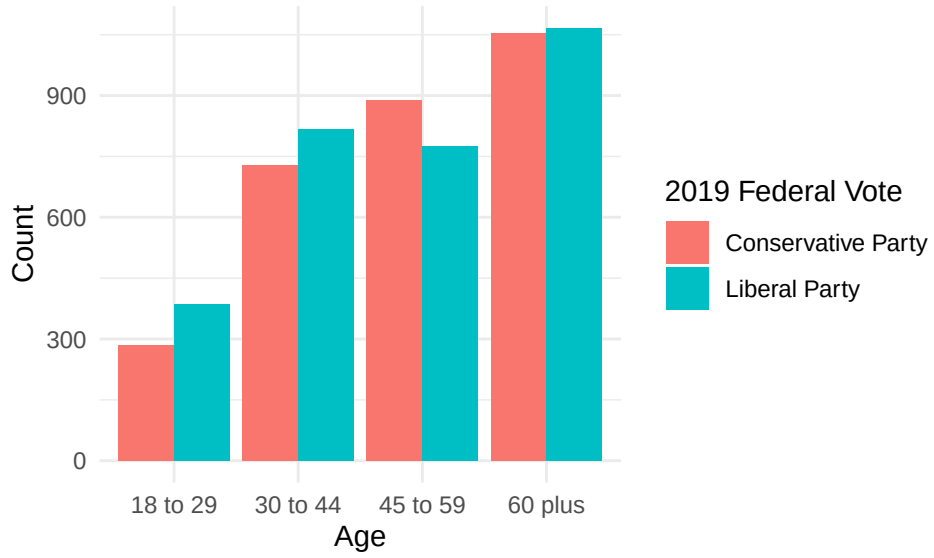

Figure 5: 2019 Presidential Votes By Age

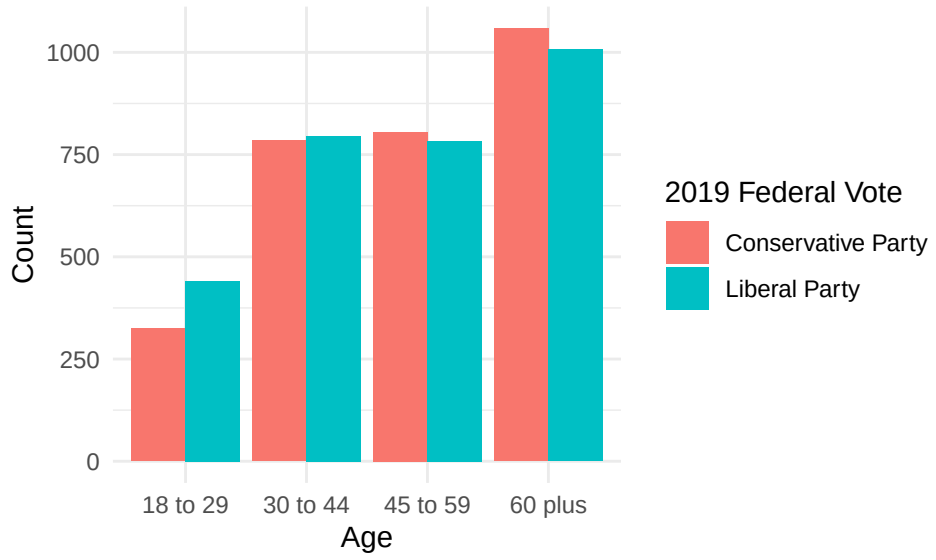

Figure 6: 2019 Presidential Votes By Age for Full Survey

Figure 7 and 8 show that the full survey and just vote survey had similar proportion breakdowns for each of the education levels. THose who graduated college had a large representation in each of the surveys, leading to possible bias for the multi-level post-stratification to correct

Figure 9 compares the proportions of each category from the full survey with the individuals who intend to vote. The proportion of age groups deviate slighlty from the full survey to the just voters whereas age and gender are very similar to each other. Figure 10 displays the the comparison of proportions for each province in the survey. As shown, there is a significatly greater proportion of respondents in the full survey rather than the just voter survey for Ontario. Ontario is a popular destination for immigrants arriving in Canada [**?**], therefore the proportion of respondents in Ontario that are immigrants nd not eligible to vote will most likely be higher.

Figure 10 compares the proportions in provinces for the full dataset and the just voting dataset. Ontario has a much higher representation in the vote dataset than the full dataset, indivating a lot of individuals who are voting are from Ontario. THere are additional deviations in other province proportions across the full survey with the just vote survey.

2.2 Post-Stratification Data

The General Social Survey collects information from individuals aged 15 and older. Since the minimum age to vote in the Canadian Federal Election is 18, all individuals not meeting this requirement were excluded from the dataset used for analysis. The target population for the GSS was the total population of Canadian residents in all ten provinces over the age of 15. Each year, around 20,000 individuals respond to the surveys. The GSS uses common telephone frames as the sampling frames provided by Statistics Canada through the Address Register, Census of Population and other administrative sources [**?**]. Interviews was conducted by telephone calls and self-completed online questionnaires. Using telephone calls for interviewing usually has lower response rates and will exclude younger, single-person households since many younger individuals do not have landlines and use cell-phones instead. Cell phone numbers are not recorded per household in the sampling frames, therefore these individuals will be excluded from the sampling pool. However, phone call interviews and self-completed online questionnaires are most cost and time effective.
Level of education attained and age groups had been adjusted to fit different subgroups in order to produce large enough cells for analysis in the Bayesian model.

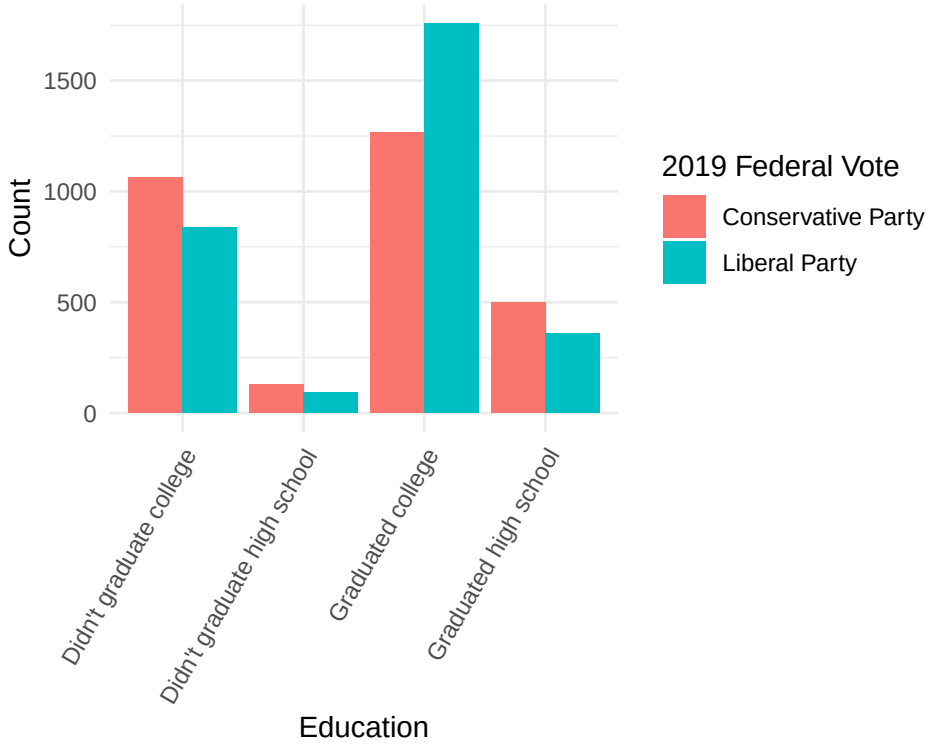

Figure 7: 2019 Presidential Voted by Education Attained

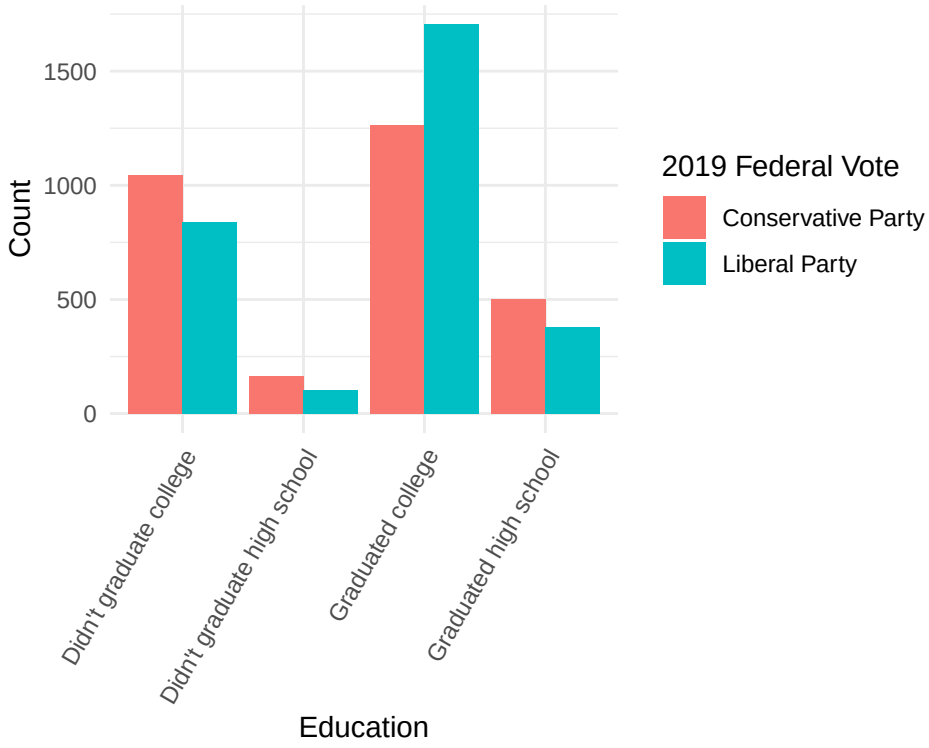

Figure 8: 2019 Presidential Voted by Education Attained for Full Survey

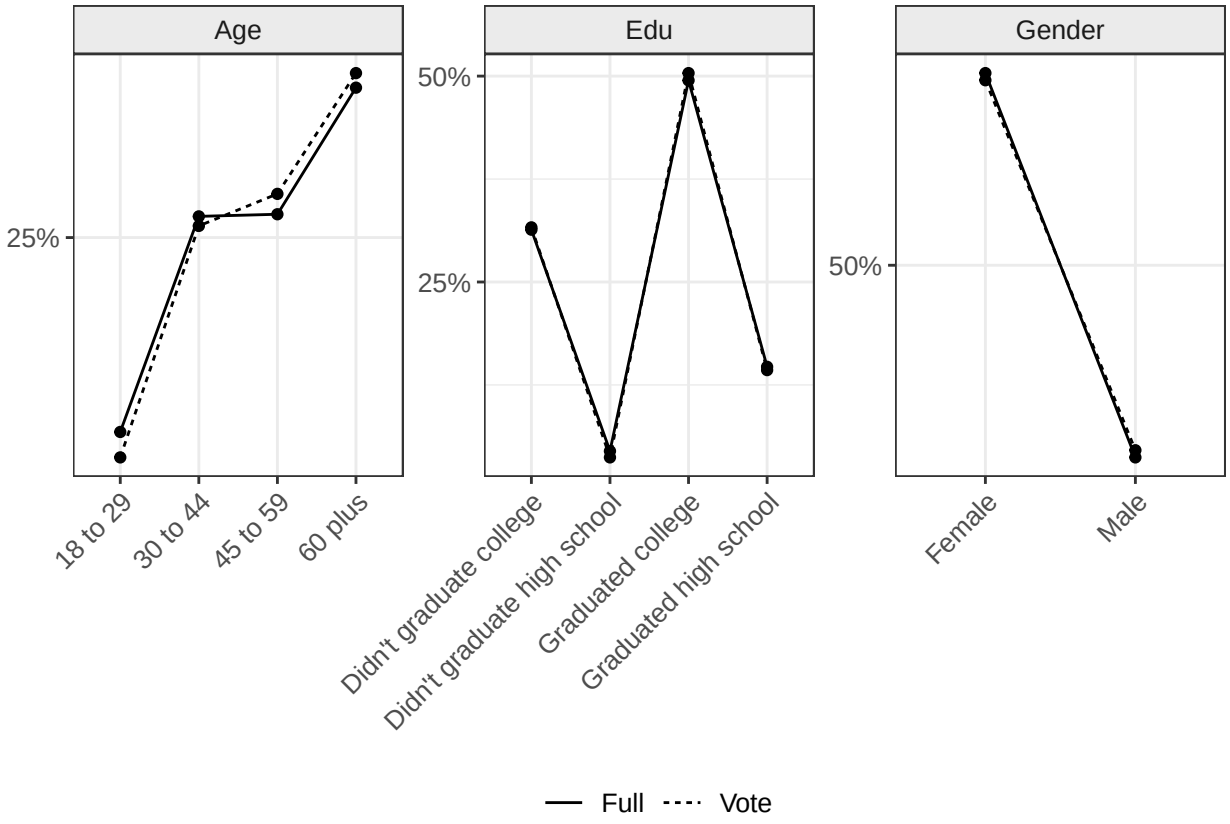

Figure 9: Demographic Proportions

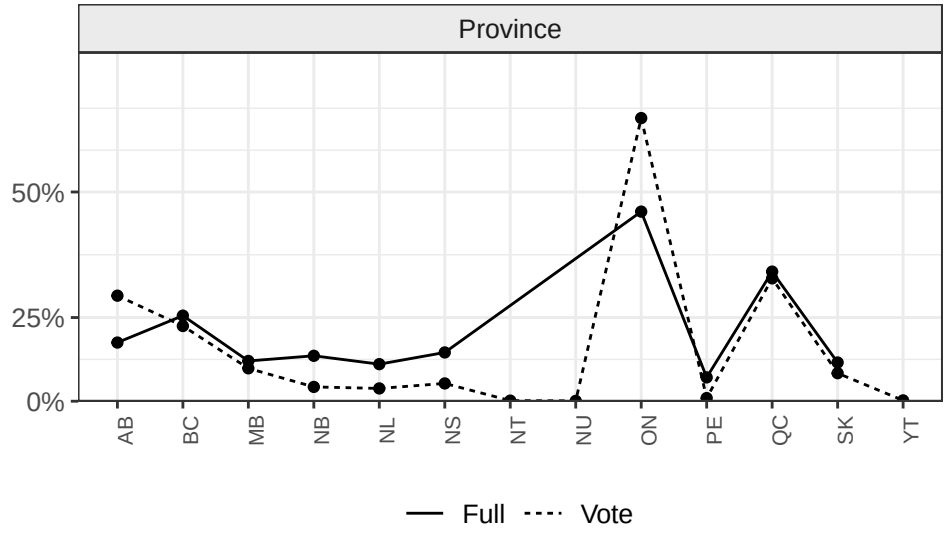

Figure 10: Proportions by Province

3 Model

The purpose of using MRP with post-stratification is to fix the misrepresentation of the different groups of respondents from the survey data. Although the CES had used strata to draw their sample, adjusted for gender for each province and reweighted by certain variables, less than 50% of the sample selected had responded to the survey. There is large possibility that the respondents who chose to respond may have similar features, therefore leading to some bias in the sample. This bias will ultimately affect the prediction of the model if we just use this sample alone. Therefore, a post-stratification dataset is used to mitigate this issue by using a more representative sample of the population from each group. This dataset does not have any information on who each individual would vote for, it is mainly used to find the proportions of each subgroup, or cells, that we are interested in from the combination of predictor variables. The results from the survey will then be reweighted by the predictor variables by post-stratifying the new dataset. Usually, census data is used for post-stratification datasets since it is the most representative of whole nation populations. This paper utilizes the General Social Survey as the post-stratification dataset. MRP is extremely useful for fixing biased samples, however can cause issues if the cells that are produced are too small for analysis. Having too small of observations in each cell will lead to poor predictions without enough statistically significant numbers to predict upon. The cleaning process of each of the survey data and the post-stratification data had been carefully fitted to avoid this issue of small cells. Only the respondents who expressed support for either the Liberal Party or Conservative Party remained in the sample data used. Therefore, the model is predicting based on a binary outcome. For this purpose, support for the Liberal Party is expressed as 1, while the Conservative Party as 0 in the binary model.
Instead of using a frequentist approach to build the model, a Bayesian model will be formed for making predictions. Considering that the GSS is acting as our post-stratification data, the estimation of the proportions of each subgroup within the population is only estimated. It is assumed when using a Bayesian approach that the actual distribution of the target population of each of the two groups being predicted on, all individuals in Canada over the age of 18 and all Canadians who are eligible to vote and intend to vote, follows a separate randomized distribution. This implies that the true proportions of the distribution are not obtainable by using data from the GSS, however we can use the GSS to find a distribution to estimate the true proportions. Therefore, this idea works into our model with the concept of Bayes Rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(A)}$$

A represents the prior distribution, which is assumed of the parameter that is being estimated of. In the instance of this paper, this is the distribution of the proportions of the Liberal party across each subgroup/cells created by the predictor variables selected: age, gender, education and province [Christian-Burkner, 2017] [Simpson, 2019].

When interpreting the results of the fitted Bayesian model of the sample data with the post-stratification data, the proportions of each of the subsets of the survey must be analyzed in addition to the final estimates from the Bayesian model. The proportions from the post-stratification must also be considered to understand the validity of the model.

4 Results

In this section, the estimates from the Bayesian model for both the full survey and the just vote subset are graphed with the 97.5% confidence intervals indicated by either red or blue bars.

Figure 11 and 12 shows the estimates of proportion of Liberal support from the full data and the vote data with the MRP estimate by age group. The estimates are quite similar to the age, considering that the age categories for both the GSS and CES were adjusted according to the Canadian census. The estimate for liberal support was higher in age group 18-29, and the MRP estimate was lower to adjust for the bias in age group participation in the survey.

Figues 13 and 14 are comparing the estimates for the provinces of each survey group. Nunavut, Northwest Territories and Yukon had no observations in the survey data, therefore there was not enough information to produce an MRP estimate. The population in these provinces are relatively low. The estimates for each are quite similar for each of the survey subsets.

Figures 15 and 16 are comparing the bayesian estimates for age. While females had reported higher support for the liberal party, the MRP estimate was slightly higher for each subset.

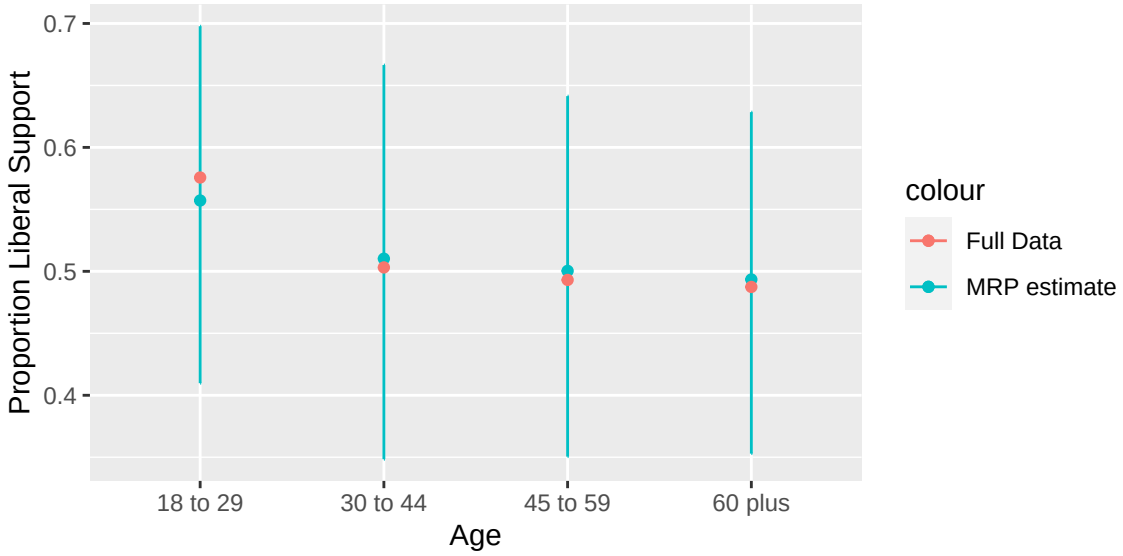

Figure 11: Comparing Full Survey Estimates with MRP Estimates by Age

Figures 17 and 18 look into the education groups. Many individuals who graduated college were more support of the liberal party, and this group had comprised most of the the individuls in the survey. This correlation is important to note in the final evalation.

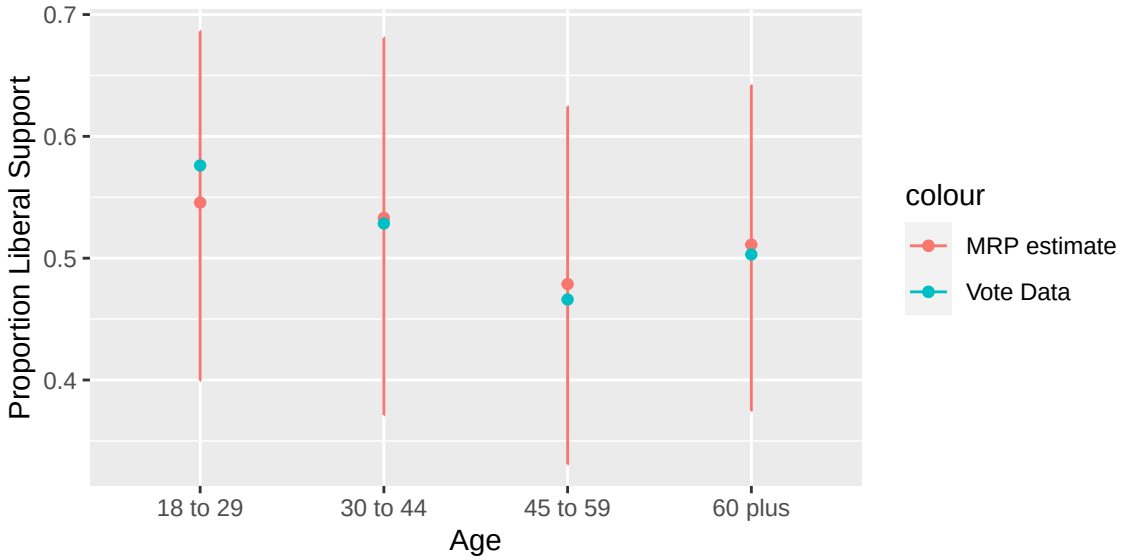

Figure 12: Comparing Just Vote Survey Estimates with MRP Estimates by Age

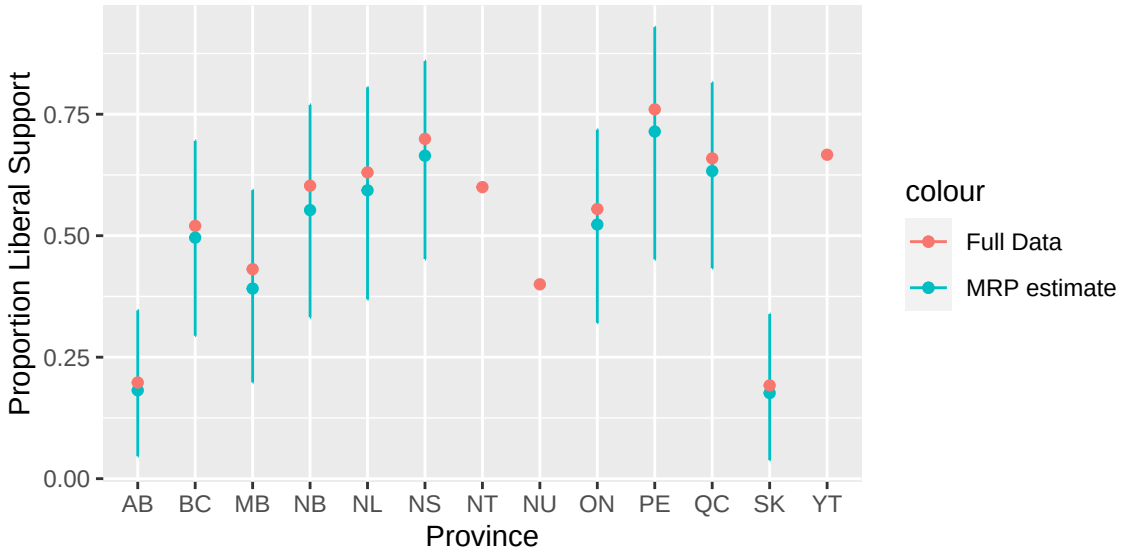

Figure 13: Comparing Just Vote Survey Estimates with MRP Estimates by Province

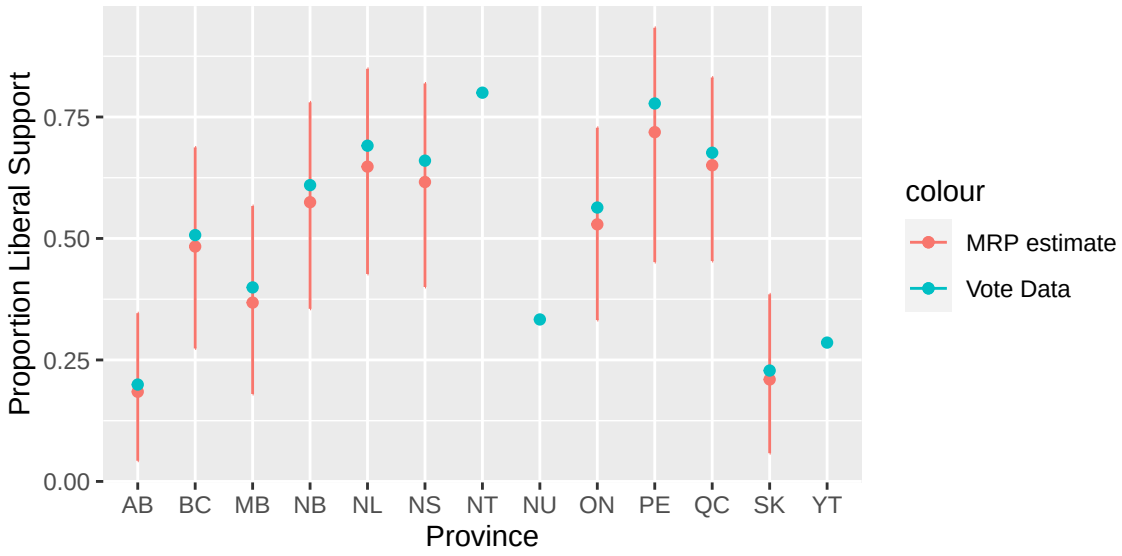

Figure 14: Comparing Just Vote Survey Estimates with MRP Estimates by Province

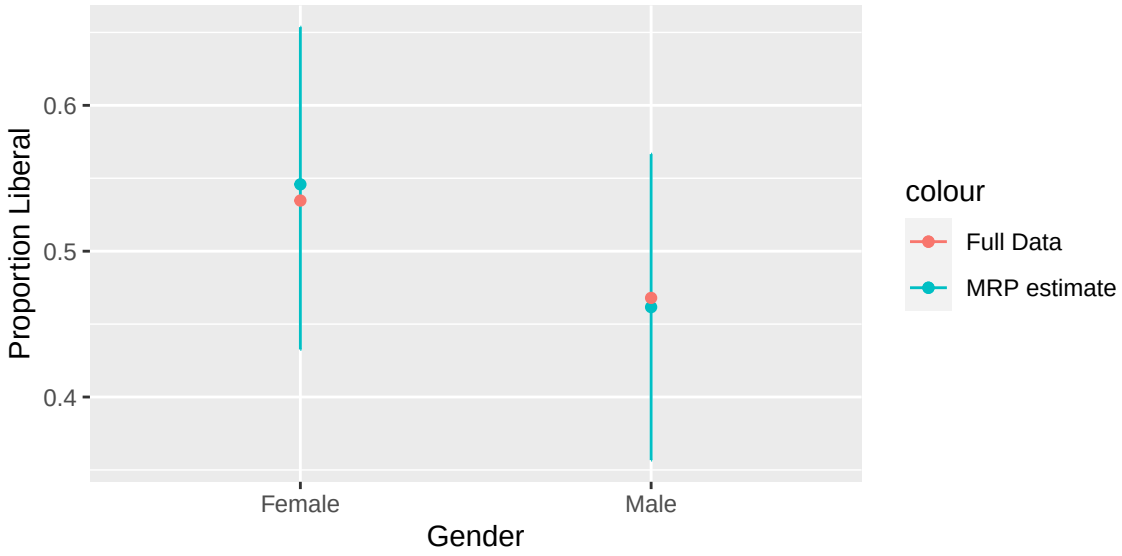

Figure 15: Comparing Survey Estimates with MRP Estimates by Gender

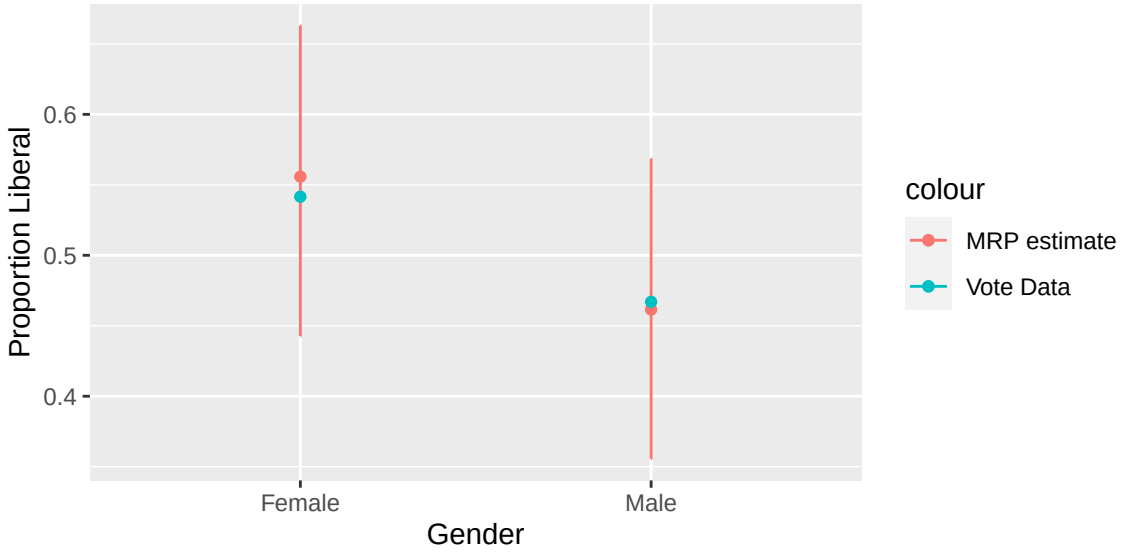

Figure 16: Comparing Survey Estimates with MRP Estimates by Gender

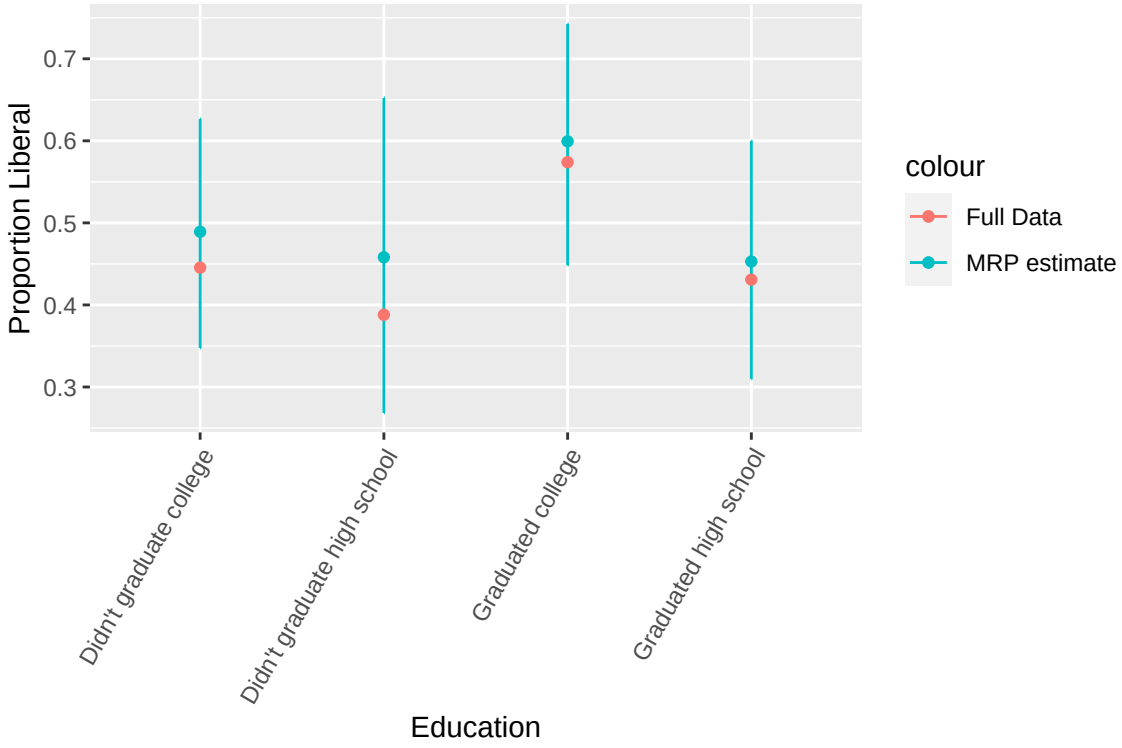

Figure 17: Comparing Survey Estimates with MRP Estimates by Education

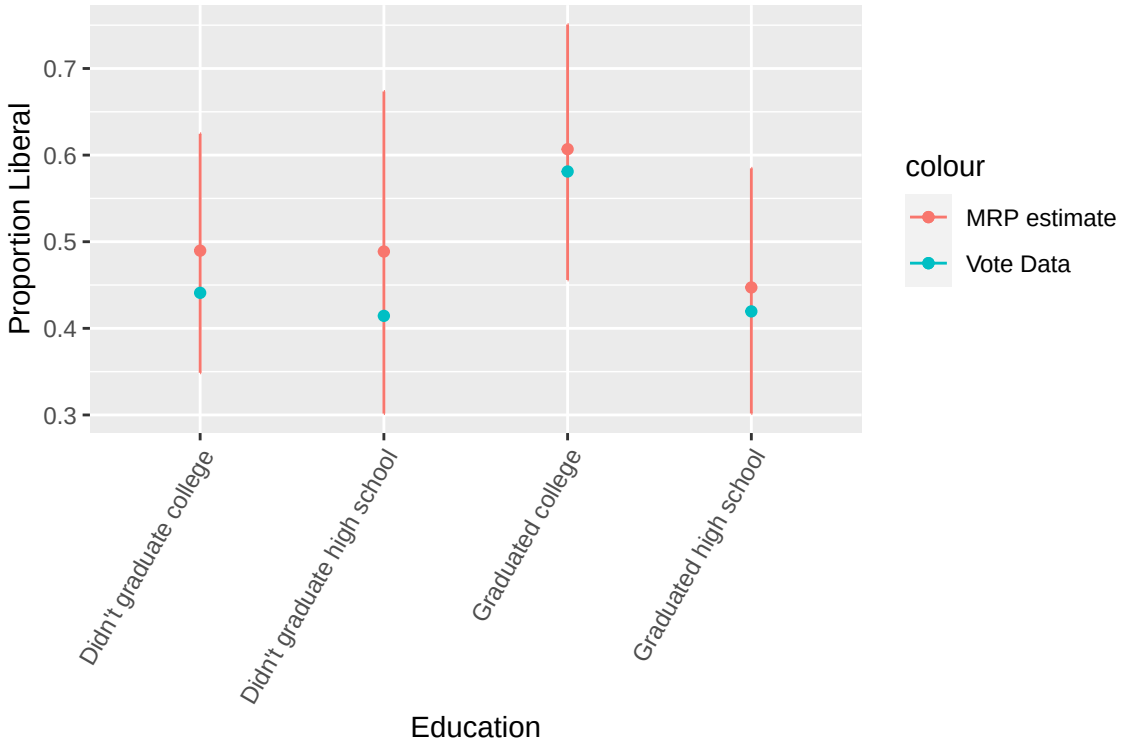

Figure 18: Comparing Survey Estimates with MRP Estimates by Education

5 Discussion

WHen conducting surveys and collecting information, it is very difficult to obtain a proper representation of the population of interest. In most cases, the population we are trying to reach is very large and not everyone you would like to obtain information on is willing or able to participate in a study. This reason causes some sort of bias. The way that studies are carried out can indicate ways that bias is formed. However, how can we exactly quantify how much bias there is when the true population is not obtainable? MRP with post-stratification is an excellent remedy to this situation since it is time and cost efficient. In order to correct bias in the Canadian Election Study, we use a more representative dataset of th Canadian population, the General Social Survey, to help create estimates for support for the Liberal Party in the 2019 Canadian Federal Election.

Our estimation found the estimates for both the full survey data and the to be very similar at 51.21% and 50.67% respectively. Although these estimates were very close, further investigation should be conducted on the non-voting parties in certain provinces and how the influence of their vote can change the result of the election. The popular vote was calculated in this instance, however, the liberal party had lost the populat vote in the 2019 Federal Election. More analysis and information should be conducted on the remaining 23% of the population to understand their possible influence in further elections should they choose to participate. Pollers every year usually just examine data from past voter preferences, but it is important to consider the trends in individuals who are being welcomed into the population, such as individuals turning the age of 18, individuals deciding they would like to vote, or Canadian residents receiving their citizenship.

MRP can be used as a tool when a population of interest is identified with key variables that can affect the variable of interest, which can be matched to a sample population to formulate inferences. Comparing the sample estimate of Liberal support with the forecasted estimates on the GSS datasetprovides further tools for research to explore nuances in the data and additional differences. MRP ultimately can identify certain variables of interest as predictions to perform smaller scaled polling to include more levels of variables of the data that were hidden in the way we categorized individuals within each of the groups, such as race and education in our case. It is relatively quick, easy and cost effective to perform on data given a survey with matching key variables with a more representative population.

MRP has lead many studies to significant and representative models on actual populations from samples, however it does not come without drawbacks. Insufficient estimates can be caused by lack of some demographic predictors, insufficient data from surveying as well as lack of regularization. In some cases, selecting certain variables as predictors lead to extremely small cell counts for certain groups, therefore producing unstable estimations. Through our cleaning process of both datasets, some tweaking of the variables and levels had to be made in order to match all variables to perform MRP. The sample data was extremely extensive and all-inclusive of any possible variable, however this is not always the case. The way that information demographic variables and in the sample and census data are sometimes collected and recorded in different forms. [Kennedy, 2020] [Gelman, 2020].

6 Weaknesses and Future Work

Using the 2016 Canadian Census data was challenging since the cells produced were very small since there were a multitude of different categories for certain variables. To remedy the small cell sizes, the General Social Survey had provided estimates for each group in the post-stratification step. However, the General Social survey collect information from individuals over the age of 15 and uses stratified sampling using telephone numbers and addresses as the frames. Depending on how the sampling was conducted and the response rates for different individuals across regions, the GSS is subject to influence of not accurately representing the Canadian population as oppose to the actual census data.
All non-responsive instances for any variable in each dataset was omitted for the purpose of running the models in statistical software R [R Core Team, 2020]. In addition, the Canadian Election Study had allowed for a third response for gender, which was non-binary. However, the GSS had only allowed two possibilities for the variable sex. Therefore, any individual who associated themselves as non-binary could not express

their gender in the GSS. Therefore, all CES respondents who had indicated they were belonging to the non-binary gender category were not included in the dataset used in order to keep consistent with the post-stratification data. The exclusion of the responses from these individuals could have affected the weights of each of the subgroups, ultimately affecting our prediction. For maintaining a binary outcome for the models, all individuals who were unsure of their vote choice or supported a different party than the Liberal or Conservative party were excluded from the data. This had cut down the observation count for the full dataset from 36,946 to 20,361. This process may have affected the proportions of the voting groups included. Perhaps, those who were not planning to vote in the election were unsure of who they would vote for if they could participate in the election since they have not given it much thought. The dynamic of individuals who had no preference in the election should be further explored in future work. After removal of all the non-responses and individuals in the non-binary gender category, the full survey data had around 20,361 observations whereas the just voting survey had 18,841 observations. After many hours in attempt to run Bayesian models on each of these datasets, I had decided to take a random sample of 6,000 observations of each of these datasets to create each of the Bayesian models. Although using the full amount of observations available would be optimal in the predictions and modelling, in the interest of my patience and sanity, the results of the prediction may not accurately represent the data as much as possible.

References

JJ Allaire, Jeffrey Horner, Yihui Xie, Vicent Marti, and Natacha Porte. *markdown: Render Markdown with the C Library 'Sundown'*, 2019. URL https://CRAN.R-project.org/package=markdown. R package version 1.1.

Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. URL https://CRAN.R-project.org/package=gridExtra. R package version 2.3.

Stefan Milton Bache and Hadley Wickham. *magrittr: A Forward-Pipe Operator for R*, 2014. URL https://CRAN.R-project.org/package=magrittr. R package version 1.5.

Paul-Christian Bürkner. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28, 2017. doi: 10.18637/jss.v080.i01.

Paul Christian-Burkner. *Advanced Bayesian MultiLevel Modeling with the R package brms*, 2017. URL https://arxiv.org/pdf/1705.11123.pdf.

Paolo Di Lorenzo. *usmap: US Maps Including Alaska and Hawaii*, 2020. URL https://CRAN.R-project.org/package=usmap. R package version 0.5.1.

Andrew Gelman. *Know your population and know your model: Using model-based regression and post-stratification to generalize findings beyond the observed sample*, 2020. URL https://arxiv.org/pdf/1906.11323.pdf.

Matthew Kay. *tidybayes: Tidy Data and Geoms for Bayesian Models*, 2020. URL http://mjskay.github.io/tidybayes/. R package version 2.1.1.

Lauren Kennedy. *Know your population and know your model: Using model-based regression and post-stratification to generalize findings beyond the observed sample*, 2020. URL https://arxiv.org/pdf/1906.11323.pdf.

Kirill Müller. *here: A Simpler Way to Find Your Files*, 2017. URL https://CRAN.R-project.org/package=here. R package version 0.1.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

David Robinson, Alex Hayes, and Simon Couch. *broom: Convert Statistical Objects into Tidy Tibbles*, 2020. URL https://CRAN.R-project.org/package=broom. R package version 0.7.1.

Dan Simpson. *Multilevel (structured) regression and post-stratification*, 2019. URL https://statmodeling.stat.columbia.edu/2019/08/22/multilevel-structured-regression-and-post-stratification/.

Stan Development Team. RStan: the R interface to Stan, 2020. URL http://mc-stan.org/. R package version 2.21.2.

Hadley Wickham and Evan Miller. *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*, 2020. URL https://CRAN.R-project.org/package=haven. R package version 2.3.1.

Hadley Wickham and Dana Seidel. *scales: Scale Functions for Visualization*, 2020. URL https://CRAN.R-project.org/package=scales. R package version 1.1.1.