Analysis on The Popular Vote Share Prediction of The Liberal Party of Canada (LPC) in The 2025 Federal Election

STA304 - Assignment 2

GROUP NUMBER: 110, by Christopher Jung, Yiyu Li, Haotong Wang & Patrick Zhou

November 29, 2022

Introduction

The popular vote phase in a Canadian federal election is a nationwide democratic practice for Canadian citizen to express their expectation of the future governing party of Canada. One of the major federal political parties, the Liberal Party of Canada (LPC), has won the last 3 consecutive federal elections from 2015-2021. On the contrary to its crushing seat rate[1], the party's leader and Canada's current Prime Minister (PM), Justin Trudeau, has set the record of "slimmest share of overall electoral support in Canadian history" twice with a popular vote share[2] of 33.12% in 2019 and an even lower share of 32.19% in 2021 (Hopper). While popular vote share is not a determining factor in winning the election of PM, it is decisive in the election of members in The House of Commons[3] and a way for the citizens to express their opinions on a variety of public issues. With the importance of the popular vote addressed, **this analysis aims to identify the major factors affecting the popular vote share of LPC using data from election surveys and polls, as well as to predict their popular vote share in the upcoming (2025) Canadian federal election.** Such predictions allow political parties to make more efficient allocation of resources for campaigning.

We made a few hypotheses regarding the selection of the factors mentioned above. This analysis focus on basic demographic factors, such as sex[4], age and province of residence. Most of those factors have supporting evidence of their influence on the LPC popular vote share, such as the popular vote ratio between the two most competitive parties, LPC and CPC (Conservative Party of Canada). In the 2021 election, this LPC to CPC ratio in most provinces was approximately 1:1 while in Alberta and Saskatchewan widened to around 1:5 (Elections Canada, "Election Night Results - Provinces & Territories"). In a web survey by Leger, this ratio is rather even (34% LPC to 35% CPC) among male voters but became 43% to 23% for females. The survey also suggests that younger voters (18-34 years old) compose a substantially smaller fraction of voting intentions in general, which aligns with the voter turnout rate by age (Elections Canada, "Voter Turnout by Sex and Age"). These pieces of evidence imply that sex, age and province of residence are all major determinants of LPC popular vote shares. **We hypothesize a similar split of voting intentions in terms of sex, age and province for the upcoming election, that is, males, younger citizens and residents of Alberta and Saskatchewan are less likely to vote for LPC.**

[1]The percentage of seats in The House of Commons voted for a specific political party. There are 338 seats in total, each held by members elected by citizens who vote in general elections or by elections. The seat rate decides the winner of a federal election (Elections Canada, "The Electoral System of Canada").

[2]The vote share of a specific political party that is cast by citizens in a federal election. Popular vote result does not determine the winner of a federal election (Elections Canada, "The Electoral System of Canada").

[3]The lower house of the Parliament of Canada. Together with the Crown and the Senate of Canada, they comprise the bicameral legislature of Canada (Elections Canada, "The Electoral System of Canada").

[4]Sex here refers to biological sex.

Data

This analysis involves model fitting using two datasets: the General Social Survey data (GSS) and the Canadian Election Study data (CES). Both contain comprehensive socio-demographic data while the latter has a focus on Canadian Election related information.

The version of CES used in this analysis is collected through telephone interviews, on behalf of the Canadian Election Study during the 2019 Canadian federal election (Stephenson et al.). This study used a modified random digit dialling procedure (RDD) that randomizes the selected phone numbers to reduce the chance of collecting biased samples. GSS comes from Canada's General Social Survey program. The program's data-collecting process also involves RDD, but with a greater focus on online questionnaires to supplement the existing telephone mode of collection (Statistics Canada).

We will refer to GSS as the census data and CES as the survey data for the rest of this report.

Data Cleaning

Since the election only concerns eligible voters, we filtered out the respondents who are underaged (< 18) or non-citizens. We also added a province abbreviation variable that makes the province name easier to read. The above procedures can be applied to the census data directly, while the survey data requires translation from number labels to plain English beforehand. We then selected the variables of our interest to keep and dropped the rest. For census data, we kept the respondent's age, sex, abbreviated province of residence, education level and household income. Note that our hypothesis does not include the respondent's education level and household income, but they are commonsensical recognized to have some impact on a political party's popular vote share. For survey data, we kept the same variables as with the census, with an extra binary indicator of whether the respondent voted for LPC. This variable is obtained by mutating the voting choice variable to be "Yes" if the voting choice is LPC and "No" otherwise. We then filtered out the respondents whose responses for any of the above variables are missing. Before the data mapping, the number of observations after cleaning for census data reduced to 18750 compared to the original 20602, whereas for survey data it is 3015 compared to the original 4021.

Data Mapping

For post-stratification purposes, we mapped the census data with the survey data by rescaling their variables. For respondent age, we split the numeric age in both datasets into 6 groups: '18-24', '25-34', '35-44', '45-54', '55-64' and '65+' years old. To match the sex variable from the census and the gender variable from the survey, we simply drop the only observation in the survey data that has the gender "Other", and the remaining categories ("Females" and "Males") align with the census sex labels. For the province of residence, both datasets contain respondents scattered across the 10 provinces and no respondents from the 3 territories. Thus this variable does not require rescaling. For household income, we applied the income level in the census to the survey, splitting the survey's numeric household income into 6 groups: 'Less than $25,000', '$25,000 to $49,999', '$75,000 to $99,999', '$100,000 to $ 124,999', '$50,000 to $74,999' and '$125,000 and more. The education levels are categorized differently in both datasets before data mapping, but both can be sorted into 3 levels: '$<$ BA', 'BA' and '$>$ BA', where BA refers to a degree of Bachelor level. After the data mapping, the number of observations after cleaning for census data is the same as before (18750), whereas for survey data it is 1 observation less than before (3014 compared to 3015).

Description of important variables

There are 5 common variables in both the census and the survey data: respondent's sex, age group, province of residence, education level and annual household income level. The survey data contains 1 unique binary variable that indicates whether the respondent voted, or was willing to vote for LPC in the 2019 election. Below are brief descriptions of the 5 common variables mentioned above and their respective summary measures:

Province A categorical variable with 10 categories: QC (Quebec), BC (British Columbia), ON (Ontario), AB (Alberta), MB (Manitoba), SK (Saskatchewan), NL (Newfoundland and Labrador), PE (Prince Edward Island), NS (Nova Scotia) and NB (New Brunswick). This variable represents the respondent's province of residence in Canada. This is the variable that we will be post-stratified with in later sections.

Age group A categorical variable with 6 levels: 18 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64 and 65+. This variable represents the range that the respondent's age (in years) falls into.

Table 1: The number of respondents in each age group

	18 to 24	25 to 34	35 to 44	45 to 54	55 to 64	65+
survey_data	140	460	567	582	605	660
census_data	882	2359	2864	2836	3970	5839

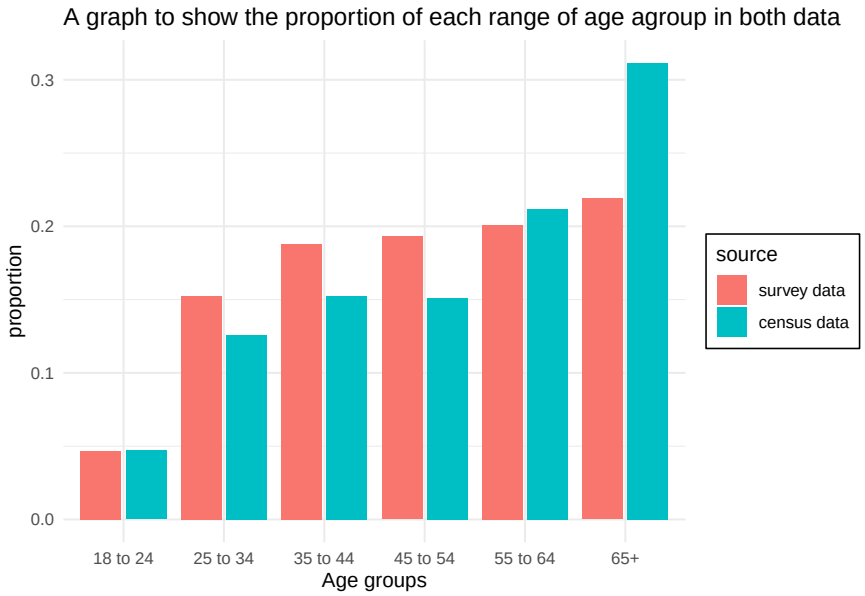

Figure 1: The distribution of respondents' age group, shown as proportions within each dataset

The above table and figure together show the discrepancy in respondent's age between census and survey data. The age of census respondents is heavily skewed to the elder generation (65+) while that of survey respondents distributes evenly across 25 – 65+ years.

Education A categorical variable with 3 levels: less than Bachelor's degree (<BA), at Bachelor's degree (BA) and more than Bachelor's degree (>BA). This variable represents the respondent's education level.

Table 2: The number of respondents in each education level

	<BA	>BA	BA
survey_data	1667	497	850
census_data	13654	1653	3443

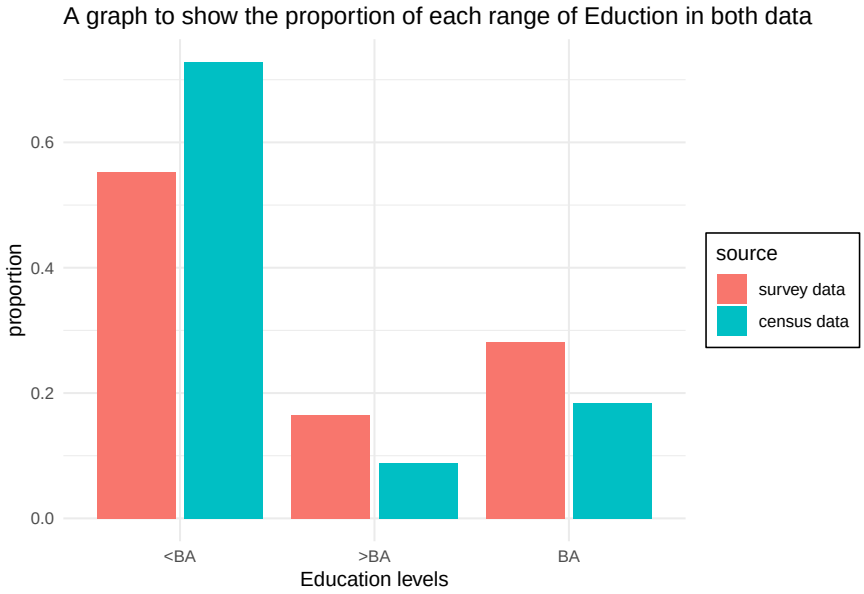

Figure 2: The distribution of respondents' education level, shown as proportions within each dataset

The ratio of proportions between each level is relatively similar within each dataset. The discrepancy of proportions, however, still exists across datasets: more than 70% of survey respondents have less than bachelor's degree, while for census this proportion becomes less than 60%. It is evident that the majority of respondents in both datasets have lower-than-Bachelor's education experience.

Household income level A categorical variable with 6 levels: Less than $25,000, $25,000 to $49,999, $50,000 to $74,999, $75,000 to $99,999, $100,000 to $124,999, $125,000 and more. This variable represents the respondent's household income level.

Table 3: The number of respondents in each household income level

	Less than $25,000	$25,000 to $49,999	$50,000 to $74,999	$75,000 to $99,999	$100,000 to $ 124,999	$125,000 and more
survey_data	345	449	536	403	390	891
census_data	2444	3954	3370	2676	2009	4297

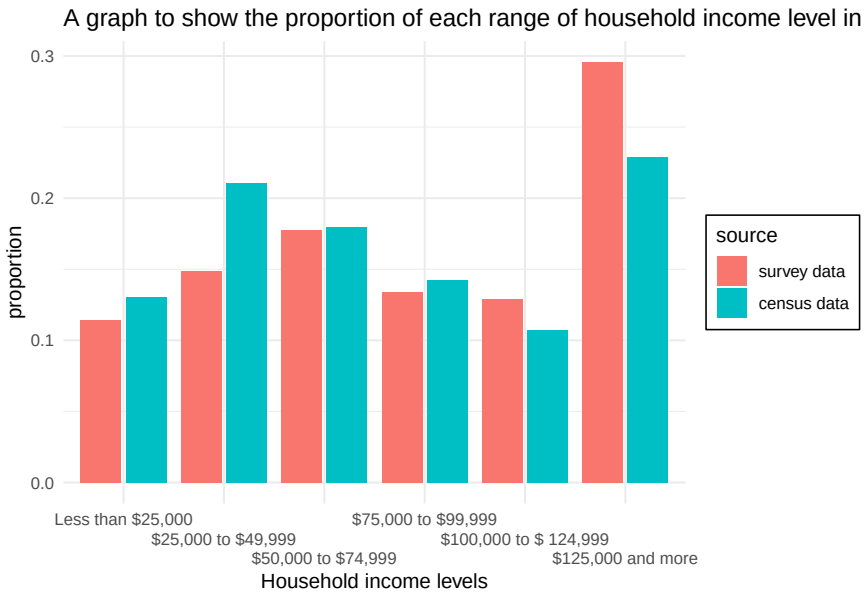

Figure 3: The distribution of respondents' household income level, shown as proportions within each dataset

The table and figure above suggest that the distribution of proportions in each household income level is relatively similar between the census and survey data.

Sex A categorical variable with 2 levels: Female and Male. Unlike with the census data, this variable is not indicative of the respondent's biological sex in the survey data. The sex variables in census and survey data are treated as the same variable for data mapping purposes.

Table 4: The number of respondents in each sex category

	Female	Male
survey_data	1266	1748
census_data	10248	8502

The figure indicates an approximate 4:6 ratio of female and male respondents in survey data, while this ratio is approximately the opposite (6:4) in census data. Along with all the discrepancies mentioned above, its potential influence on the result will be further addressed in the Result section.

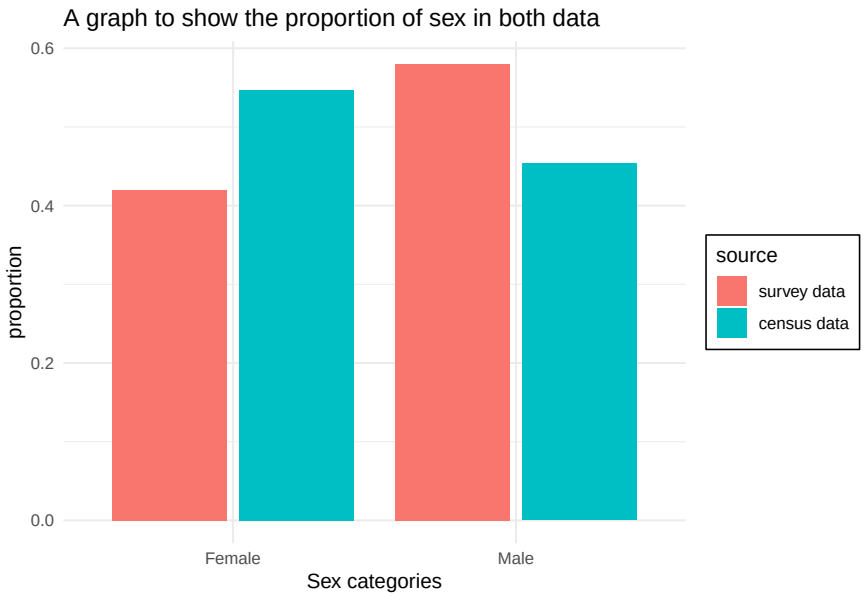

Figure 4: The distribution of respondents' sex, shown as proportions within each dataset

Methods

We plan to use the strategy of poststratification and multilevel regression, which has been popularly used to obtain accurate small-area subgroup estimates for voters' turnout based on their province.

Model Selection

Since we want to predict the vote share for LPC by a few predictors and the vote share is a binary variable in the data, we choose to use logistic regression. Because the linear regression provide the continuous output and the logistic regression provide the discrete output. So we choose the generalized linear model (GLM), which allows linear regression to relate the response variables through a link function so that the sum of the variance in each measure is related to its predictor. the price. It provides other statistical models such as linear regression, logistic regression and Poisson regression.

There are many criteria to select a model and many things that need to consider. And there can be many incorrect models, while there are many correct model that can be choose. To choose a model, the Akaike Information Criterion (AIC) can be a good choice. AIC is a mathematical method that calculated from the number of predictors and the maximum likelihood estimation (MLE) of the model to evaluate how well a model fits the data it was generated from.

We want to predict the vote share of LPC by the following pool of predictors: province_abbr, age_group, sex, income_family, education. We applied the AIC test on models built with all possible combinations of the above predictors.

A lower AIC score is better, and AIC penalizes models that use too many parameters. Therefore, if two models explain a significant amount of variance, the non-significant model has a lower AIC score and is the better model. In this case, we would like to choose the `vote ~ province_abbr + age_group + sex+education` model, which has the lowest AIC.

Model Specifics

Applying MRP in our setting comprises two steps. First of all, we fit a frequentist model. This method calculates the probability that the experiment would have the same outcomes if you were to replicate the same conditions again. And it only uses data from the current experiment when evaluating outcomes. In a

frequentist model, probability is the limit of the relative frequency of an event after many trials. Moreover, for frequentist approaches are more computationally efficient, which is much less intensive and available in a wider range of software packages that are good to use.

$$Pr(y_i = 1) = logit^{-1}(\alpha_0 + \alpha_{p[i]}^{province} + \alpha_{e[i]}^{education} + \alpha_{a[i]}^{age} + \alpha_{s[i]}^{sex})$$

Where y_i determines whether or not a voter votes for LPC, when y_i is 1 that means a voter vote for LPC.

α_0 represents the fix slope for LPC's fraction of vote share in the 2019 election.

$\alpha_{p[i]}^{province}$, for p = 1,. . . ,10

$\alpha_{e[i]}^{education}$, for e = 1,. . . ,3

$\alpha_{a[i]}^{age}$, for a = 1, . . . , 6

$\alpha_{s[i]}^{sex}$, for s = 1,2

Post-Stratification

We plan to use the strategy of poststratification and multilevel regression, which has been popularly used to obtain accurate small-area subgroup estimates for voters turnout base on their individual province. And poststratification is often used when a simple random sample does not reflect the distribution of some known variable in the population. Poststratification is a common method for correcting for the differences between survey and census populations (Little, 1993). The poststratification estimate can be define by

$$\hat{y}^{ps} = \frac{\sum_{j=1}^{J} N_i \hat{y}_j}{\sum_{j=1}^{J} N_i}$$

where $\hat{y}_j$ is the estimate of y in cell j, and N_i is the size of the j^{th} cell in the overall population. Post-stratification estimation is a technique used in statistical models to improve forecasting performance. The survey weight test is adjusted to force the unit predictions in each group of estimates to be balanced among known population totals.

Post-stratification Table

Table 5: Post-stratification Table

age_group	sex	province_abbr	education	income_family	Num_of_Cell	prob_cell
18 to 24	Female	AB	<BA	$100,000 to $ 124,999	3	0.0001600
18 to 24	Female	AB	<BA	$125,000 and more	14	0.0007467
18 to 24	Female	AB	<BA	$25,000 to $49,999	7	0.0003733
18 to 24	Female	AB	<BA	$50,000 to $74,999	3	0.0001600
18 to 24	Female	AB	<BA	$75,000 to $99,999	3	0.0001600
18 to 24	Female	AB	<BA	Less than $25,000	8	0.0004267

We end up with 1700 post-stratification cell, but based on the levels 10 provinces, 6 age categories, 3 education categories, and 2 genders would have expected $10 * 6 * 3 * 2$ cells. This difference is common in most post-stratification analysis.

All analysis for this report was programmed using `R version 4.0.2`.

Assumptions

Firstly, we check the assumptions for logistic regression.

- Outcome is binary. We let the outcome as vote or not for LPC, so the outcome is binary.
- Linearity in the log odds for numeric predictor variables. The numerical for the logistic model is We can use the Box-Tidwell test to check the linearity assumptions.
- Lack of strongly influential outliers Since influential values are extreme individual data points, which can be examined by visualizing the Cook's distance values, to alter the quality of the logistic regression model.

Here we can label the top 3 largest values:1705, 1725, 840

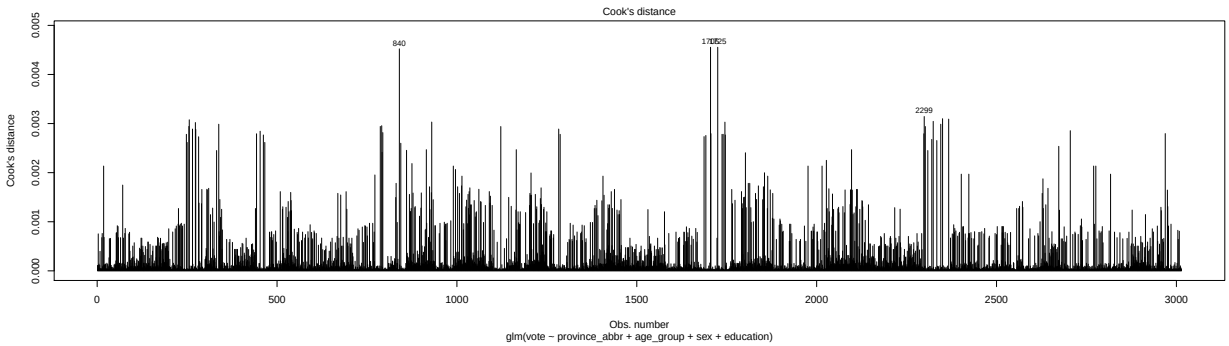

Figure 5: Cook's Distance concerning observations from the cleaned survey data

(Moreover, not all outliers are influential observations. To check whether the data contains potential influential observations, the standardized residual error can be inspected. Data points with an absolute standardized residuals above 3 represent possible outliers and may deserve closer attention.)

-Absence of multicollinearity By the definition of multicollinearity, it corresponds to a situation where the data contain highly correlated predictor variables.In this case our model has 4 predictor. To compute the variance inflation factors:

Table 6: VIF of the model fitted with the cleaned survey data

	GVIF	Df	GVIF^(1/(2*Df))
province_abbr	1.051980	9	1.002819
age_group	1.072309	5	1.007006
sex	1.009074	1	1.004527
education	1.053090	2	1.013016

As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity. And by the VIF value, we could see there is no predictor exceeds 10.

Results

We can see some interesting results in the overall Generalized Linear Model.

Table 7: Summary of the fitted model

term	estimate	std.error	statistic	p.value
(Intercept)	-2.2819685	0.3129290	-7.2922885	0.0000000
province_abbrBC	0.7023567	0.2496251	2.8136465	0.0048983
province_abbrMB	0.8168381	0.2850656	2.8654391	0.0041643

term	estimate	std.error	statistic	p.value
province_abbrNB	0.8834837	0.2974877	2.9698160	0.0029798
province_abbrNL	1.1752265	0.2936170	4.0025836	0.0000627
province_abbrNS	1.1020282	0.2903301	3.7957759	0.0001472
province_abbrON	1.3228268	0.2441407	5.4182967	0.0000001
province_abbrPE	1.1464027	0.2901436	3.9511564	0.0000778
province_abbrQC	0.8844324	0.2468018	3.5835737	0.0003389
province_abbrSK	-0.0355505	0.3209472	-0.1107674	0.9118008
age_group25 to 34	-0.1135585	0.2415846	-0.4700570	0.6383143
age_group35 to 44	-0.0782944	0.2359872	-0.3317739	0.7400600
age_group45 to 54	-0.2331237	0.2358177	-0.9885761	0.3228706
age_group55 to 64	0.0849707	0.2318011	0.3665672	0.7139419
age_group65+	0.2359575	0.2294103	1.0285394	0.3036962
sexMale	-0.0208950	0.0882718	-0.2367115	0.8128806
education>BA	0.7078315	0.1159388	6.1052167	0.0000000
educationBA	0.4471147	0.1016955	4.3966044	0.0000110

Sex is not a statistically significant predictor of liberal voting, there are not strong enough differences between males and females when controlling for the other variables. We could remove that variable from the model and check if the AIC improves. Province and Education appear to be significant at various levels, and we will look at the estimated effects in more detail below. Age does not appear to be significant at any level, compared to the baseline level (less than 25). Since this is a categorical variable we can check the significance of the variable overall with a χ^2 test.

Table 8: Chi-square tests of the fitted model

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
3001	3215.116	NA	NA	NA
2996	3200.736	5	14.38007	0.0133671

With a p-value of 0.01 we can conclude that age group is a significant feature even when controlling for the other features, it should remain in the model.

Province appears to be a significant feature in the model, with most levels being significantly different from the baseline province. The signs are also not the same, meaning they are flucating between being more liberal and less liberal depending on the province. below is a table and plot of the estimated liberal percentage in each province after the post stratification. We are controlling for the other variables here.

Table 9: Post-stratified Weighted Estimate of Province on LPC Vote Share, shown as table

province_abbr	weighted_odds	pop	weighted_prob
AB	-2.1517244	1548	0.1041702
BC	-1.3862146	2247	0.2000128
MB	-1.3161728	1064	0.2114558
NB	-1.2459230	1249	0.2234067
NL	-0.9779906	1034	0.2732907
NS	-1.0111865	1344	0.2667477
ON	-0.7635939	5064	0.3178665
PE	-0.9856231	650	0.2717775
QC	-1.2468746	3516	0.2232416

province_abbr	weighted_odds	pop	weighted_prob
SK	-2.1907826	1034	0.1005813

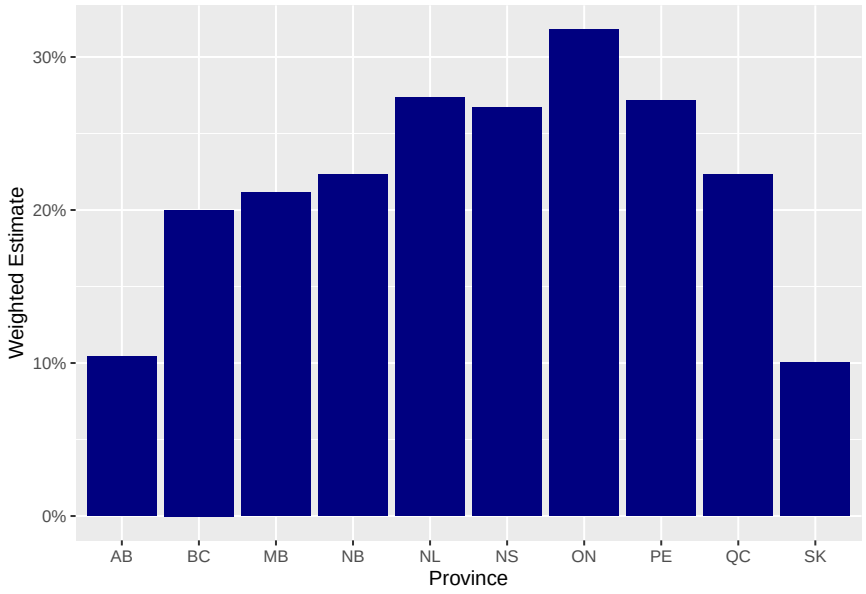

Figure 6: Post-stratified Weighted Estimate of Province on LPC Vote Share, shown in bar plot

It is clear to see that Ontario (ON) had the strongest estimated liberal voting, controlling for the other variables in the model. AB and SK were on the other end, with the lowest liberal voting after controlling for the other features of the respondents.

Age is a little bit different, most of the groups seems to be about the same until age 55, then there appears to be a monotonic increase in liberal voting in the next 2 groups after controlling for the other features.

Table 10: Post-stratified Weighted Estimate of Age Groups on LPC Vote Share, shown as table

age_group	weighted_odds	pop	weighted_prob
18 to 24	-1.360865	882	0.2040997
25 to 34	-1.367193	2359	0.2030737
35 to 44	-1.280549	2864	0.2174569
45 to 54	-1.440974	2836	0.1913946
55 to 64	-1.186234	3970	0.2339332
65+	-1.034488	5839	0.2622148

Education seems to be obviously a monotone relationship with liberal voting. As education increases, so does the likelihood of voting liberal even controlling for the other features.

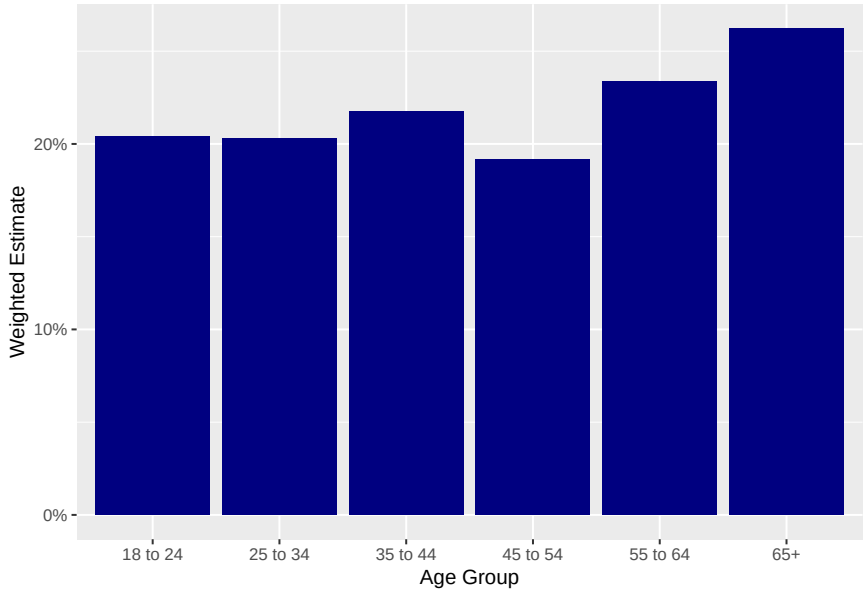

Figure 7: Post-stratified Weighted Estimate of Age groups on LPC Vote Share, shown in bar plot

Table 11: Post-stratified Weighted Estimate of Education on LPC Vote Share, shown as table

education	weighted_odds	pop	weighted_prob
<BA	-1.3668934	13654	0.2031222
BA	-0.9433492	3443	0.2802243
>BA	-0.6157275	1653	0.3507538

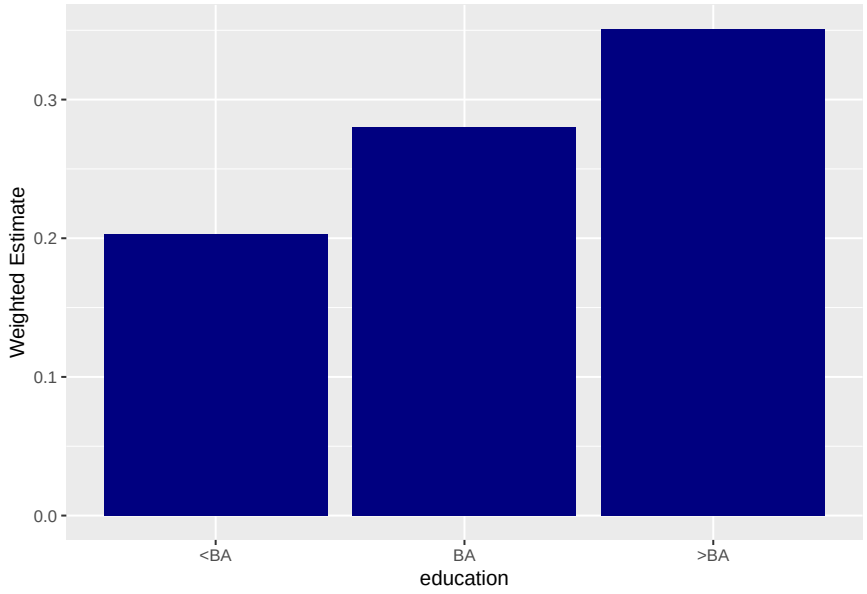

Figure 8: Post-stratified Weighted Estimate of Education on LPC Vote Share, shown in bar plot

Conclusions

Our main hypothesis was that males, younger citizens and residents of Alberta and Saskatchewan are less likely to vote for LPC. That hypothesis held up, aside from male voters being less likely. Males did have a negative estimated effect in the model, but it was not statistically significant when we controlled for other variables, even if the observed difference is larger (23.74% for males, 24.96% for females).

Another finding was how liberal Ontario was compared to the other Provinces. Education level also had a significant impact on liberal voting, with more educated people being more likely to vote liberal on average, controlling for the other model features.

This analysis is not perfect, ideally we would be able to input the age value, rather than an age group. We also might not have the most representative sample, but we did the best with what was provided by applying the post-stratification technique. Future work here might be collecting more data about the respondents, such as how strongly they felt about certain issues, or how likely they think they would be to vote liberal in the future to gauge how many people were strongly liberal versus on the fence.

Bibliography (MLA 8)

1. Bricker, Darrell. *Significant Gender Gap in Voting Intentions among Younger Canadians; Boomers Vote as Block, regardless of Gender.* Ipsos, 4 Oct. 2019, [www.ipsos.com/en-ca/news-polls/Significant-Gender-Gap-in-Voting-Intentions-Among-Younger-Canadians]. Accessed 19 Nov. 2022.
2. —.*Election Night Results - Provinces & Territories.* Enr.elections.ca, Elections Canada, enr.elections.ca/Provinces.aspx?lang=e. Accessed 19 Nov. 2022.
3. —.*The Electoral System of Canada.* Www.elections.ca, Elections Canada, 15 Apr. 2021, [www.elections.ca/content.aspx?section=res&dir=ces&document=part1&lang=e#p13]. Accessed 18 Nov. 2022.
4. —.*Voter Turnout at Federal Elections and Referendums.* Elections.ca, Elections Canada, 2019, [www.elections.ca/content.aspx?section=ele&dir=turn&document=index&lang=e]. Accessed 18 Nov. 2022.
5. —.*Voter Turnout by Sex and Age.* Www.elections.ca, Elections Canada, 6 Aug. 2020, [www.elections.ca/content.aspx?section=res&dir=rec/eval/pes2019/vtsa&document=index&lang=e]. Accessed 19 Nov. 2022.
6. Hopper, Tristin. *FIRST READING: The Least Popular Canadian Government Ever Elected.*" National Post, 22 Sept. 2021, [nationalpost.com/news/politics/election-2021/first-reading-the-least-popular-canadian-government-ever-elected.Leger]. NUMÉRO de PROJET. Association for Canadian Studies, 22 June 2020.
7. Parliament of Canada. *Elections and Candidates.* Lop.parl.ca, Parliament of Canada, [lop.parl.ca/sites/ParlInfo/default/en_CA/ElectionsRidings/Elections]. Accessed 19 Nov. 2022.
8. Statistics Canada. *The General Social Survey: An Overview.* Statcan.gc.ca, 2010, [www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2013001-eng.htm].
9. Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, *2019 Canadian Election Study - Phone Survey Technical Report*,2019 Canadian Election Study (CES) - Phone Survey,[https://doi.org/10.7910/DVN/8RHLG1/1PBGR3], Harvard Dataverse, V1