Forecasting Canadian Elections, an attemp utilizing post-stratification

Group 68: Yinzhou Liu, Junke Hou, Chuxin Chen

November 27, 2022

Introduction

Following the increase of democratic countries in the last century, elections have become ever so important. Election results could decide the outcome of a nation, or even influence the political environment of the world. Naturally, forecasting such events would be a hot topic of interest.

A common method of forecasting still used today is polling, but forecasting by simple polling had its flaws. In the past, polls are carried out by mail or by contacting individuals via phone book or car registration list(Lusinchi, 2012). While this method, a form of convenience sampling, has the potential to reach a lot of respondents, the data collected are often not very representative of the overall voting population. By only using the sample data to forecast, we would arrive at very biased and incorrect results.

While simple sampling with no considerations for representation is flawed, statisticians have found ways of correcting such mistakes. Methods such as multilevel regression and post-stratification that could solve the "unrepresentative" problem of polling can be utilized. For this paper we will be applying such methods on Canadian election data. While elections in Canada are required to take place every 5 years, because of how the Canadian political system works, the prime minister can call an election any time. Such uncertainty makes forecasting the outcome of Canadian elections more important.

Canada has a unique political voting system, instead of voting for a candidate, citizens votes for a candidate in their ridings(Canadian electoral district), and the candidate with the most votes becomes a member of parliament. Then within the parliament, the party with the most seats will win the final election(Schwartz, 2021). Our research question is to forecast the Canadian election by predicting the popular vote percentage a Canadian political party receives using survey data.

Due to the complexity, we will be working with the popular vote percentage instead, while historically there are occasions where the party with higher votes received fewer seats in the parliament thus losing elections, popular vote percentage is still a very important indicator of how well a party performs during the election. The problem of this survey is the responses are not representative of the entire Canadian voting population as displayed through various graphs in later sections.

This paper uses survey data during the 2019 campaign phase to "forecast" future elections. Hypothetically, this means our forecasted result should correspond to the actual outcome of the 2019 election. If our hypothesis is correct, that means our method and models have the potential to forecast future elections. By changing the data used, we can also extend our models to forecast other elections.

Data

We are using two data sets, "GSS" or general social survey, representative of the Canadian voting population and GES2019, or Canadian Election Study.

GSS2017 serves as a census data, representative of the overall population. GSS targets Canadian population 15 years of age and older, the respondents are also screened in based on representative criteria. Provinces and CMAs(Census metropolitan areas) are divided into stratas and the resulting survey is re-weighted by Statistics Canada, all to ensure that the survey is representative(Government of Canada, 2019).

GES 2019 is conducted during the 2019 election campaign period by phone, the respondents were asked basic demographic information such as age and gender, and also were asked a series of question regarding the election such as preferred party and opinions on matters such as being eco-friendly or on the matter of refugees(Stephenson et.al, 2022).

Cleaning

Upon importing our data sets, we clean our data set. For GSS 2017, The data cleaning process consists of three parts. First, we separate the age variable into five categories: less than 25, between 25 to 40, between 40 to 55, between 55 to 70, and over 70. Second,we use 1 to represent those who has religious affiliation, and 0 to represent those who does not have religious affiliation. Finally, we only keep our interested variable, that includes age, sex, province and whether one has religious affiliation, and drop all the NA values.

For GES 2019, the data cleaning process consists of several parts. First, We create a new variable indicating age using 2019 subtracting the year of birth in the survey data. Second, we create a new variable indicating whether one votes for liberal using q11 from survey data, and use 1 to represent those who would vote for liberal and 0 to represent those who does not, we repeat this process for all the other major Canadian political parties and create several new columns. Third, we rename the q2 column which stands for gender, replace 1 with male, 2 with female, and 3 with others, we simplified this column because inorder to perform post-stratification, the information has to correspond to our data in GSS2017. Then for q4 in the survey data, which represents the location of the respondent, replace each number with the province it represents accordingly. Afterwards, we create a new variable that indicates whether the respondent is religious or not, using the data of q62 in the survey data, and use 0 to represent those who are not religious and 1 to represent those who are. At last, we proceed to categorize our age variable similarly to what we did for our GSS 2017, we select our interested columns and drop all NA values.

Our main variables are the four demographic variables which we are going to perform post-stratification on, that includes age, sex, province that the respondent is in and whether they are religious. Our response variables or variables of interest are whether a respondent votes for a Canadian political party, contained only in GES 2019.

A summary table can be seen below.
*Note: We will be using census to indicate GSS2017 and survey to indicate GES 2019 for convenience.

Visualizations

Figure 1: Barplot of main parties vote per province participating in the survey

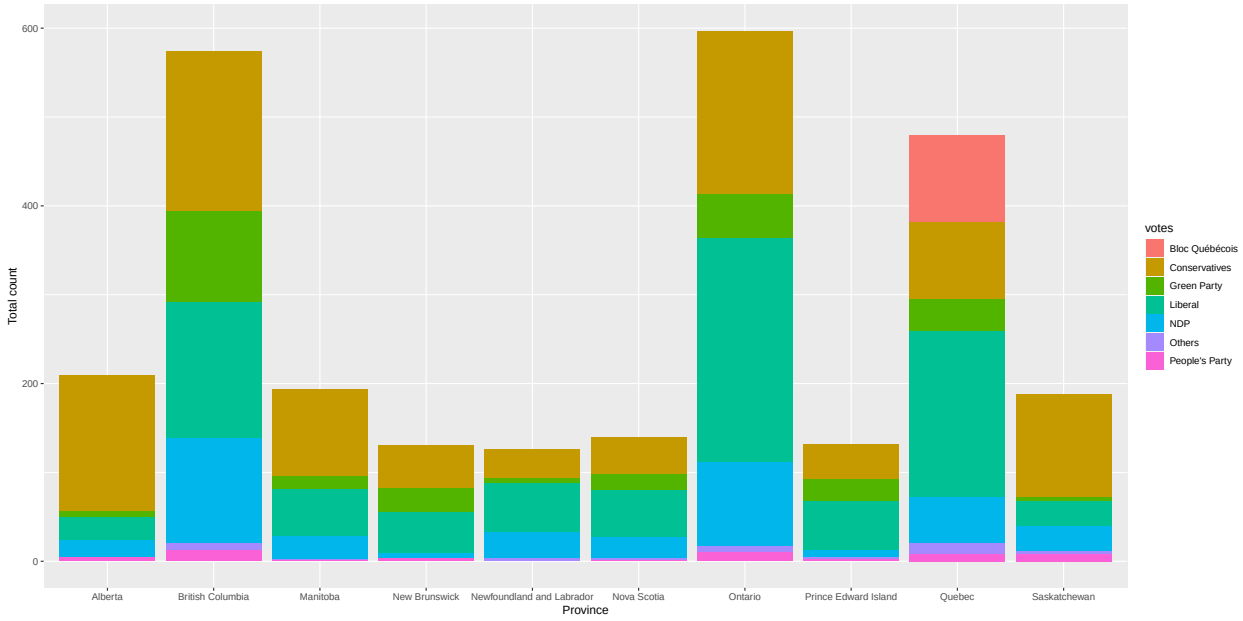

Figure 1 is a barplot showing number of voters for different parties for each province, with x axis showing name of the province and y axis showing numbers of vote. The total number of voters are not same across different provinces. Provinces such as Ontario, Quebec, and British Columbia have significantly more voters than other provinces, due to their high population. Each province has different ratio of voting to different parties. For example, in Quebec voting for Bloc Québécois party is significantly higher than other province; in Alberta and Saskatchewan, most voters prefer conservative party; in Ontario, most voters prefer liberal party. Such difference indicates that we are likely to get a biased estimate if we assume every province has same preference of the political parties and model using only survey data.

Figure 2: Density of age in GES2019 and GSS2017

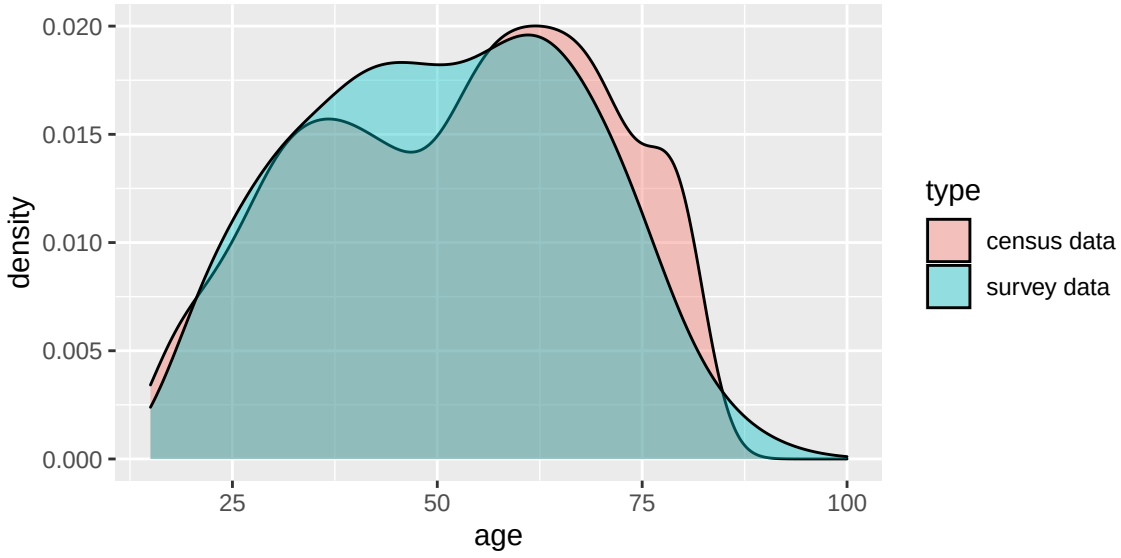

Figure 2 is a density plot(Holtz, n.d.) of age in census and survey data. The purpose of plotting this graph is to compare the distribution of age between census and survey data. The distribution of age below 30 is very similar in both data sets; for age between 30 and 55, the survey data has higher density; for age between 55 and 80, the census data has higher density. The plot shows that distribution of age is not same across census and survey data. This means if we are to only model using survey data, our results will be biased due to the difference in the distribution of age among the respondents and the public.

Figure 3: Boxplot of the age of support for each main party

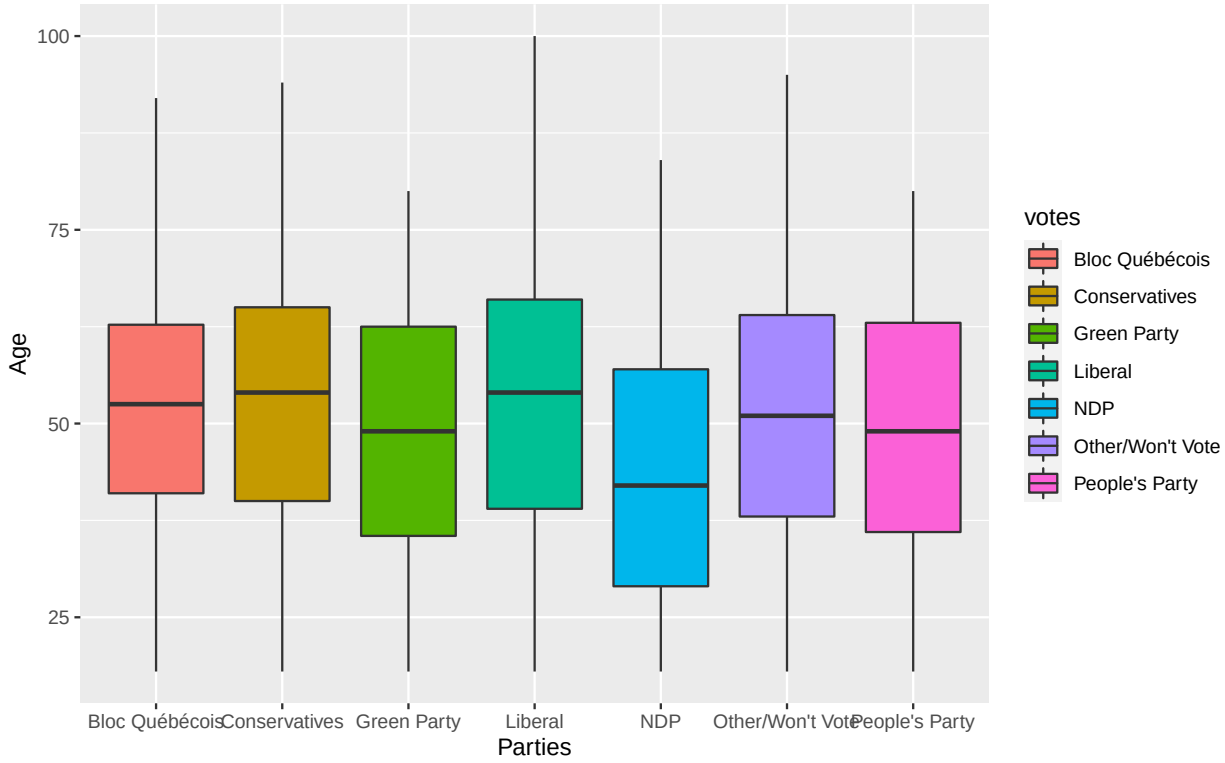

Figure 3 is a side-by-side boxplot comparing the age of people voting for different parties. The median of each group is mostly above 50, with a few exceptions: Green party, NDP, and people's party. NDP has the lowest median and quartiles, possibly indicating that this party is preferred among younger generations. For liberal and conservation, their age medians are slightly higher than other parties, indicating those two parties are preferred among older generations. Provided with those differences, the plot shows that age could be a useful predictor of popular vote for different parties. These differences paired with the differences in density observed in Figure 2 all serves as motivations for post-stratification.

Numerical Summary

Table 1: Summary of variable Age by Gender

votes	Count	Mean	Median	Variance
Bloc Québécois	98	52.25510	52.5	277.8002
Conservatives	980	52.75408	54.0	272.8434
Green Party	287	48.96864	49.0	275.3242
Liberal	909	52.62046	54.0	302.8987
NDP	405	43.59259	42.0	271.3064
Other/Won't Vote	1209	50.97932	51.0	261.6742
People's Party	49	48.93878	49.0	276.3087

Table 1 is a summary of our numerical variable age of respondents from the survey data. We see the average age of all voters is around 50, with a lot of variations between voters of each political party. For example, those who voted for NDP from our survey data has an average age of 43, almost a 10-year difference from those who voted for the Conservative party. This suggests that the distribution of age might be different for those who vote for different political parties and as such, could be used as a predictor when forecasting their voting preference.

Methods

We will first run a regression model on our GES2019 data, than perform a post-stratification by re-weighting the GES 2019 data to be representative of the overall population. Our regression model uses a logit regression, as our responding variables are binary and categorical. The results obtained from our regression model and after transformation, is the probability of a party getting voted. This corresponds with the percentage of votes they receive from the public. Linear model is not used as our outcome variable here is binary(whether a respondent votes for a party).

Model Specifics

We will be creating 6 logistic regression models to model whether people will vote for 6 main parties or not. We used a method called AIC to evaluate the goodness of fit of the model to the data generated from it. In particular, we mainly compare three different types of models for each party. The results are presented in the table following, and the different model types are listed below the table.

Table 2: AIC Tabel For Model Selection

	Model Type 1	Model Type 2	Model Type 3
Liberal	3496.5666	3501.6346	3495.0135
Conservatives	3469.6860	3524.2281	3540.8321
NDP	2187.5645	2211.1629	2215.3853
Bloc Québécois	858.6798	854.7515	856.7307
Green Party	1800.7887	1799.8348	1854.1788
People's party	502.6657	499.7582	500.7201

[a] Type 1 = Including Age, Sex, Religion
[b] Type 2 = Including Age, Religion
[c] Type 3 = Including Age, Sex

From the table, we can observe that the AIC of Model Type 1 is the smallest when predicting the voting rate of Conservatives and NDP. In addition, although the AIC of Model Type 1 is not the smallest among the other parties, it is similar to the AIC of the other two Model Types. Since the AIC of Model Type 1 includes

one more variable, we conclude that Model Type 1 is the best choice for predicting the voting rate for each party.

After deciding the model, we will be using age, gender, and whether they are religious as independent variables, which are recorded as categorical data. We will use these models to predict the probability of voting for 6 different main Canadian political parties. The logistic regression model we are using is:

$$y = \beta_0 + \beta_1 x_{age} + \beta_2 x_{sex} + \beta_3 x_{religion} + \epsilon$$

Where y represents whether the individual will support a particular party or not, β_0 represents when all independent variables are 0, the output of logistic regression will be equal to β_0. In addition, β_1, β_2, β_3 are the coefficients of the three independent variables age gender and the presence or absence of religion, respectively.

*Since there are too many models used, we will put the summary of these models in the Appendix.

Post-Stratification

We will be using post-stratification to correct the representation problem present in our survey data. As observed in graph in the data section, many demographics aspects of the survey sample, not just age as we showed above, doesn't match with our survey. So any results of models conducted solely on our survey data is biased and incorrect.

By performing post-stratification, we are taking each samples from our survey data, categorizing them by their demographics (age, sex and religion), and re-weighting them based on the distribution of their demographic data according to the census so they correctly represents the actual proportion for the entire Canadian population.

In Post-Stratification, we want to observe the voting rate of each main party in terms of province, so we choose "province" as our group variable, and then we group the census data by province and other independent variables to enumerate different cells in order to predict their $\hat{y_j}$.

Next, we use the survey data to create a logistic regression for each of the main party to predict the voting rate, then apply the model to Post-stratification data which is formed by census to obtain the regression output for each cell. After that, in order to derive the probabilities which is $\hat{y_j}$, we need to perform the following calculations on the outputs obtained from the logistic regression.

$$\hat{y_j} = \frac{exp(y)}{1 + exp(y)}$$

In the final step, we will use $\hat{y}^{PS} = \frac{\sum N_j \hat{y_j}}{\sum N_j}$ to infer how the entire population in each province will vote.

Results

Table 3: Result of Post-Stratification

province	Liberal	Conservatives	NDP	Québécois	Green	Peoples
Alberta	0.3338	0.3471	0.1523	0.0341	0.0986	0.0178
British Columbia	0.3401	0.3314	0.1503	0.0353	0.1106	0.0170
Manitoba	0.3423	0.3631	0.1363	0.0346	0.0914	0.0171
New Brunswick	0.3458	0.3701	0.1290	0.0353	0.0867	0.0164
Newfoundland and Labrador	0.3424	0.3785	0.1278	0.0349	0.0806	0.0167
Nova Scotia	0.3433	0.3630	0.1329	0.0351	0.0918	0.0169
Ontario	0.3411	0.3586	0.1391	0.0350	0.0926	0.0168
Prince Edward Island	0.3458	0.3642	0.1328	0.0348	0.0867	0.0165
Quebec	0.3424	0.3757	0.1298	0.0345	0.0824	0.0170
Saskatchewan	0.3404	0.3593	0.1393	0.0338	0.0918	0.0177

Table 4: Mean of Post-Stratification

	Liberal	Conservatives	NDP	Québécois	Green	Peoples
Mean	0.34174	0.3611	0.13696	0.03474	0.09132	0.01699

We can observe the results of our forecast by province for each part in the above table.

Overall, we see roughly similar votes among the provinces for Liberal party, at around 34%. Votes for the Conservative party varies within provinces, ranging from 33.14% to 37.57%, slightly more votes compared to the Liberal party. For NDP, the results show a much smaller percentage, ranging from 12.78% to 15.23%. For Bloc Québécois, the forecasted results sit round 3.4%, for Green Party, the forecasted results vary but is almost always close but below 10%, at last for the People's party of Canada, the votes fluctuates around 1.6% and 1.7%.

All of our results value are within logical ranges and can be interpreted. For example, our forecasting model suggests that in the province of Ontario, the liberal party received 34.11% of votes.

Conclusions

Across the provinces, we observe on average, that 34.174% voted for the Liberal party, 36.11% voted for the Conservative party, 13.696% voted for the NDP, 3.474% voted for the Bloc Quebecois, 9.132% voted for the Green party, and 1.699% voted for the People's party of Canada. Comparing this to the actual overall votes percentage in 2019, where 33.1% voted for the Liberal party, 34.3% voted for the Conservative party, 16% voted for the NDP, 7.6% voted for the Bloc Quebecois, 6.5% voted for the Green party, and 1.699% voted for the People's party of Canada(Heard, n.d.).

We see small differences among the lesser voted parties, but very similar data for both Liberal and Conservative party. If we look at graph 1, we do see that votes for Liberal and Conservatives takes up a lot of the survey responses, thus after reweighting by post-stratification, the votes forecasted for these two parties will be more accurate. Overall, results from our model successfully captures the overall trend of this election.

Our original hypothesis stated that our forecasted results should correspond to the actual results in 2019.While our results are somewhat promising and contains explainable values, our forecasted results by province do not match the actual data by province. For example, Bloc Quebecois, a Quebec based political

party actually received over 32% of votes in the province of Quebec but 0% of votes else were(Elections Canada, 2019). However our model forecasts around 3.4% votes for Bloc Quebecois for all provinces. This raises a flag about the interpretabality by province. The cause of this is simple, because we are utilizing post-stratification by re-weighting according to demographic data of each province, we are leaving out the province data during our first stage models. In other words, our original logistic model neglects the effect of provinces on voting in-order to later perform post-stratification with provinces as groups. This is a drawback on the interpretability aspect of our model, meaning we should interpret with care. Regardless, our results by provinces still provides very useful insights on the election outcomes.

We conclude we can extend our model to forecast other elections, with potential future steps such as using more data sets, finding more overlapping variable to perform post-stratification on or applying more models in hopes of achieveing more accurate estimates.

Bibliography APA style

Lusinchi, D. (2012). "president" Landon and the 1936 literary digest poll: Were automobile and telephone owners to blame? Social Science History, 36(1), 23–54. https://doi.org/10.1215/01455532-1461650

Government of Canada, Statistics Canada. (2019, February 6). General Social Survey - Family (GSS). Surveys and statistical programs. Retrieved November 24, 2022, from https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4501

Elections Canada. (2019). Forty-third general election 2019. Official Voting Results. Retrieved November 26, 2022, from https://www.elections.ca/res/rep/off/ovr2019app/51/table9E.html

Holtz, Y. (n.d.). Density Chart with several groups. – the R Graph Gallery. Retrieved November 26, 2022, from https://r-graph-gallery.com/135-stacked-density-graph.html

Heard, A. (n.d.). Canadian election results by party 1867 to 2021. Canadian Election Results: 1867-2021. Retrieved November 26, 2022, from https://www.sfu.ca/~aheard/elections/1867-present.html

Schwartz, M. (2021, August 26). How Canada's Electoral System Works. CIC News. Retrieved November 26, 2022, from https://www.cicnews.com/2021/08/how-canadas-electoral-system-works-0819016.html#gs.jsq4vu

Stephenson, L. B., Harell, A., Rubenson, D., & Loewen, P. J. (2022, March 21). 2019 Canadian Election Study (CES) - phone survey. Harvard Dataverse. Retrieved November 29, 2022, from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2F8RHLG1&version=1.1

Appendix

Due to the large number of models we use, the six models summery in the "Method" section are placed here.

```
##
## Call:
## glm(formula = vote_liberal ~ age + sex + have_religion, family = "binomial",
##     data = survey_data_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0740  -0.9034  -0.8386   1.4092   1.6587
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.84663    0.16418  -5.157 2.52e-07 ***
## age25to40       0.14273    0.17294   0.825  0.40918
## age40to55       0.16132    0.17009   0.948  0.34292
## age55to70       0.25749    0.16891   1.524  0.12739
## ageover70       0.53952    0.18351   2.940  0.00328 **
## sexMale        -0.23784    0.08219  -2.894  0.00381 **
## sexOther      -10.86216  196.96771  -0.055  0.95602
## have_religion   0.05897    0.08828   0.668  0.50419
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3505.3  on 2768  degrees of freedom
## Residual deviance: 3480.6  on 2761  degrees of freedom
## AIC: 3496.6
##
## Number of Fisher Scoring iterations: 10
```

```
##
## Call:
## glm(formula = vote_conservatives ~ age + sex + have_religion,
##     family = "binomial", data = survey_data_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1898  -0.9325  -0.8056   1.2041   2.0228
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.90748    0.17724 -10.762  < 2e-16 ***
## age25to40       0.31194    0.18001   1.733  0.08312 .
## age40to55       0.44144    0.17563   2.513  0.01195 *
## age55to70       0.53316    0.17436   3.058  0.00223 **
## ageover70       0.40944    0.19067   2.147  0.03176 *
## sexMale         0.63666    0.08478   7.510 5.92e-14 ***
## sexOther       -9.97051  196.96771  -0.051  0.95963
## have_religion   0.76672    0.09174   8.357  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3598.8  on 2768  degrees of freedom
## Residual deviance: 3453.7  on 2761  degrees of freedom
## AIC: 3469.7
##
## Number of Fisher Scoring iterations: 10
```

```
##
## Call:
## glm(formula = vote_NDP ~ age + sex + have_religion, family = "binomial",
##     data = survey_data_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0582  -0.5862  -0.5003  -0.3845   2.4996
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.2871     0.1712  -1.677   0.0936 .
## age25to40      -0.5199     0.1811  -2.870   0.0041 **
## age40to55      -0.8342     0.1849  -4.513 6.40e-06 ***
## age55to70      -1.1049     0.1910  -5.784 7.29e-09 ***
## ageover70      -1.6162     0.2553  -6.332 2.43e-10 ***
## sexMale        -0.5529     0.1111  -4.976 6.48e-07 ***
## sexOther       13.3730   324.7437   0.041   0.9672
## have_religion  -0.6230     0.1136  -5.483 4.18e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2304.8  on 2768  degrees of freedom
## Residual deviance: 2171.6  on 2761  degrees of freedom
## AIC: 2187.6
##
## Number of Fisher Scoring iterations: 11
```

```
##
## Call:
## glm(formula = vote_bloc_québécois ~ age + sex + have_religion,
##     family = "binomial", data = survey_data_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3095  -0.2975  -0.2694  -0.2250   2.7454
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.43482    0.42209  -8.138 4.03e-16 ***
## age25to40      -0.26018    0.46470  -0.560    0.576
## age40to55       0.38919    0.42543   0.915    0.360
## age55to70       0.18698    0.43184   0.433    0.665
## ageover70       0.15578    0.47599   0.327    0.743
```

```
## sexMale         0.03093    0.21006   0.147    0.883
## sexOther       -9.87107  535.41126  -0.018    0.985
## have_religion  -0.05011    0.22175  -0.226    0.821
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 847.38  on 2768  degrees of freedom
## Residual deviance: 842.68  on 2761  degrees of freedom
## AIC: 858.68
##
## Number of Fisher Scoring iterations: 12
```

```
##
## Call:
## glm(formula = vote_green_party ~ age + sex + have_religion, family = "binomial",
##     data = survey_data_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6565  -0.5555  -0.3993  -0.3607   2.3754
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.445433   0.223077  -6.480 9.20e-11 ***
## age25to40      -0.134684   0.242285  -0.556   0.5783
## age40to55       0.020227   0.238186   0.085   0.9323
## age55to70      -0.074590   0.241122  -0.309   0.7571
## ageover70      -0.007073   0.272365  -0.026   0.9793
## sexMale        -0.210662   0.127538  -1.652   0.0986 .
## sexOther      -10.985945 324.743733  -0.034   0.9730
## have_religion  -0.968971   0.130572  -7.421 1.16e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1844.3  on 2768  degrees of freedom
## Residual deviance: 1784.8  on 2761  degrees of freedom
## AIC: 1800.8
##
## Number of Fisher Scoring iterations: 11
```

```
##
## Call:
## glm(formula = vote_peoples_party ~ age + sex + have_religion,
##     family = "binomial", data = survey_data_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.2537  -0.2056  -0.1751  -0.1639   3.0393
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)    -4.47476    0.63752  -7.019 2.23e-12 ***
## age25to40       0.74991    0.63150   1.187    0.235
## age40to55      -0.06262    0.67371  -0.093    0.926
## age55to70       0.32423    0.64544   0.502    0.615
## ageover70       0.18100    0.71862   0.252    0.801
## sexMale         0.30481    0.30365   1.004    0.315
## sexOther       -9.84122  535.41129  -0.018    0.985
## have_religion  -0.07130    0.30497  -0.234    0.815
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 492.50  on 2768  degrees of freedom
## Residual deviance: 486.67  on 2761  degrees of freedom
## AIC: 502.67
##
## Number of Fisher Scoring iterations: 12
```