Using an MRP model to predict the 2019 Canadian election's full voter turnout results

Danny Xu

12/4/2020

The code for this paper is located at the following repository: https://github.com/valuemeta/STA304-Final. The data is located separately and can be found from the CES website and Statistics Canada.

Abstract

Multilevel Models with Poststratification (MRPs) are being used frequently in forecasting election results. We seek to apply these models instead to past elections to gain insight into the election results if every eligible person had voted. Specifically, the 2019 Canadian election will be explored, using data from the CES 2019 Web Survey to fit our model and the Canadian Census to poststratify. A frequentist logistic multilevel model with random intercepts for each riding is used to predict the proportion of people who vote for each of the 5 parties in Canada that won seats in the House of Commons in the election. The results of the analysis would be used to justify or reject the idea that everybody's vote matters.

Introduction

Eligible voters in Canada are encouraged to go out and vote in all elections. They are constantly told that their vote matters and changes the country for the next few years. This paper will focus on examining if there is indeed a change in the number of seats that a political party has in the House of Commons if every person in Canada had voted during the 2019 Canadian election. These results are important because the truth of one of the basic principles of democracy would depend on them.

A popular method of predicting political results are MRP models. MRP models will be used in this paper to predict which seats go to which party for each of the Canadian federal electoral districts. The results of the MRP models will be compared to the actual results of the 2019 Canadian election to determine the change in seats for each of the political parties. In addition, further investigation will be conducted on which federal electoral districts changed political party due to the extra voter turnout.

Two datasets will be used to conduct the analysis in this paper. They consist of a sample dataset containing information about political leanings of individuals in Canada as well as a few other variables, and another more representative dataset that is missing political information. The Methodology section elaborates further on the 2 datasets used, as well as the MRP models in-depth. The results of the model are placed in the Results section. Conclusions drawn from these results are located in the Conclusion Section.

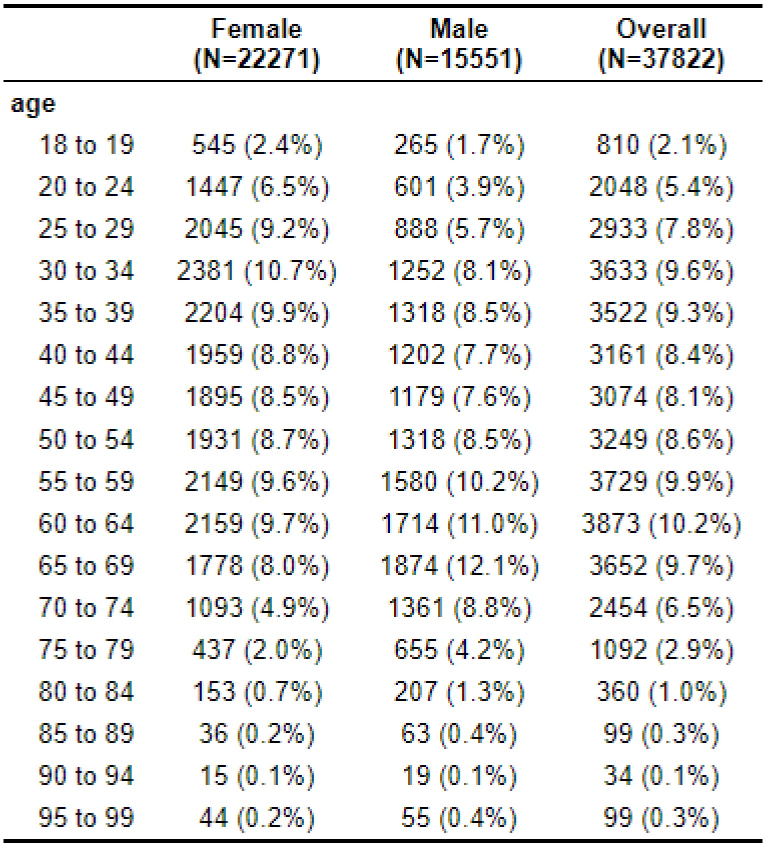
Keywords

MRP,Poststratification,Multilevel Model,Canadian Politics,Frequentist,Logistic Regression

Methodology

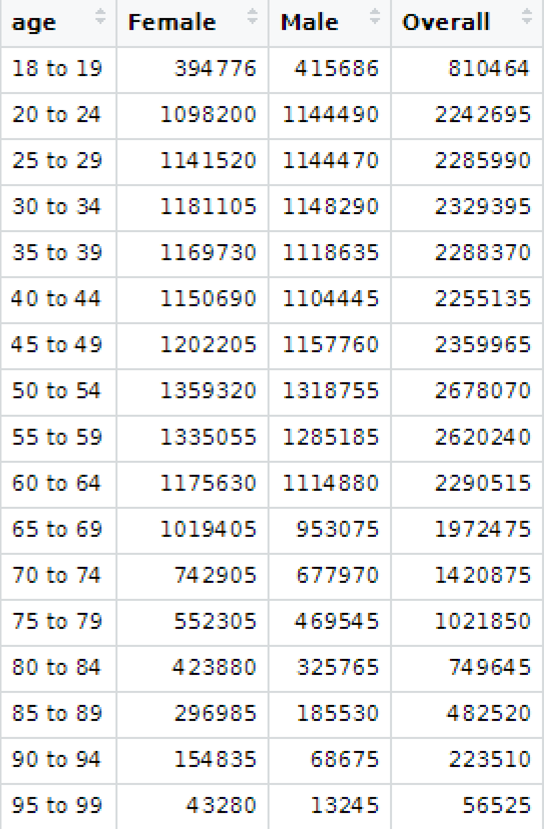
Data:

Table 1.



	Female (N=22271)	Male (N=15551)	Overall (N=37822)
age			
18 to 19	545 (2.4%)	265 (1.7%)	810 (2.1%)
20 to 24	1447 (6.5%)	601 (3.9%)	2048 (5.4%)
25 to 29	2045 (9.2%)	888 (5.7%)	2933 (7.8%)
30 to 34	2381 (10.7%)	1252 (8.1%)	3633 (9.6%)
35 to 39	2204 (9.9%)	1318 (8.5%)	3522 (9.3%)
40 to 44	1959 (8.8%)	1202 (7.7%)	3161 (8.4%)
45 to 49	1895 (8.5%)	1179 (7.6%)	3074 (8.1%)
50 to 54	1931 (8.7%)	1318 (8.5%)	3249 (8.6%)
55 to 59	2149 (9.6%)	1580 (10.2%)	3729 (9.9%)
60 to 64	2159 (9.7%)	1714 (11.0%)	3873 (10.2%)
65 to 69	1778 (8.0%)	1874 (12.1%)	3652 (9.7%)
70 to 74	1093 (4.9%)	1361 (8.8%)	2454 (6.5%)
75 to 79	437 (2.0%)	655 (4.2%)	1092 (2.9%)
80 to 84	153 (0.7%)	207 (1.3%)	360 (1.0%)
85 to 89	36 (0.2%)	63 (0.4%)	99 (0.3%)
90 to 94	15 (0.1%)	19 (0.1%)	34 (0.1%)
95 to 99	44 (0.2%)	55 (0.4%)	99 (0.3%)

Table 2.



age	Female	Male	Overall
18 to 19	394776	415686	810464
20 to 24	1098200	1144490	2242695
25 to 29	1141520	1144470	2285990
30 to 34	1181105	1148290	2329395
35 to 39	1169730	1118635	2288370
40 to 44	1150690	1104445	2255135
45 to 49	1202205	1157760	2359965
50 to 54	1359320	1318755	2678070
55 to 59	1335055	1285185	2620240
60 to 64	1175630	1114880	2290515
65 to 69	1019405	953075	1972475
70 to 74	742905	677970	1420875
75 to 79	552305	469545	1021850
80 to 84	423880	325765	749645
85 to 89	296985	185530	482520
90 to 94	154835	68675	223510
95 to 99	43280	13245	56525

The sample dataset from the CES 2019 Web Survey is summarized by Table 1. The 2016 Canadian Census dataset that was used for poststratification is summarized in Table 2. The datasets are recorded 3 years apart. In Table 1, it is important to note that the number of female respondents is far larger than the number of male respondents, even though the authors intended to make it a 50/50 split. Meanwhile in Table 2, the number of respondents of both sexes is about equal. In addition, Table 1 shows us that the distribution of females and males surveyed is not the same, 30-34 aged females respond the most within their sex category whereas 65-69 aged males responded the most within their sex category. From Table 2, we can see that the 50-54 age bracket contains the most amount of people, which is different from the data in Table 1.

Model:

The model is a logistic multilevel model with 2 predictor variables and a random intercept. We generate 5 models, one for each political party in Canada that won seats in the 2019 election. Our independent variables are sex and age, and we are predicting the probability of voting within a federal district.

The individual level equation is $log(\frac{Y_{ij}}{1-Y_{ij}}) = \beta_{0j} + \sum_{k=1}^{16} \beta_k X_{ij,k} + \beta_{male} X_{ij,male} + r_{ij}$. Y_{ij} refers to the probability of voting for the specified party which implies that $log(\frac{Y_{ij}}{1-Y_{ij}})$ are the log-odds of voting for the specified party. β_{0j} is the intercept of the jth district. β_k is the slope of the relationship between age brackets and the log-odds of voting for the given party, within the jth district, with respect to the smallest age bracket (18 to 19 years old). β_{male} is the difference in the log-odds of voting for a given party between male and female voters. r_{ij} is the error term. $X_{ij,k}$ are indicator variables that denote the age bracket category of the observation. Thus, at most only one of these terms are non-zero. $X_{ij,male}$ is the indicator variable that is 1 if the observation is male, and 0 otherwise.

The district level equation is $\beta_{0j} = r_{00} + r_{01} W_j + u_{0j}$. r_{00} is the overall intercept, which is the mean of all the log-odds of voting for the specified party across all districts for females of 18 to 19 years of age. r_{01} is the slope between the log-odds of voting for the given party and the jth district. W_j is an indicator variable that is 1 if the observation is located in the jth district and 0 otherwise. u_{0j} is the error term.

Results

We found that our model predicts that the Liberal Party would win 164 seats, the Conservative Party would win 141 seats, the Bloc Québécois would win 21 seats, the NDP would win 11 seats and the Green Party would win a single seat. In comparison, the actual results of the election were that the Liberal Party would win 157 seats, the Conservative Party would win 121 seats, the Bloc Québécois would win 32 seats, the NDP would win 24 seats, and the Green Party would win 3 seats [12]. Thus, the difference that occurred when everybody in the country voted was that the Liberal Party would gain 7 seats, the Conservative Part would gain 20 seats, the Bloc Québécois would lose 11 seats, the NDP would lose 13 seats, and the Green Party would lose 2 seats. Notice that in the actual election, one of the seats was not from any of the 5 big parties, which we have not accounted for in the models.

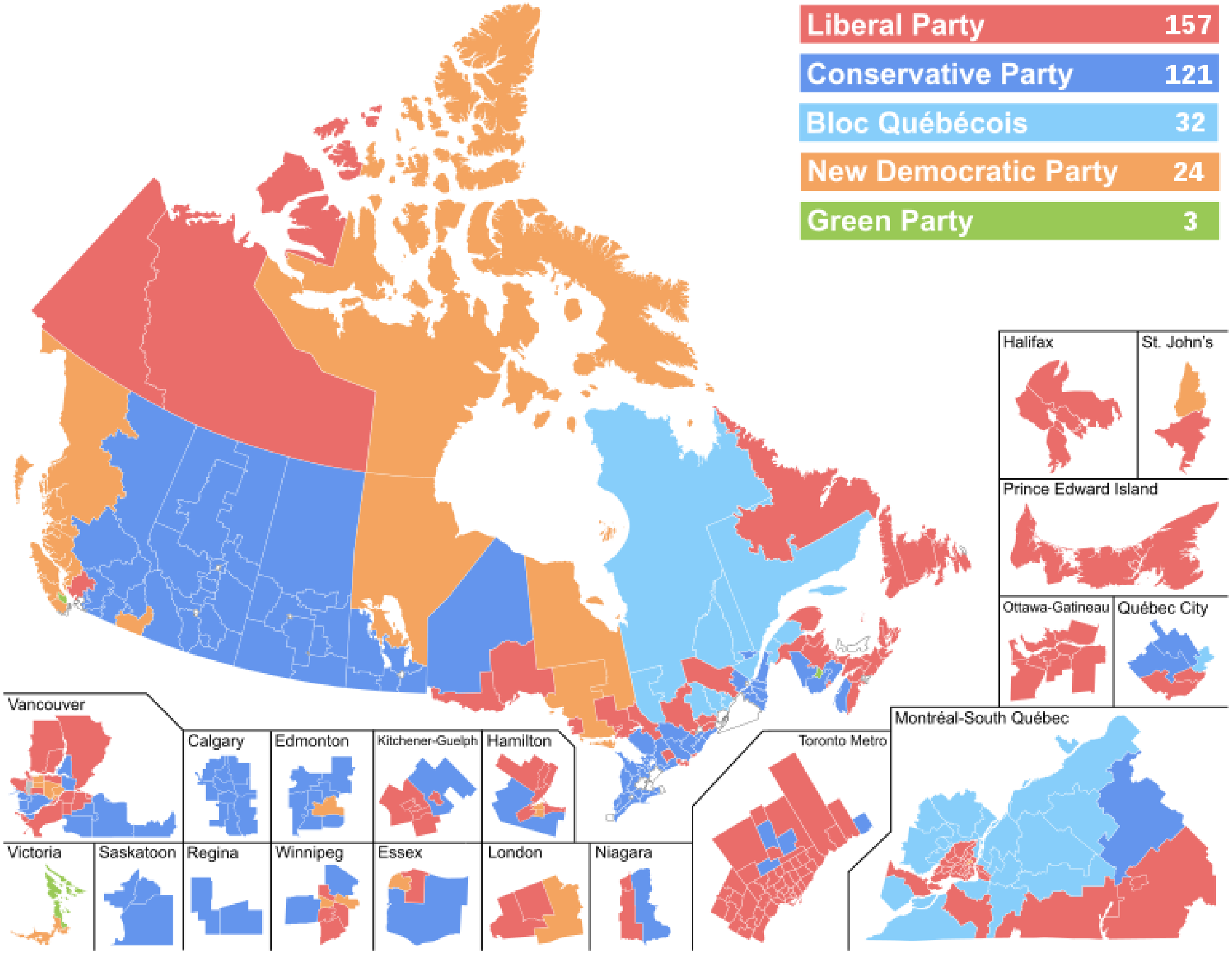


Figure 1: Actual election results by district

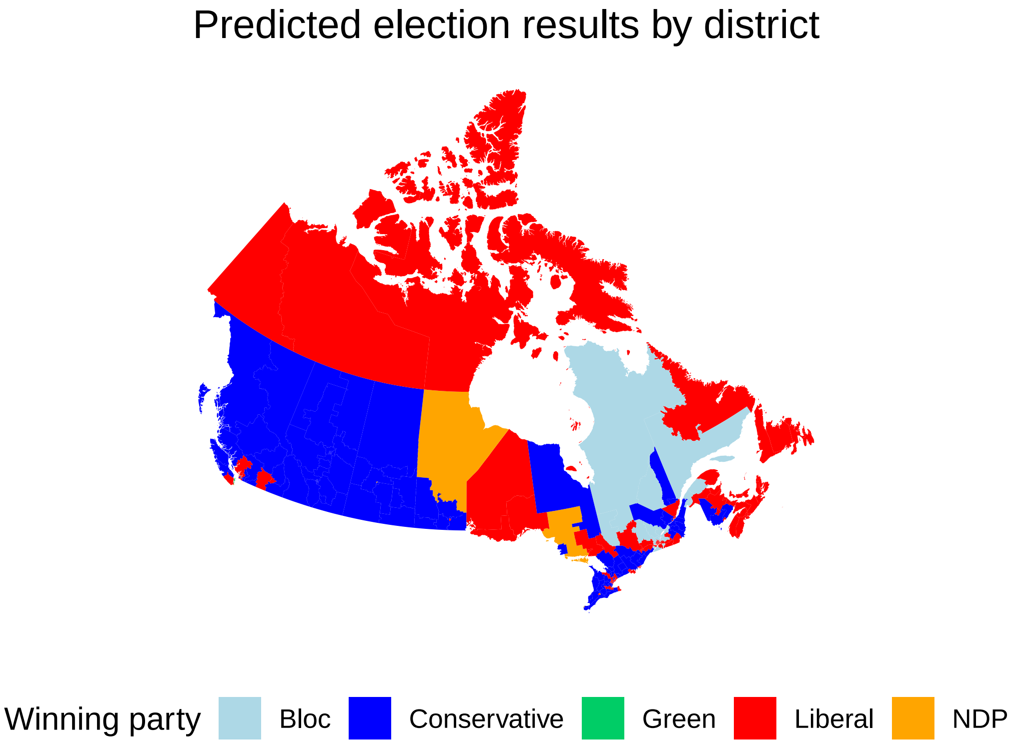


Figure 2

The summaries for the actual models are too lengthy to display in this section, but are located in the Appendix. Our models have two kinds of R^2 [11]. The marginal R^2 is the proportion of variance of the data explained by the fixed effects of the model, whereas the conditional R^2 is the proportion of variance of the data explained by the combined fixed and random effects of the model.

The liberal model does not have many statistically significant coefficients, and has very poor R^2 values using both the delta and theoretical methods of computing R^2 for random intercept logistic models. However, it and all the other models do significantly increase their R^2 when we factor in the random effects. In the conservative model, we have most of our coefficients are statistically significant at the 5 significance level. Also, we have a much better fit to the sample data than in the liberal case as the conditional R^2 is 0.184 or 0.130 depending on the computation method chosen. In the Bloc Québécois model, many of the coefficients are statistically significant, and we have an extremely good fit as the conditional R^2 is 0.851 or 0.777 depending on the computation method chosen. In the NDP model, most of our coefficients are statistically significant, and we have an R^2 of 0.160 or 0.070 depending on the computation method chosen. In the Green model, most of our coefficients are statistically significant, and we have an R^2 of 0.102 or 0.027 depending on the computation method chosen.

Discussion

Summary:

In this paper we created 5 random intercept logistic models to predict the party who wins ridings across Canada. We used 5 models to account for the fact that some of the people in our sample would not respond to the vote choice question. By making models for each party, we can find the party with the highest proportion of votes in any given riding, which allows us to determine a winner. A post-stratification process on these models fitted to the sample data to obtain the number of seats won for each party if everyone eligible had voted in the election.

Conclusions:

We find that sex and age bracket are somewhat statistically significant in determining who people will vote for. However, they do not explain much of the variation of the data, so more predictors would be necessary to generate a good statistical model for vote predictions. Despite this, we found that most ridings did not flip to other parties. This hints that the election is mainly determined by a few ridings that are competitive. Everyone's vote indeed counts in these competitive ridings. However, for ridings where it is not so competitive, adding everybody's vote does not change the outcome of the election. For example, most of the ridings in Quebec that were won by the Bloc were still won by the Bloc even after post-stratification.

In terms of the global impacts, the idea that everyone's vote counts, seems to falter in regions where ridings are not competitive. If a riding is significantly Liberal for example, there aren't enough people from other parties to outnumber the Liberals. This is an artifact of having first past the post voting (FPTP). The main problem is that there are wasted votes, that don't count towards the party that lost the riding but possibly received a large amount of votes still, which means the backing in that riding is still significant. This leads to misrepresentation of the support for parties in Canada in the House of Commons. Our result is that FPTP voting in Canada leads to smaller parties having less support if everyone votes.

Weakness & Next Steps:

The census dataset from Statistics Canada has some significant flaws. Firstly, the data is from 2016, which is 3 years apart from 2019 (which is when our sample dataset was procured). In addition, due to Canadian law, the census data is not available on a per observation basis. The variables are only available with summarized data, with each variable and sex distributions, but not distributions across variables. This means we can only select 1 variable apart from sex to use in our analysis, which significantly limits the scope and the models we can use. This is why the models' R^2 is low.

If a more recent post stratification dataset with observation by observation results is created, this problem can be solved and the analysis can be recreated using that dataset, but with the models updated to include many different variables to predict the votes in a riding for each party. Further analysis can be carried out on other elections as well, to get a more conclusive idea about how FPTP negatively impacts smaller parties when everyone's votes are counted. Looking at other countries who use a different method of determining winners in elections such as Australia's ranking system will be useful to compare FPTP and other systems to see which one is better in what way, and the trade offs.

References

1. Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, "2019 Canadian Election Study - Online Survey", https://doi.org/10.7910/DVN/DUS88V, Harvard Dataverse, V1

2. Statistics Canada. (2018). 2016 Census of Population for various levels of geography, including provinces and territories, census metropolitan areas, communities and census tracts. (Catalogue number 98-316-X2016001). Retrieved December 18, 2020 from Statistics Canada: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/download-telecharger/comp/page_dl-tc.cfm?Lang=E

3. R Core Team. (2020). The R project for statistical computing. R. https://www.r-project.org/

4. Paul A. Hodgetts and Rohan Alexander (2020). cesR: Access the CES Datasets a Little Easier.. R package version 0.1.0.

5. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

6. Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

7. Benjamin Rich (2020). table1: Tables of Descriptive Statistics in HTML. R package version 1.2.1. https://CRAN.R-project.org/package=table1

8. Andrew McCormack and Aaron Erlich (2020). mapcan: Tools for Plotting Canadian Choropleth Maps and Choropleth Alternatives. R package version 0.0.1.

9. Ben Bolker and David Robinson (2020). broom.mixed: Tidying Methods for Mixed Models. R package version 0.2.6. https://CRAN.R-project.org/package=broom.mixed

10. Kamil Barton (2020). MuMIn: Multi-Model Inference. R package version 1.43.17. https://CRAN.R-project.org/package=MuMIn

11. Nakagawa Shinichi, Johnson Paul C. D. and Schielzeth Holger. 2017 The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded J. R. Soc. Interface. 14:20170213 http://doi.org/10.1098/rsif.2017.0213

12. Forty-third general election 2019. (n.d.). Retrieved December 20, 2020, from https://www.elections.ca/res/rep/off/ovr2019app/51/table11E.html

Appendix

Liberal Model Summary

```
## # A tibble: 19 x 6
##    effect   group              term            estimate std.error   p.value
##    <chr>    <chr>              <chr>              <dbl>     <dbl>     <dbl>
##  1 fixed    <NA>               (Intercept)     -0.961      0.0975  6.35e-23
##  2 fixed    <NA>               age20 to 24     -0.204      0.110   6.28e- 2
##  3 fixed    <NA>               age25 to 29     -0.122      0.104   2.40e- 1
##  4 fixed    <NA>               age30 to 34     -0.106      0.102   2.98e- 1
##  5 fixed    <NA>               age35 to 39     -0.00116    0.102   9.91e- 1
##  6 fixed    <NA>               age40 to 44     -0.130      0.103   2.06e- 1
##  7 fixed    <NA>               age45 to 49     -0.0625     0.103   5.43e- 1
##  8 fixed    <NA>               age50 to 54     -0.147      0.102   1.49e- 1
##  9 fixed    <NA>               age55 to 59     -0.0696     0.101   4.91e- 1
## 10 fixed    <NA>               age60 to 64      0.0261     0.100   7.95e- 1
## 11 fixed    <NA>               age65 to 69      0.134      0.101   1.84e- 1
## 12 fixed    <NA>               age70 to 74      0.163      0.104   1.17e- 1
## 13 fixed    <NA>               age75 to 79      0.132      0.117   2.61e- 1
## 14 fixed    <NA>               age80 to 84      0.0845     0.156   5.89e- 1
## 15 fixed    <NA>               age85 to 89     -0.0731     0.261   7.79e- 1
## 16 fixed    <NA>               age90 to 94      1.07       0.414   9.56e- 3
## 17 fixed    <NA>               age95 to 99     -0.140      0.266   5.98e- 1
## 18 fixed    <NA>               sexMale         -0.00659    0.0267  8.05e- 1
## 19 ran_pars constituencynumber sd__(Intercept)  0.554     NA      NA
```

R^2 Matrix:

```
##                     R2m        R2c
## theoretical 0.003479520 0.08840945
## delta       0.002393677 0.06081977
```

Conservative Model Summary

```
## # A tibble: 19 x 6
##    effect   group              term            estimate std.error   p.value
##    <chr>    <chr>              <chr>              <dbl>     <dbl>     <dbl>
##  1 fixed    <NA>               (Intercept)       -2.06     0.126   8.19e-60
##  2 fixed    <NA>               age20 to 24        0.249    0.137   6.93e- 2
##  3 fixed    <NA>               age25 to 29        0.494    0.130   1.39e- 4
##  4 fixed    <NA>               age30 to 34        0.621    0.127   1.00e- 6
##  5 fixed    <NA>               age35 to 39        0.671    0.127   1.17e- 7
##  6 fixed    <NA>               age40 to 44        0.852    0.127   1.89e-11
##  7 fixed    <NA>               age45 to 49        0.757    0.127   2.79e- 9
##  8 fixed    <NA>               age50 to 54        0.961    0.126   2.44e-14
##  9 fixed    <NA>               age55 to 59        0.864    0.125   5.44e-12
## 10 fixed    <NA>               age60 to 64        0.875    0.125   2.58e-12
## 11 fixed    <NA>               age65 to 69        0.829    0.125   3.80e-11
## 12 fixed    <NA>               age70 to 74        0.892    0.128   3.91e-12
## 13 fixed    <NA>               age75 to 79        1.13     0.138   4.39e-16
## 14 fixed    <NA>               age80 to 84        1.27     0.172   1.50e-13
## 15 fixed    <NA>               age85 to 89        1.04     0.266   8.92e- 5
## 16 fixed    <NA>               age90 to 94        0.547    0.495   2.70e- 1
## 17 fixed    <NA>               age95 to 99        0.938    0.284   9.44e- 4
## 18 fixed    <NA>               sexMale            0.497   0.0278   2.44e-71
## 19 ran_pars constituencynumber sd__(Intercept)    0.788   NA      NA
```

R^2 Matrix:

```
##                    R2m       R2c
## theoretical 0.02986077 0.1840290
## delta       0.02108326 0.1299341
```

Bloc Model Summary

```
## # A tibble: 19 x 6
##    effect   group              term            estimate std.error   p.value
##    <chr>    <chr>              <chr>              <dbl>     <dbl>     <dbl>
##  1 fixed    <NA>               (Intercept)       -7.08     0.490   2.54e-47
##  2 fixed    <NA>               age20 to 24        0.288    0.444   5.16e- 1
##  3 fixed    <NA>               age25 to 29        0.765    0.423   7.01e- 2
##  4 fixed    <NA>               age30 to 34        0.725    0.418   8.29e- 2
##  5 fixed    <NA>               age35 to 39        0.524    0.418   2.10e- 1
##  6 fixed    <NA>               age40 to 44        1.06     0.411   1.01e- 2
##  7 fixed    <NA>               age45 to 49        0.955    0.414   2.11e- 2
##  8 fixed    <NA>               age50 to 54        1.03     0.412   1.23e- 2
##  9 fixed    <NA>               age55 to 59        1.45     0.407   3.53e- 4
## 10 fixed    <NA>               age60 to 64        1.50     0.406   2.29e- 4
## 11 fixed    <NA>               age65 to 69        1.62     0.407   6.60e- 5
## 12 fixed    <NA>               age70 to 74        1.53     0.411   2.09e- 4
## 13 fixed    <NA>               age75 to 79        1.47     0.428   5.74e- 4
## 14 fixed    <NA>               age80 to 84        1.81     0.517   4.68e- 4
## 15 fixed    <NA>               age85 to 89       -0.235    1.11    8.33e- 1
## 16 fixed    <NA>               age90 to 94        3.48     1.39    1.23e- 2
## 17 fixed    <NA>               age95 to 99        0.801    0.755   2.89e- 1
## 18 fixed    <NA>               sexMale            0.215   0.0640   8.04e- 4
## 19 ran_pars constituencynumber sd__(Intercept)    4.30    NA      NA
```

R^2 Matrix:

```
##                      R2m       R2c
## theoretical 0.010058370 0.8505368
## delta       0.009190408 0.7771419
```

NDP Model Summary

```
## # A tibble: 19 x 6
##    effect   group              term            estimate std.error   p.value
##    <chr>    <chr>              <chr>              <dbl>     <dbl>     <dbl>
##  1 fixed    <NA>               (Intercept)     -1.11       0.104   2.65e-26
##  2 fixed    <NA>               age20 to 24      0.144      0.115   2.09e- 1
##  3 fixed    <NA>               age25 to 29     -0.00847    0.111   9.39e- 1
##  4 fixed    <NA>               age30 to 34     -0.319      0.110   3.85e- 3
##  5 fixed    <NA>               age35 to 39     -0.489      0.111   1.11e- 5
##  6 fixed    <NA>               age40 to 44     -0.646      0.114   1.56e- 8
##  7 fixed    <NA>               age45 to 49     -0.736      0.115   1.58e-10
##  8 fixed    <NA>               age50 to 54     -0.900      0.116   7.53e-15
##  9 fixed    <NA>               age55 to 59     -1.01       0.115   1.82e-18
## 10 fixed    <NA>               age60 to 64     -1.28       0.118   2.66e-27
## 11 fixed    <NA>               age65 to 69     -1.30       0.119   2.00e-27
## 12 fixed    <NA>               age70 to 74     -1.24       0.127   1.03e-22
## 13 fixed    <NA>               age75 to 79     -1.52       0.171   5.80e-19
## 14 fixed    <NA>               age80 to 84     -2.09       0.313   2.39e-11
## 15 fixed    <NA>               age85 to 89     -0.812      0.357   2.28e- 2
## 16 fixed    <NA>               age90 to 94     -2.21       1.03    3.21e- 2
## 17 fixed    <NA>               age95 to 99     -0.954      0.372   1.03e- 2
## 18 fixed    <NA>               sexMale         -0.352      0.0369  1.23e-21
## 19 ran_pars constituencynumber sd__(Intercept)  0.577     NA      NA
```

R^2 Matrix:

```
##                    R2m        R2c
## theoretical 0.07471936 0.15985664
## delta       0.03261770 0.06978318
```

Green Model Summary

```
## # A tibble: 19 x 6
##    effect   group              term            estimate std.error   p.value
##    <chr>    <chr>              <chr>              <dbl>     <dbl>     <dbl>
##  1 fixed    <NA>               (Intercept)       -1.71      0.122   8.65e-45
##  2 fixed    <NA>               age20 to 24       -0.195     0.141   1.66e- 1
##  3 fixed    <NA>               age25 to 29       -0.539     0.138   9.47e- 5
##  4 fixed    <NA>               age30 to 34       -0.473     0.134   4.07e- 4
##  5 fixed    <NA>               age35 to 39       -0.764     0.137   2.59e- 8
##  6 fixed    <NA>               age40 to 44       -0.875     0.141   5.34e-10
##  7 fixed    <NA>               age45 to 49       -0.792     0.140   1.48e- 8
##  8 fixed    <NA>               age50 to 54       -0.963     0.141   8.02e-12
##  9 fixed    <NA>               age55 to 59       -1.02      0.139   1.76e-13
## 10 fixed    <NA>               age60 to 64       -1.08      0.139   1.08e-14
## 11 fixed    <NA>               age65 to 69       -0.955    0.138   4.60e-12
## 12 fixed    <NA>               age70 to 74       -0.929    0.145   1.71e-10
## 13 fixed    <NA>               age75 to 79       -1.27     0.189   1.82e-11
## 14 fixed    <NA>               age80 to 84       -1.21     0.279   1.48e- 5
## 15 fixed    <NA>               age85 to 89       -1.22     0.481   1.14e- 2
## 16 fixed    <NA>               age90 to 94       -1.66     1.05    1.11e- 1
## 17 fixed    <NA>               age95 to 99       -0.889    0.447   4.67e- 2
## 18 fixed    <NA>               sexMale           -0.106    0.0454  1.99e- 2
## 19 ran_pars constituencynumber sd__(Intercept)    0.540   NA      NA
```

R^2 Matrix:

```
##                     R2m        R2c
## theoretical 0.021883004 0.10163215
## delta       0.005793048 0.02690489
```