Mini-project 2

Ankit Mathur & Nitesh Jaswal

3/17/2019

Variable definitions:

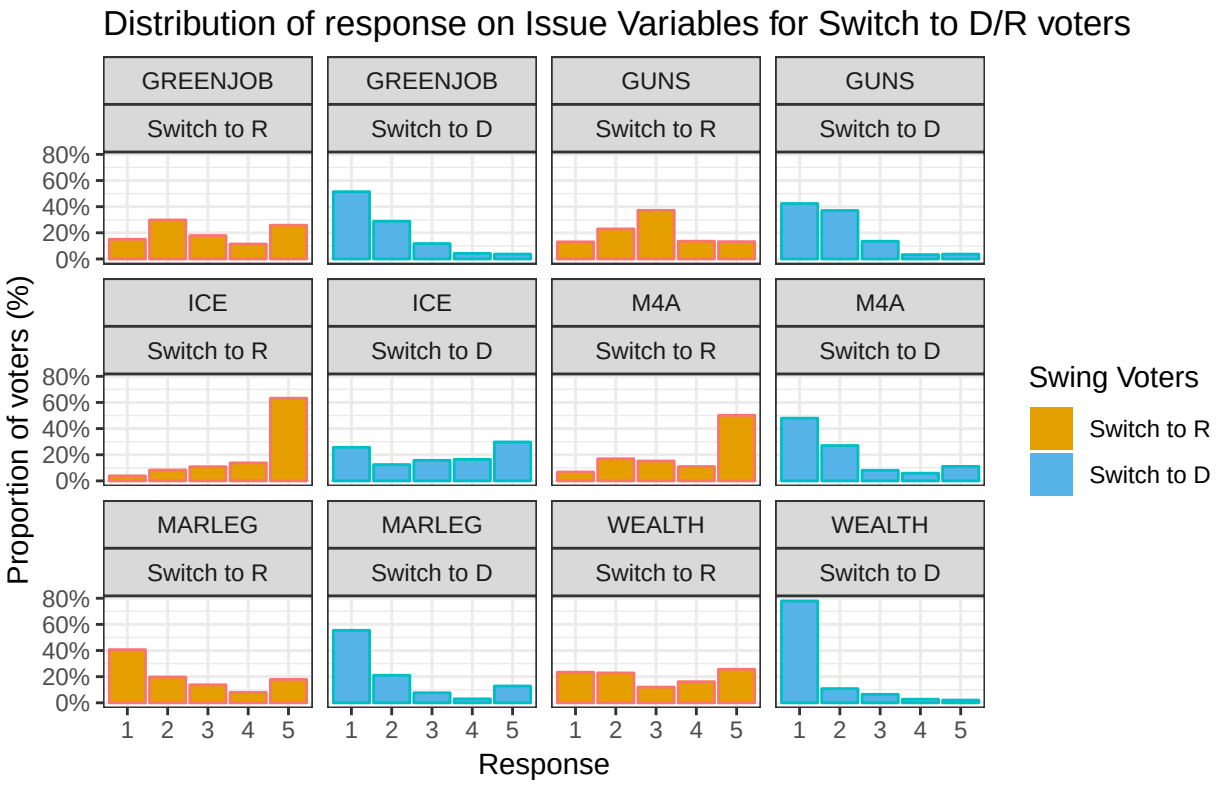
Issue variables:
Respondents were asked to give their support for the following programs on a 1–5 scale, where 1 means *strongly support* and 5 means *strongly oppose*.

- **M4A**: Medicare for All
- **GREENJOB**: A Green Jobs program
- **WEALTH**: A tax on wealth over $100 million
- **MARLEG**: Legalizing marijuana
- **ICE**: Defunding Immigration and Customs Enforcement
- **GUNS**: Gun control

Voter groups:

- **Loyal Democrats**: People who voted for Hillary Clinton in 2016 and a Democratic House candidate in 2018.
- **Loyal Republicans**: People who voted for Donald Trump in 2016 and a Republican House candidate in 2018.
- **Swing voters**: All other people who voted in 2018. In addition, define the following two subsets of swing voters:
- **Switch to D**: People who didn't vote for Hillary Clinton in 2016 but voted for a Democratic House candidate in 2018.
- **Switch to R**: People who didn't vote for Donald Trump in 2016 but voted for a Republican House candidate in 2018.

Q1: How do Switch to D and Switch to R voters differ on the issue variables?



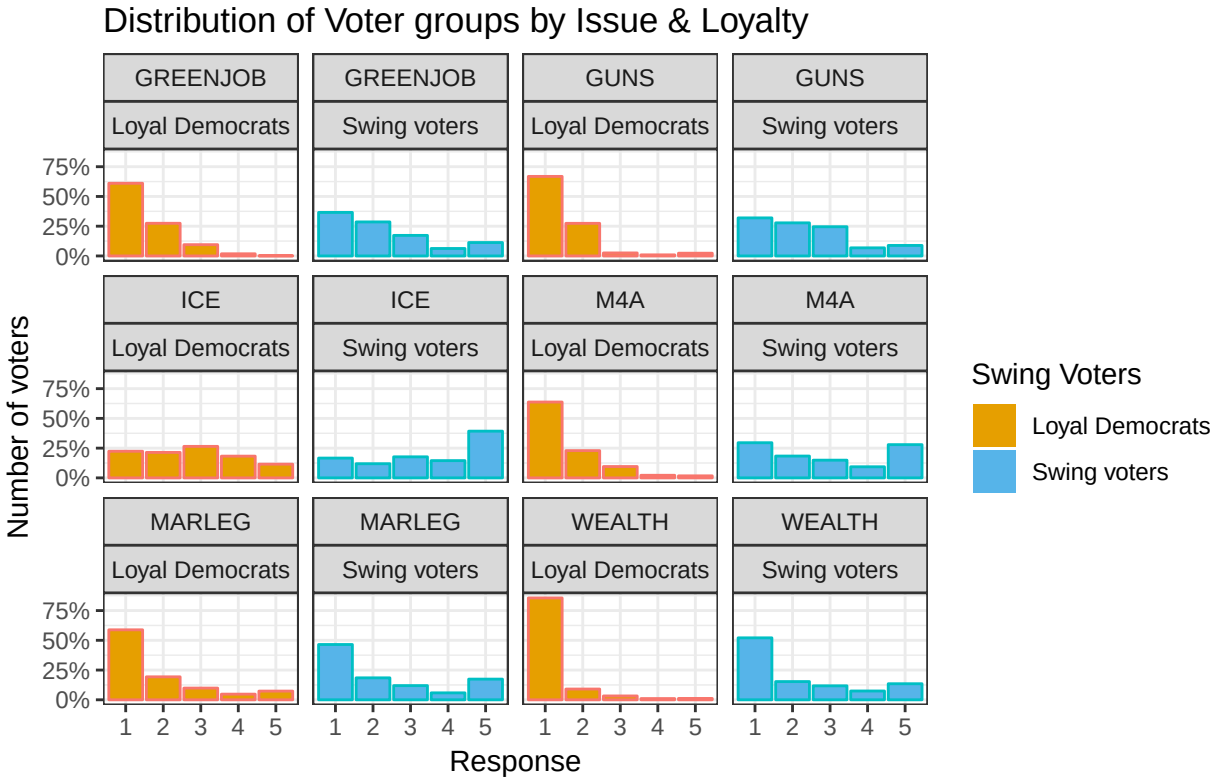
Response Values: 1. Strongly support 2. Somewhat support 3. Neither support nor oppose 4. Somewhat oppose 5. Strongly oppose

From the plot above, it can be observed that the **Switch to R** and **Switch to D** voters differ strongly on the Medicare for all (M4A) issue. While a majority of **Switch to R** voters seem to strongly oppose this issue, the **Switch to D** voters tend to support it.

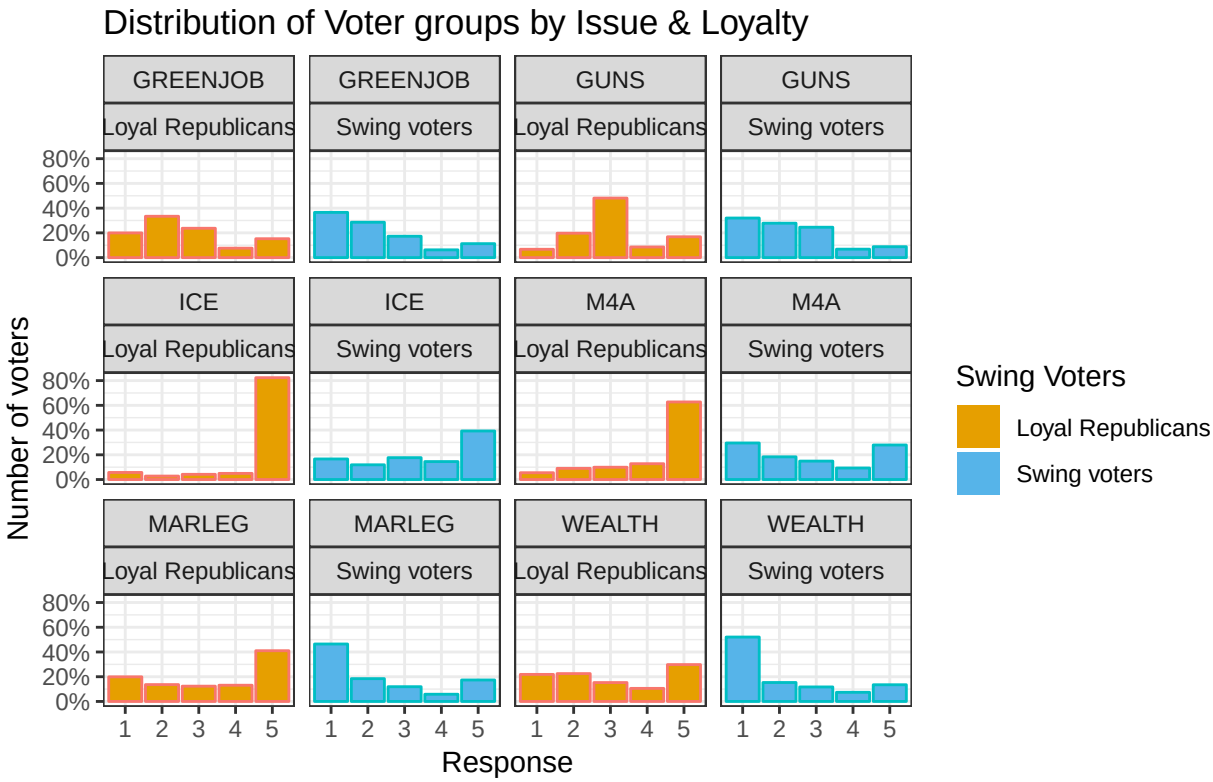
Additional insights:

- The **Switch to D** voters have a strong opinion on all issues except the immigration issue (ICE). On the other hand, the **Switch to R** voters have a strong opinion on issues such as ICE, MARLEG and M4A.
- The **Switch to R** and **Switch to D** voters behave rather similarly on the issue of marijuana legalization (MARLEG) with both the cohorts showing support towards the legalization.
- It is interesting to note that while the **Switch to R** voter group does not seem to have a strong opinion on multiple issues (GREENJOB, GUNS, and WEALTH), the **Switch to D** cohort does demonstrate a strong opinion on all issues except immigration (ICE).

Q2: How do swing voters differ from loyal Democrats and loyal Republicans on the issue variables?



Values: 1. Strongly support 2. Somewhat support 3. Neither support nor oppose
4. Somewhat oppose 5. Strongly oppose



Values: 1. Strongly support 2. Somewhat support 3. Neither support nor oppose
4. Somewhat oppose 5. Strongly oppose

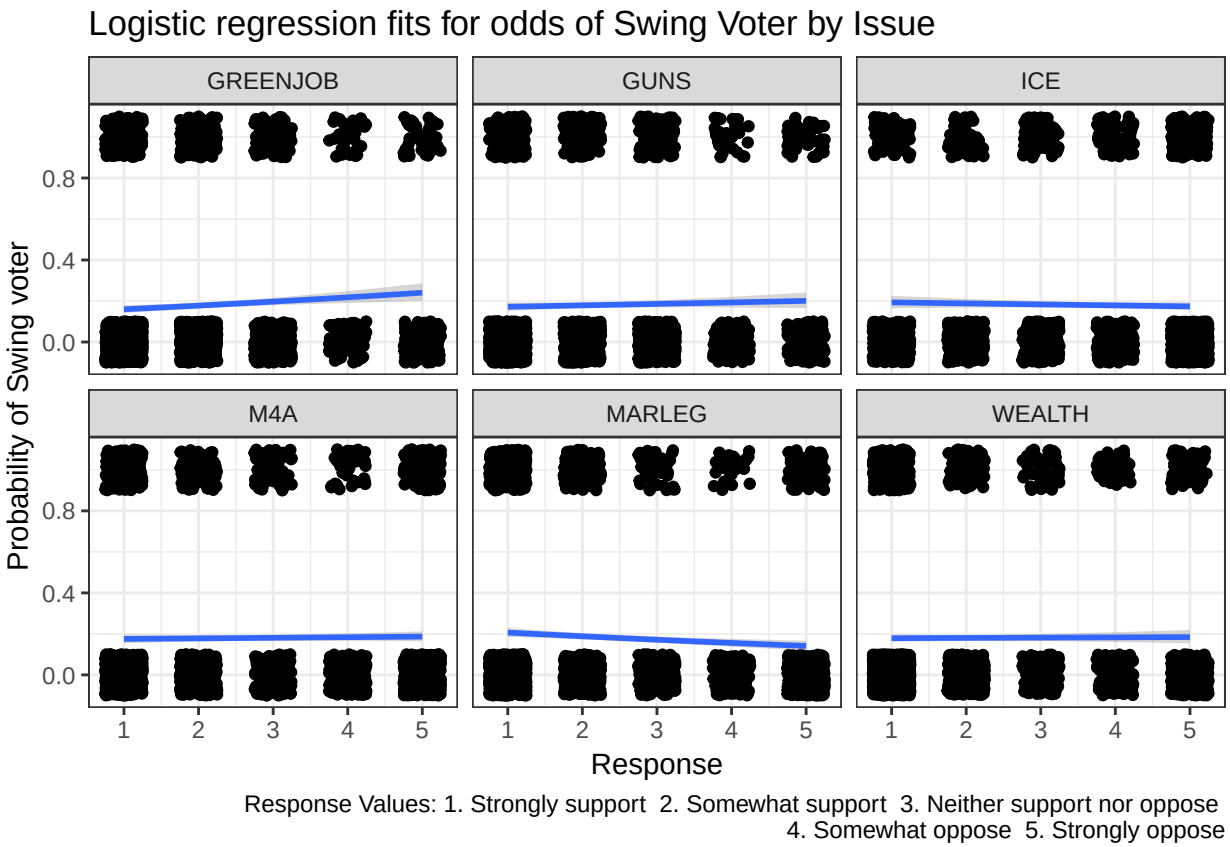
- With the exception of immigration issue (ICE), **Swing voters** tend to share *similar* opinions with **Loyal Democrats** across all of the remaining five issues. Interestingly, both these cohorts voted in favor of all five issues with a majority of voters choosing the *Strongly support* option.

- On the immigration issue, while **Loyal Democrats** do not seem to have any coherent opinion, both **Loyal Republicans** and **Swing voters** cohorts voted against the defunding of ICE with a majority of them choosing the *Strongly oppose* option.

Additional insights:

- An equivalent number of **Swing voters** chose *Strongly support* and *Strongly oppose* on the Medicare for all (M4A) issue, thereby showcasing a slightly different opinion than the **Loyal Democrats**.

Q3: What predicts being a swing voter?



The plots above showcase how the predicted probability of a registered voter being a swing voter varies with responses to each of the six issues. With the exception of GREENJOB and MARLEG, all the other fitted lines are fairly flat suggesting that we include just these two issues in our final model.
However, just to be sure, we did explicitly fit a logistic regression model including all six issue variables without interaction and looked at the coefficients for each of these issues (see appendix). As expected, the coefficients for all issues, except GREENJOB and MARLEG, turned out to be close to zero.
Hence, in our final model, we will focus only on these two issues. Furthermore, for the sake of simplicity more than anything else, we will avoid including any interactions between our predictors.
Finally, let us fit a logistic regression model to predict swing voters using GREENJOB and MARLEG issues:

```
## glm(formula = swing_voter ~ GREENJOB + MARLEG, family = "quasibinomial",
##     data = WTHH.issue, weights = weight_DFP)
```

```
##             coef.est coef.se
## (Intercept) -1.34     0.11
## GREENJOB     0.21     0.04
## MARLEG      -0.20     0.03
## ---
##   n = 2648, k = 3
##   residual deviance = 2589.1, null deviance = 2635.7 (difference = 46.5)
##   overdispersion parameter = 1.0
```

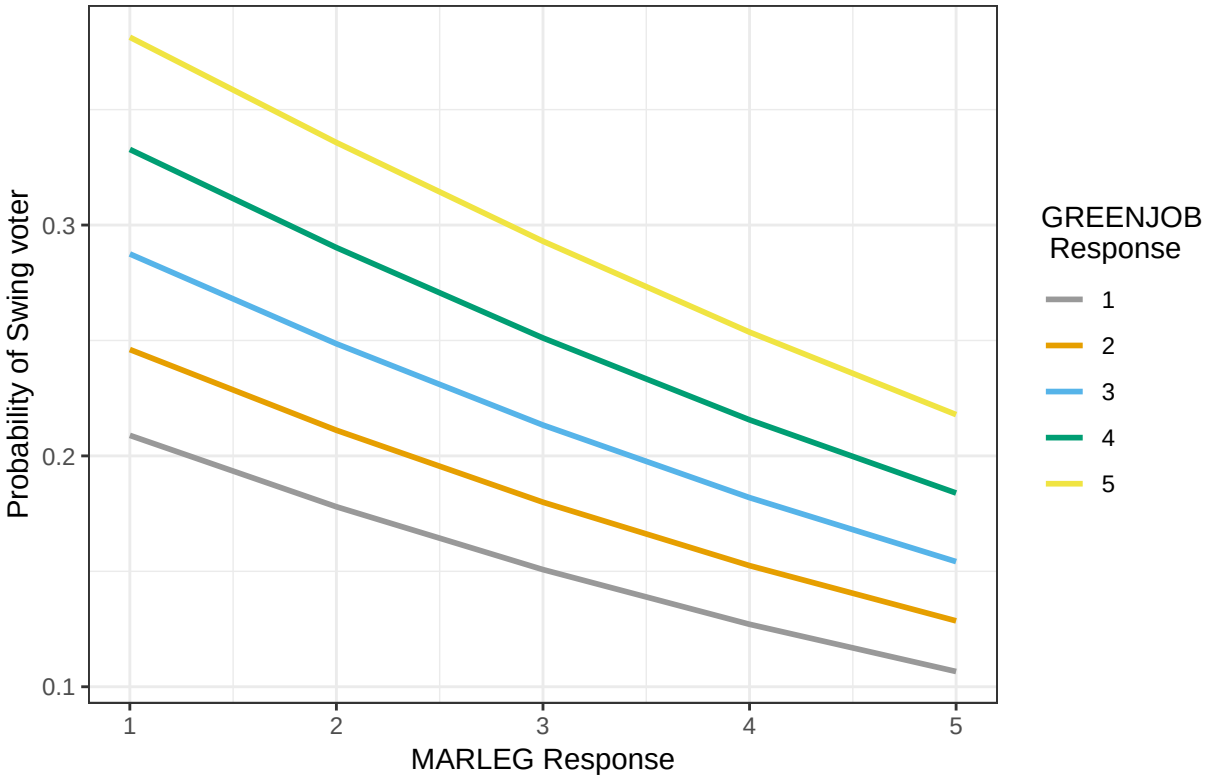
Interpreting the model above, we get:

$$\text{logit}[\text{P(Swing voter)}] = -1.34 + 0.21 \times \text{GREENJOB} - 0.20 \times \text{MARLEG}$$

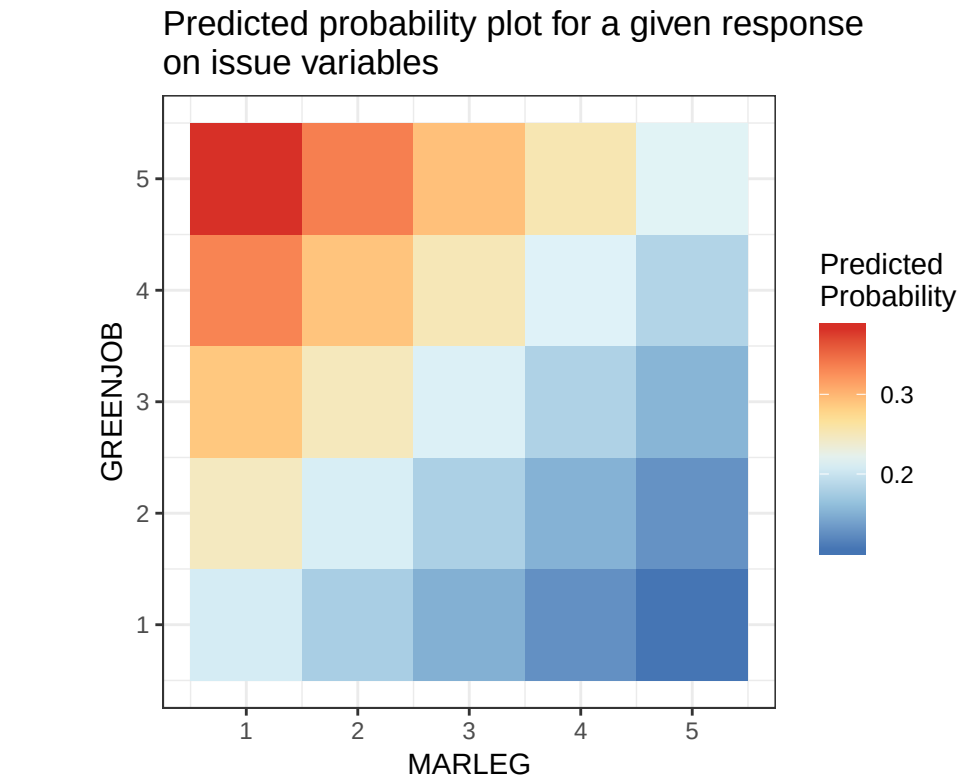
Using the "divide by 4" rule, we can say that:
• As per our model, keeping the opinion on MARLEG constant and moving one level closer to opposing the GREENJOB issue *increases* the probability of swing voter by 5.25%
• Similarly, keeping the opinion on GREENJOB constant and moving one level closer to opposing the MARLEG issue *decreases* the model probabilty by 5%

Looking at the residual deviance, we can say that this model is not good enough to predict the probability of a voter being a swing voter as it is able to explain only about $\frac{46.5}{2635.7} = 1.8\%$. However, in our opinion, this is the best we can do if we are to use just the issue variables for prediction because by including any other interactions or other issue variables, we ae just chasing noise rather than actually understanding the true picture here.

Now, let's visualize this model:

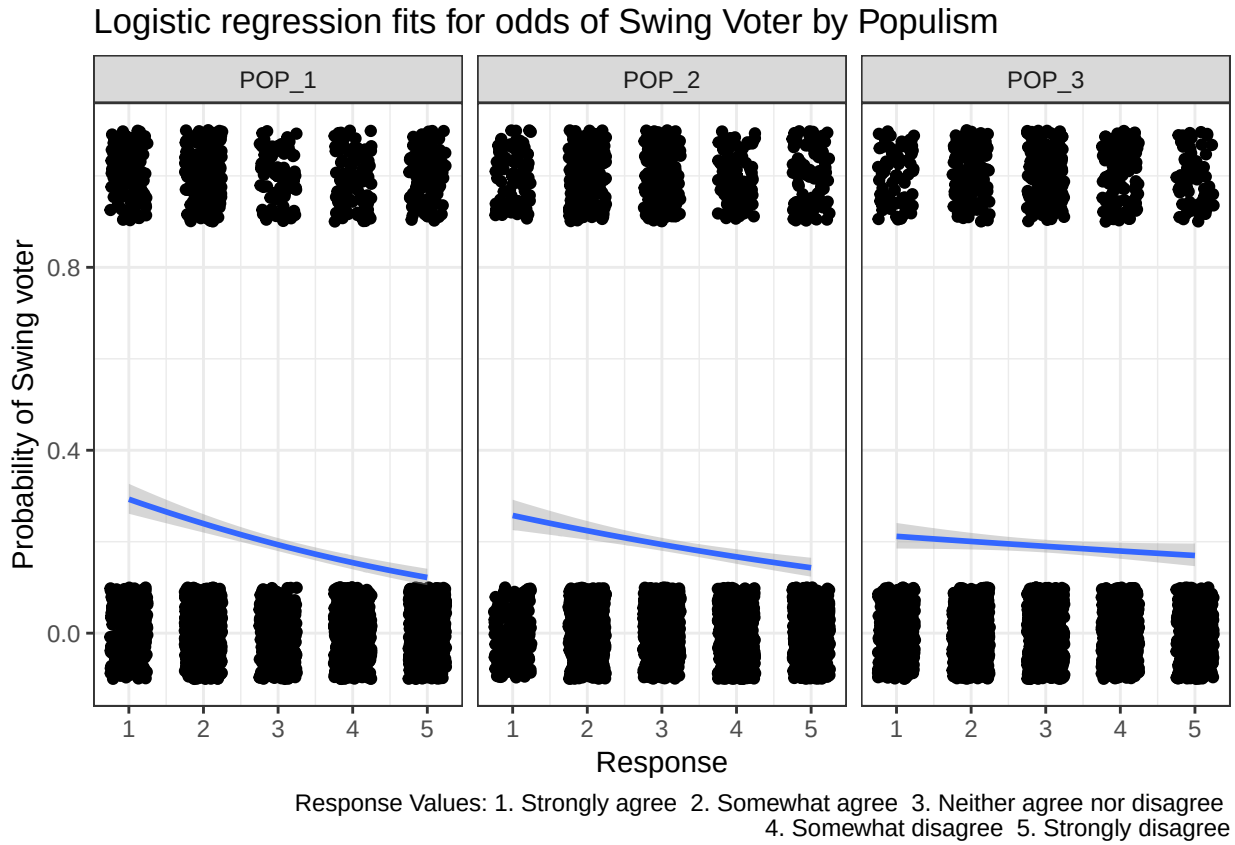


Response Values: 1. Strongly support 2. Somewhat support 3. Neither support nor oppose
4. Somewhat oppose 5. Strongly oppose



In the above plots, it can be clearly visualized that keeping the opinion on GREENJOB constant and moving towards opposing the MARLEG issue *decreases* the model probabilty of being a swing voter. Furthermore, for a given opinion on MARLEG, the more a voter opposes GREENJOB, higher are the chances of him/her being a swing voter.

Moving on to *populism* variables, let's do a faceted plot (as before) to get a feel of how the odds of swing voter vary with each populism variable.



The slope of the fitted lines decrease as we move from *POP_1* to *POP_2* ultimately flattening out as we reach *POP_3*. As before, just for our satisfaction, we looked at the slopes for each of these populism variables by explicitly fitting a logistic regression model for predicting the swing voters and observed that the coefficients for *POP_2* and *POP_3* variables turned out to be close to zero (see appendix).
Hence, let us fit our second model based on just the *POP_1* populism variable:

```
## glm(formula = swing_voter ~ POP_1, family = "quasibinomial",
##     data = WTHH.pop, weights = weight_DFP)
##             coef.est coef.se
## (Intercept) -0.54     0.11
## POP_1       -0.25     0.03
## ---
##   n = 3049, k = 2
##   residual deviance = 3100.3, null deviance = 3159.5 (difference = 59.2)
##   overdispersion parameter = 1.0
```

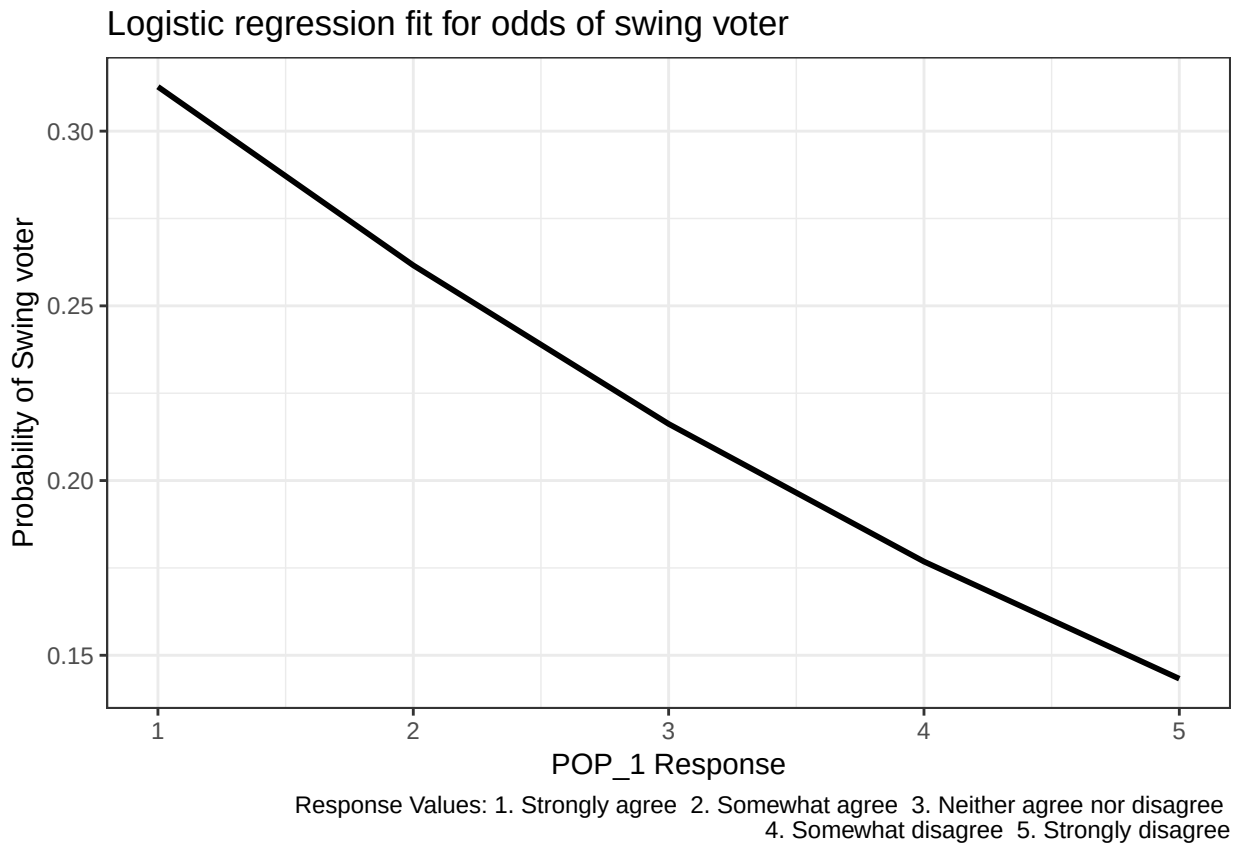
We interpret the coefficients above as representing the odds of a registered voter being a swing voter as below:

$$\text{logit}[\text{P}(\text{Swing voter})] = -0.54 - 0.25 \times POP_1$$

Again, as per the "divide by 4" rule, we can say that an increment of one degree of disagreement with the sentiment shared by *POP_1* variable decreases the probability of swing voter by 6.25%.

Looking at the difference between residual deviance and the null deviance, we see that just POP_1 explains $\frac{59.2}{3159.5} = 1.9\%$ variation in our response variable. Furthermore, adding other populism variables (or their interactions) does not help explaining much of the residual deviance here. Hence, we'll go ahead with this model for now.

Visualizing this model we get:



Training set classification error

```
## [1] "Classification Error = 20.275776"
```

We drew the decision boundary at the probability of 0.13 and classified all inputs whose probability was greater than this as swing voters. The classification error for this model turned out to be 20.27578% which is slightly worse than simply guessing all the voters to not be swing voters.
Since the distribution of our data is skewed such that there are very few swing voters in our data, accuracy is not a good metric in this case.

Appendix

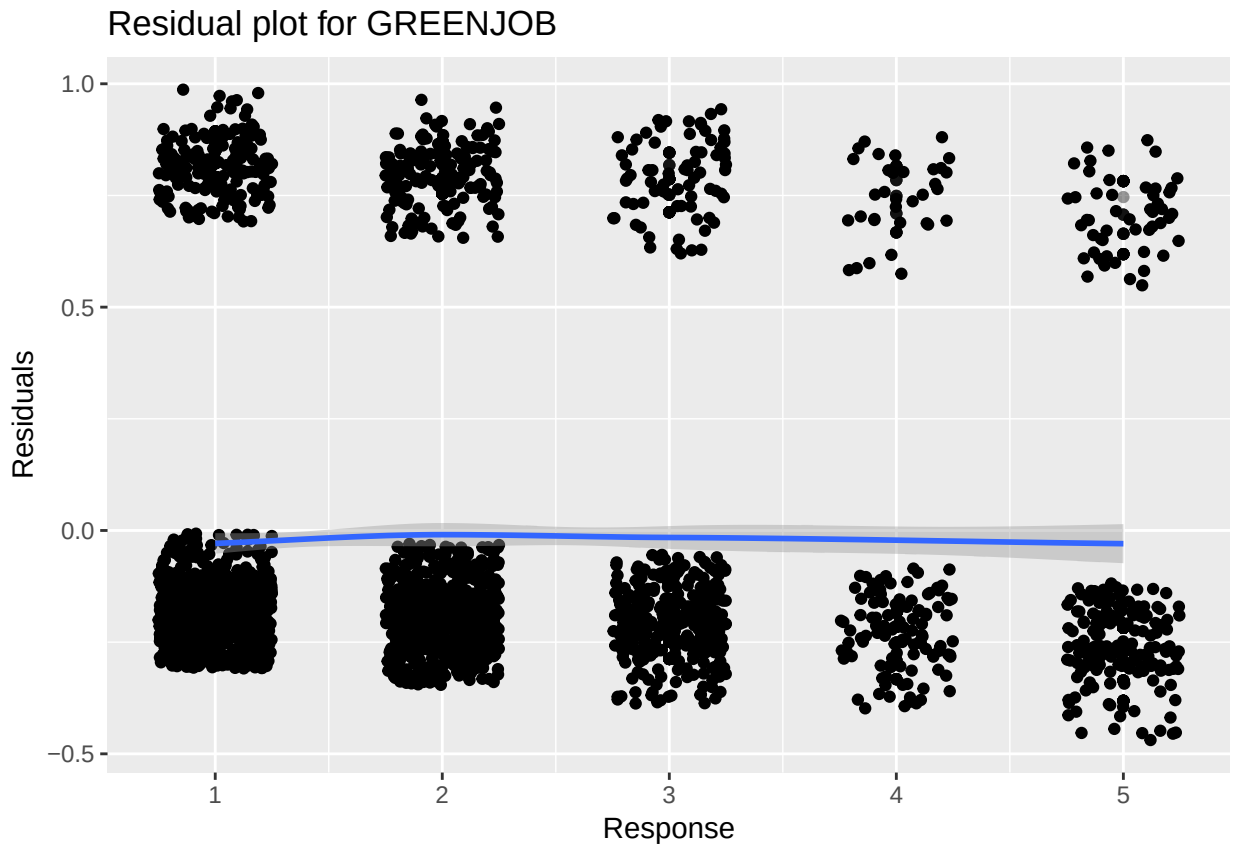
```
## glm(formula = swing_voter ~ GREENJOB + GUNS + ICE + M4A + MARLEG +
##     WEALTH, family = "quasibinomial", data = WTHH.issue, weights = weight_DFP)
##             coef.est coef.se
## (Intercept) -1.25     0.14
## GREENJOB     0.19     0.05
## GUNS         0.05     0.05
## ICE         -0.09     0.04
## M4A          0.06     0.05
## MARLEG      -0.20     0.04
## WEALTH      -0.01     0.05
## ---
##   n = 2648, k = 7
##   residual deviance = 2582.5, null deviance = 2635.7 (difference = 53.1)
##   overdispersion parameter = 1.0
```

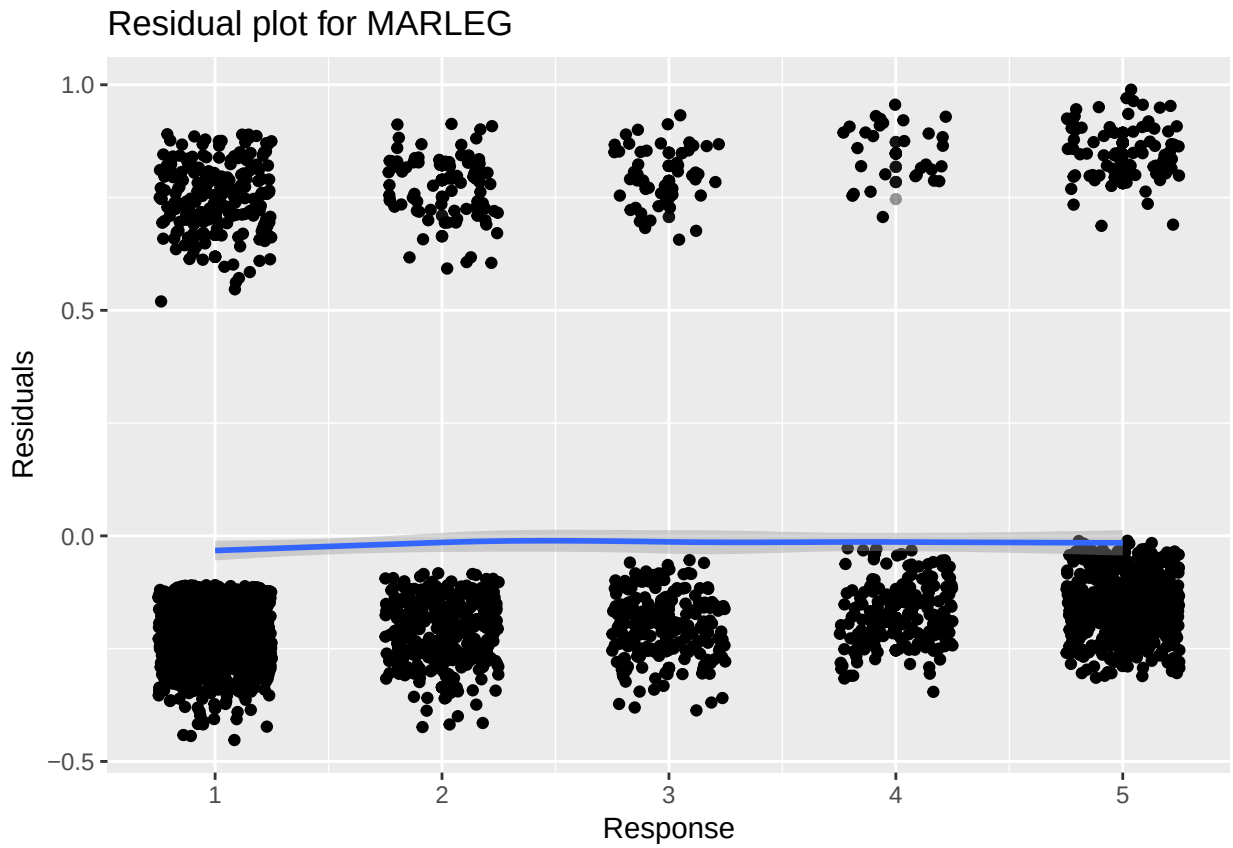
```
## glm(formula = swing_voter ~ GREENJOB * GUNS * ICE * M4A * MARLEG *
##     WEALTH, family = "quasibinomial", data = WTHH.issue, weights = weight_DFP)
##                                    coef.est coef.se
## (Intercept)                         0.34     4.19
## GREENJOB                           -3.36     2.51
## GUNS                               -1.35     2.00
## ICE                                -1.54     1.13
## M4A                                 1.82     2.11
## MARLEG                             -0.70     2.05
## WEALTH                             -2.83     3.00
## GREENJOB:GUNS                       1.63     1.19
## GREENJOB:ICE                        1.11     0.62
## GUNS:ICE                            0.75     0.52
## GREENJOB:M4A                        0.36     1.02
## GUNS:M4A                           -0.57     0.89
## ICE:M4A                             0.11     0.49
## GREENJOB:MARLEG                     0.72     1.08
## GUNS:MARLEG                         1.07     1.14
## ICE:MARLEG                          0.30     0.55
## M4A:MARLEG                         -0.25     0.75
## GREENJOB:WEALTH                     3.01     1.59
## GUNS:WEALTH                         1.42     1.24
## ICE:WEALTH                          1.02     0.76
## M4A:WEALTH                         -0.63     1.33
## MARLEG:WEALTH                       1.35     1.34
## GREENJOB:GUNS:ICE                  -0.50     0.28
## GREENJOB:GUNS:M4A                  -0.24     0.42
## GREENJOB:ICE:M4A                   -0.24     0.24
## GUNS:ICE:M4A                       -0.06     0.20
## GREENJOB:GUNS:MARLEG               -0.73     0.56
## GREENJOB:ICE:MARLEG                -0.23     0.27
## GUNS:ICE:MARLEG                    -0.34     0.28
## GREENJOB:M4A:MARLEG                -0.07     0.37
## GUNS:M4A:MARLEG                    -0.07     0.35
## ICE:M4A:MARLEG                     -0.04     0.18
## GREENJOB:GUNS:WEALTH               -1.17     0.70
## GREENJOB:ICE:WEALTH                -0.76     0.36
## GUNS:ICE:WEALTH                    -0.41     0.31
```

```
## GREENJOB:M4A:WEALTH                -0.40     0.47
## GUNS:M4A:WEALTH                     0.27     0.48
## ICE:M4A:WEALTH                     -0.05     0.29
## GREENJOB:MARLEG:WEALTH             -0.99     0.58
## GUNS:MARLEG:WEALTH                 -1.07     0.65
## ICE:MARLEG:WEALTH                  -0.37     0.34
## M4A:MARLEG:WEALTH                   0.06     0.44
## GREENJOB:GUNS:ICE:M4A               0.12     0.09
## GREENJOB:GUNS:ICE:MARLEG            0.21     0.13
## GREENJOB:GUNS:M4A:MARLEG            0.12     0.16
## GREENJOB:ICE:M4A:MARLEG             0.06     0.08
## GUNS:ICE:M4A:MARLEG                 0.06     0.08
## GREENJOB:GUNS:ICE:WEALTH            0.29     0.16
## GREENJOB:GUNS:M4A:WEALTH            0.16     0.17
## GREENJOB:ICE:M4A:WEALTH             0.14     0.10
## GUNS:ICE:M4A:WEALTH                 0.00     0.11
## GREENJOB:GUNS:MARLEG:WEALTH         0.65     0.29
## GREENJOB:ICE:MARLEG:WEALTH          0.26     0.14
## GUNS:ICE:MARLEG:WEALTH              0.26     0.16
## GREENJOB:M4A:MARLEG:WEALTH          0.11     0.17
## GUNS:M4A:MARLEG:WEALTH              0.08     0.18
## ICE:M4A:MARLEG:WEALTH               0.03     0.10
## GREENJOB:GUNS:ICE:M4A:MARLEG       -0.05     0.04
## GREENJOB:GUNS:ICE:M4A:WEALTH       -0.05     0.04
## GREENJOB:GUNS:ICE:MARLEG:WEALTH    -0.16     0.07
## GREENJOB:GUNS:M4A:MARLEG:WEALTH    -0.10     0.07
## GREENJOB:ICE:M4A:MARLEG:WEALTH     -0.04     0.04
## GUNS:ICE:M4A:MARLEG:WEALTH         -0.04     0.04
## GREENJOB:GUNS:ICE:M4A:MARLEG:WEALTH  0.03     0.01
## ---
##   n = 2648, k = 64
##   residual deviance = 2452.0, null deviance = 2635.7 (difference = 183.7)
##   overdispersion parameter = 1.0
```

```
## [1] 0.1805136
```

```
##    0    1
## 2170  478
```

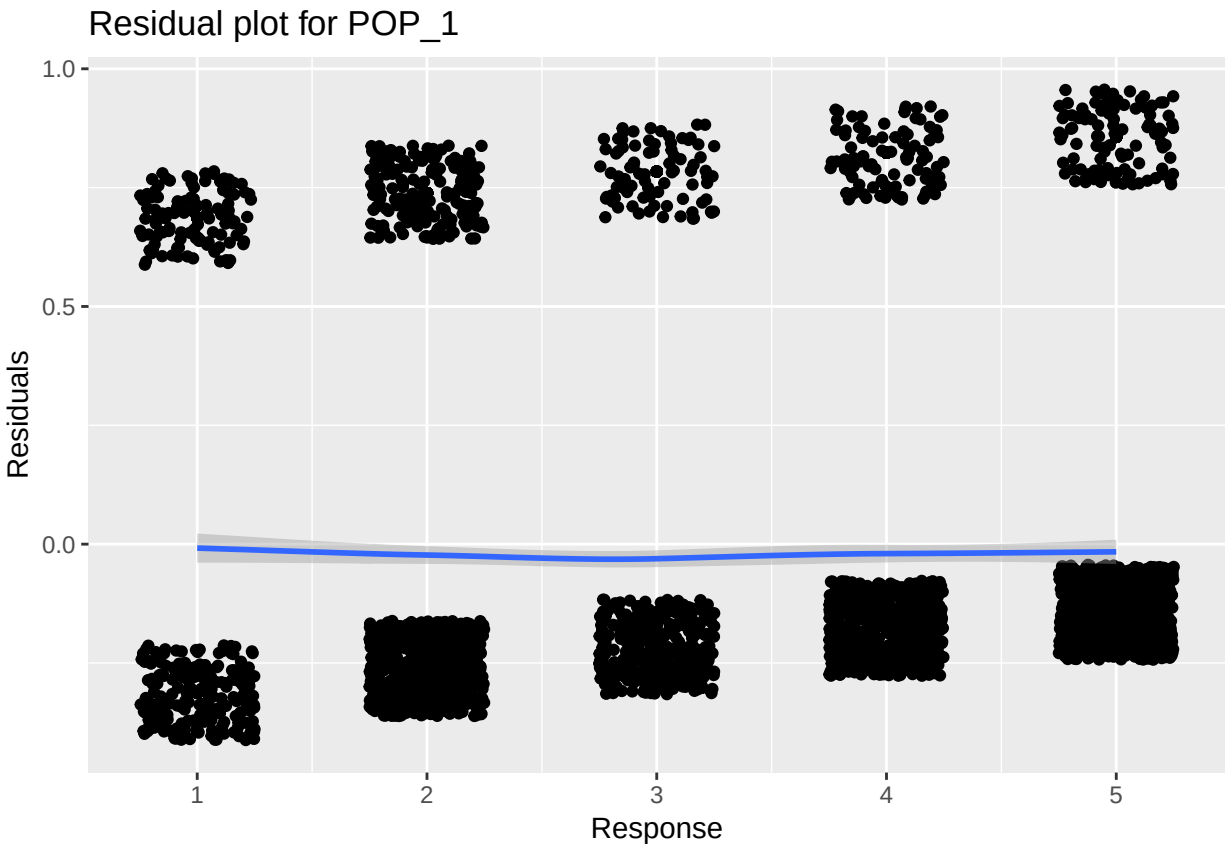

Residual plot for GREENJOB
Residuals
1.0
0.5
0.0
−0.5
1
2
3
4
5
Response




```
## glm(formula = swing_voter ~ POP_1 * POP_2 * POP_3, family = "quasibinomial",
##     data = WTHH.pop, weights = weight_DFP)
##                   coef.est coef.se
## (Intercept)       -0.62     0.57
## POP_1             -0.33     0.21
## POP_2              0.13     0.19
## POP_3              0.19     0.19
## POP_1:POP_2       -0.01     0.06
## POP_1:POP_3        0.00     0.06
## POP_2:POP_3       -0.11     0.06
## POP_1:POP_2:POP_3  0.02     0.02
## ---
##   n = 3049, k = 8
##   residual deviance = 3089.0, null deviance = 3159.5 (difference = 70.4)
##   overdispersion parameter = 1.0
```

```
## [1] 0.1902263
```

```
##    0    1
## 2469  580
```


Residual plot for POP_1
1.0
0.5
0.0
Residuals
1
2
3
4
5
Response