DATA606 - Data Project

Adam Douglas

12/9/2018

Goal

The goal of this project is to investigate whether a correlation exists between the types of media typically consumed by a voter (or the sources of such media), and their attitude towards the two major party candidates of the 2016 United States presidential election.

Data Collection

My data was courtesy of the American National Election Studies (ANES), a survey conducted by is a collaboration between Stanford University and the University of Michigan, with funding by the National Science Foundation[1].

I loaded the data from a pipe-delimited file with the assistance of the "readr" package, part of the Hadley Wickham Tidyverse[2] series of packages.

The raw data had 1,290 variables in total, well more than one could reasonably work with. So, I wanted to narrow our data down to some of the most interesting variables that can help reach the stated goal.

By using the provided user guide and codebook (also on the website[3]), I was able to select some appropriate variables into a new (and smaller) data frame. The variables I selected are:

Variable	Description
`ID`	Unique identifier for the survey respondent
`surveyMethod`	Defines if the respondent did only a pre-election survey or a pre and post-election survey
`sex`	Observational gender variable
`state`	Current state of residence for the respondent
`payAttPol`	How often the respondent pays attention to politics
`TVNews`	Respondent heard about the campaign via television news
`newspapers`	Respondent heard about the campaign via newspapers
`TVTalk`	Respondent heard about the campaign via television talk shows
`internet`	Respondent heard about the campaign via internet web sites
`radio`	Respondent heard about the campaign via television news
`none`	Respondent heard about the campaign via none of these methods
`yahoo`	Respondent frequents Yahoo.com for news about the campaign
`CNN`	Respondent frequents CNN.com for news about the campaign
`NBC`	Respondent frequents NBC.com for news about the campaign
`huff`	Respondent frequents huffingtonpost.com for news about the campaign
`CBS`	Respondent frequents cbsnews.com for news about the campaign
`USA`	Respondent frequents usatoday.com for news about the campaign
`nyt`	Respondent frequents newyorktimes.com for news about the campaign
`FOX`	Respondent frequents foxnews.com for news about the campaign

[1]The American National Election Studies (www.electionstudies.org).
[2]https://www.tidyverse.org
[3]The American National Election Studies (www.electionstudies.org).

Variable	Description
`wapo`	Respondent frequents washingtonpost.com for news about the campaign
`BBC`	Respondent frequents bbcnews.com for news about the campaign
`guardian`	Respondent frequents theguardian.com for news about the campaign
`ABC`	Respondent frequents abcnews.com for news about the campaign
`other`	Respondent frequents some other website for news about the campaign
`clinton`	Respondent's opinion of candidate Hillary Clinton (0-100)
`trump`	Respondent's opinion of candidate Donald Trump (0-100)
`bothCandidates`	Derived variable. Variable `trump` minus variable `clinton`

Exploration

First I looked at the distribution of scores for the candidates, since these will be the variables we're trying to find correlations to.

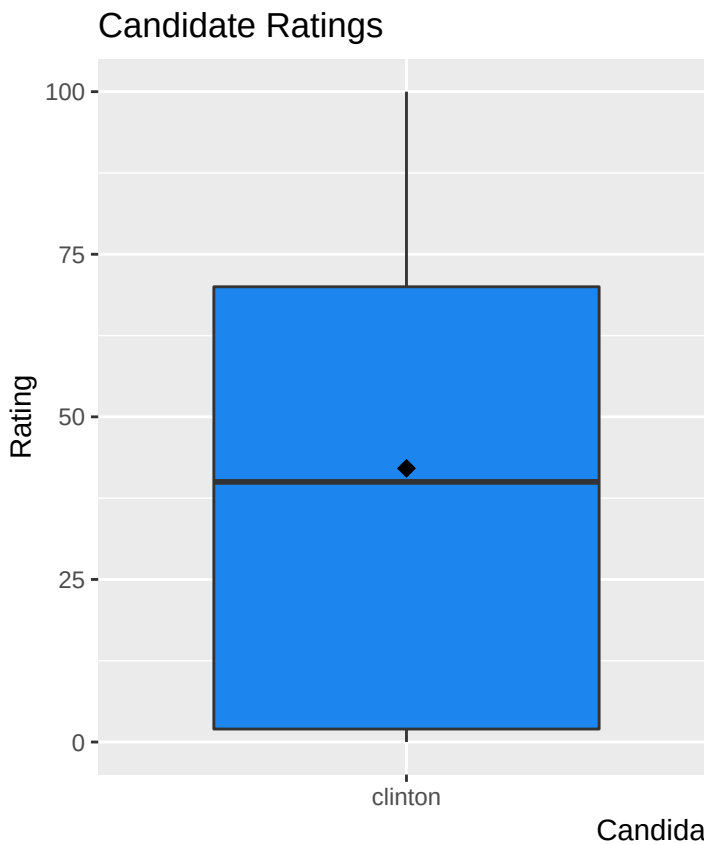
```
## Trump:
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    0.00   30.00   36.95   70.00  100.00      41
```

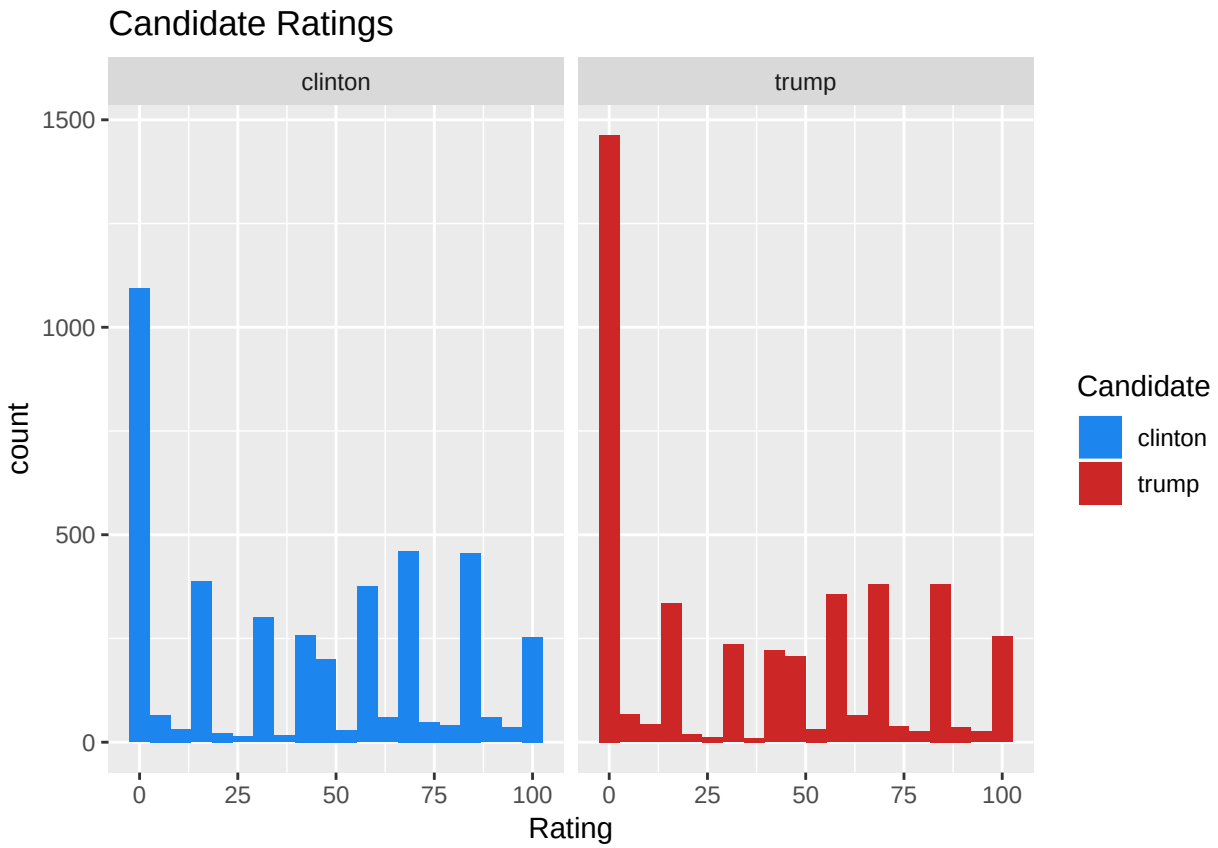
```
## Clinton:
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    2.00   40.00   42.15   70.00  100.00      38
```

The summaries show pretty similar distributions. However, there is a difference in median and mean values, suggesting that Mr. Trump did not enjoy as high popularity scores amongst respondents as his opponent did.

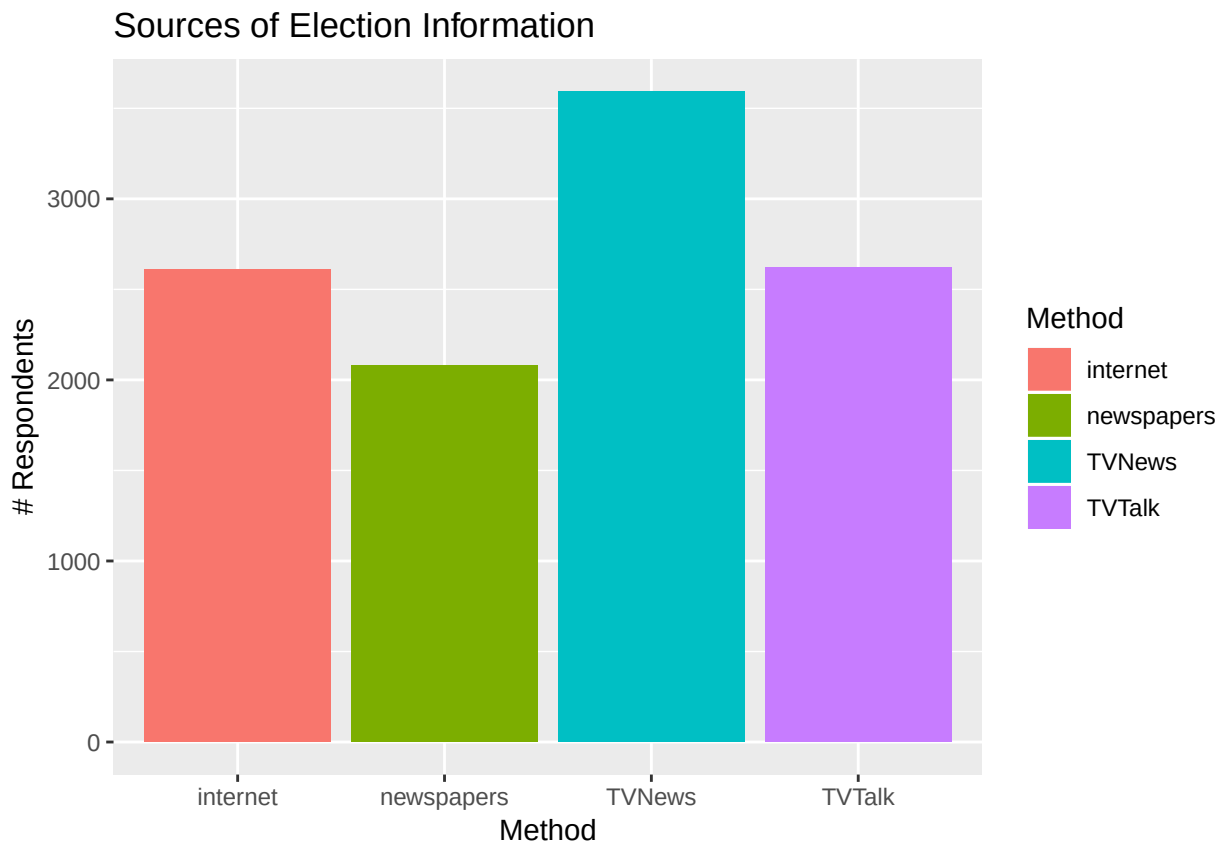


This is slightly more evident with the help of a few visualizations:



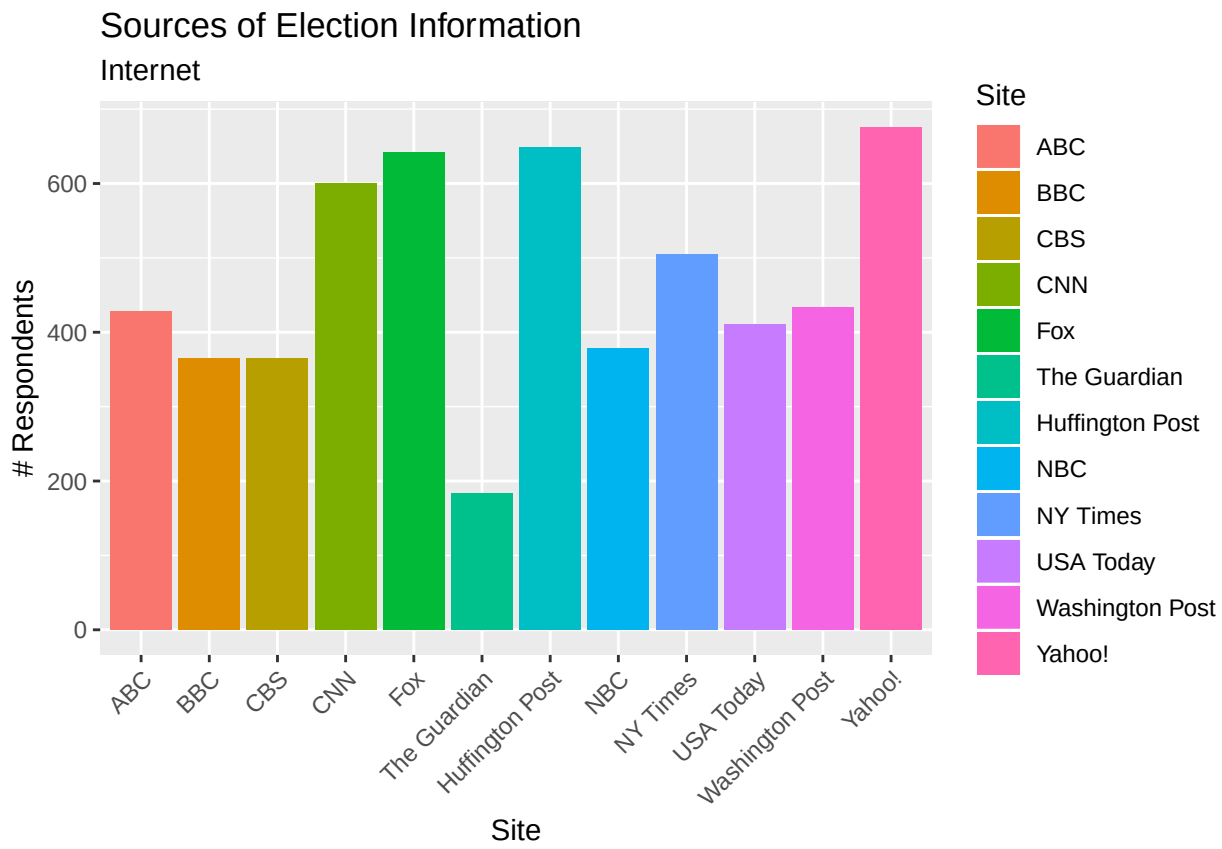
The election has been characterized in the media as one where people didn't vote *for* a candidate they liked, but rather voted *against* the one they did not like. The distribution of scores seems to support that theory to some degree, judging by the large number of 0 scores for each candidate.

Turning to the explanatory variables, I looked next at how the respondents to the survey got their election news:



Sources seem to have been evenly spread across all 4 sources.

Because of the rise in the importance of the internet as a source of news, I wanted to focus on web sources. This graph shows the breakouts of which sources of election news were visited by respondents:



Interestingly, a lot of respondents use Yahoo! news. Admittedly, I was surprised to see that result, but less surprised to see CNN, Fox, and Huffington Post as frequently-visited sites for election news.

Analysis

Next I began looking at building linear regression models to see how well correlated these news sources were to respondents' ratings of the candidates.

The dependent variable I used was `bothCandidates`, which is simply the rating of `trump` minus the rating of `clinton`. So, a score of < 0 represents preference to candidate Clinton while > 0 represents preference to candidate Trump.

The first model used variables representing the source of election news: `TVNews`, `newspapers`, `TVTalk`, and `internet`.

```
##
## Call:
## lm(formula = bothCandidates ~ TVNews + newspapers + TVTalk +
##     internet, data = survey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -101.414  -59.958   -1.414   60.042  115.281
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -5.932      2.847  -2.084 0.037222 *
## TVNewsYes         5.351      2.916   1.835 0.066608 .
## newspapersYes    -7.892      2.123  -3.717 0.000204 ***
## TVTalkYes         1.996      2.247   0.888 0.374434
## internetYes      -1.456      2.110  -0.690 0.490091
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.98 on 4142 degrees of freedom
##   (124 observations deleted due to missingness)
## Multiple R-squared:  0.004033,   Adjusted R-squared:  0.003071
## F-statistic: 4.193 on 4 and 4142 DF,  p-value: 0.002169
```

Looking at the above summary, this model does very little to show any linear correlation between these variables and candidate preference. Even when I removed more variables (*not shown here*), the adjusted R-squared never rises above 0.01.

I surmised that maybe the delivery format of the news was less important than the *source* of the information. So, I took the website variables and built a full linear regression model:

```
##
## Call:
## lm(formula = bothCandidates ~ yahoo + CNN + NBC + huff + CBS +
##     USA + NYT + FOX + wapo + BBC + guardian + ABC, data = survey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -146.418  -46.225   -1.135   45.955  155.888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.9554     1.6323  -0.585   0.5584
## yahooYes      6.5244     2.6121   2.498   0.0126 *
## CNNYes      -12.4750     2.8755  -4.338 1.49e-05 ***
## NBCYes       -4.9560     3.6848  -1.345   0.1788
## huffYes     -16.0214     2.9693  -5.396 7.45e-08 ***
## CBSYes       -5.8724     3.8776  -1.514   0.1300
## USAYes       -2.0457     3.4125  -0.599   0.5489
## NYTYes      -27.7628     3.4267  -8.102 8.28e-16 ***
## FOXYes       44.8493     2.7649  16.221  < 2e-16 ***
## wapoYes      -5.5774     3.6631  -1.523   0.1280
## BBCYes       -6.0684     3.5608  -1.704   0.0885 .
## guardianYes   0.3794     4.8536   0.078   0.9377
## ABCYes      -10.5132     3.6185  -2.905   0.0037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.98 on 2561 degrees of freedom
##   (1697 observations deleted due to missingness)
## Multiple R-squared:  0.187,  Adjusted R-squared:  0.1832
## F-statistic:  49.1 on 12 and 2561 DF,  p-value: < 2.2e-16
```

Looking at the adjusted R-squared, this model is a bit more usable than the first one. Since some of the parameter estimates appear to not be statistically significant than others, I decided to remove one variable at a time and see how that impacts the model.

Below is the best model based on adjusted R-squared (*interim models not shown*):

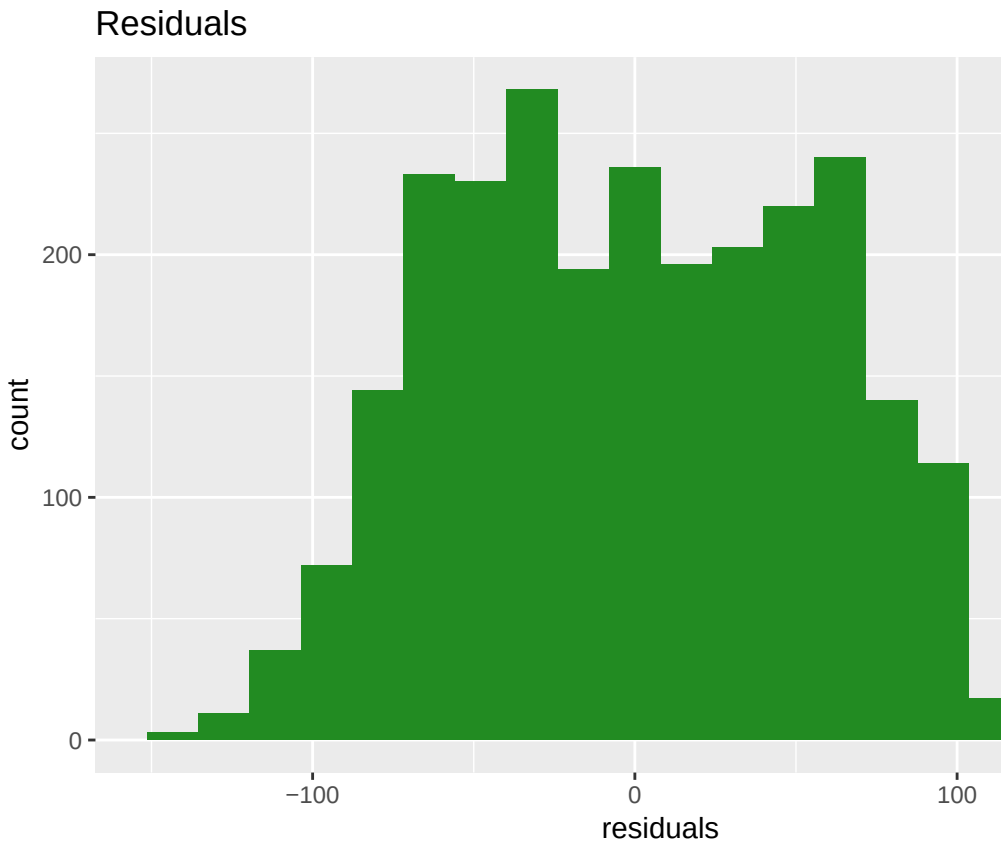
```
##
## Call:
## lm(formula = bothCandidates ~ yahoo + CNN + NBC + huff + CBS +
##     NYT + FOX + wapo + BBC + ABC, data = survey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -146.113  -46.457   -0.939   45.990  156.785
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.9904     1.6308  -0.607  0.54371
## yahooYes      6.3993     2.6013   2.460  0.01396 *
## CNNYes      -12.5241     2.8732  -4.359 1.36e-05 ***
## NBCYes       -5.1424     3.6678  -1.402  0.16102
## huffYes     -16.1631     2.9404  -5.497 4.25e-08 ***
## CBSYes       -6.0707     3.8623  -1.572  0.11613
## NYTYes      -27.7890     3.3991  -8.175 4.59e-16 ***
## FOXYes       44.7044     2.7523  16.242  < 2e-16 ***
## wapoYes      -5.7178     3.5995  -1.589  0.11230
## BBCYes       -6.1161     3.5013  -1.747  0.08079 .
## ABCYes      -10.8632     3.5689  -3.044  0.00236 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.97 on 2563 degrees of freedom
##   (1697 observations deleted due to missingness)
## Multiple R-squared:  0.1869, Adjusted R-squared:  0.1837
## F-statistic: 58.92 on 10 and 2563 DF,  p-value: < 2.2e-16
```

At an adjusted R-squared of 0.1837, it certainly isn't anything to write home about. However the signs of the parameter estimates are interesting in and of themselves, as they show a possible correlation between source of internet news and candidate preference.

Model Assumptions

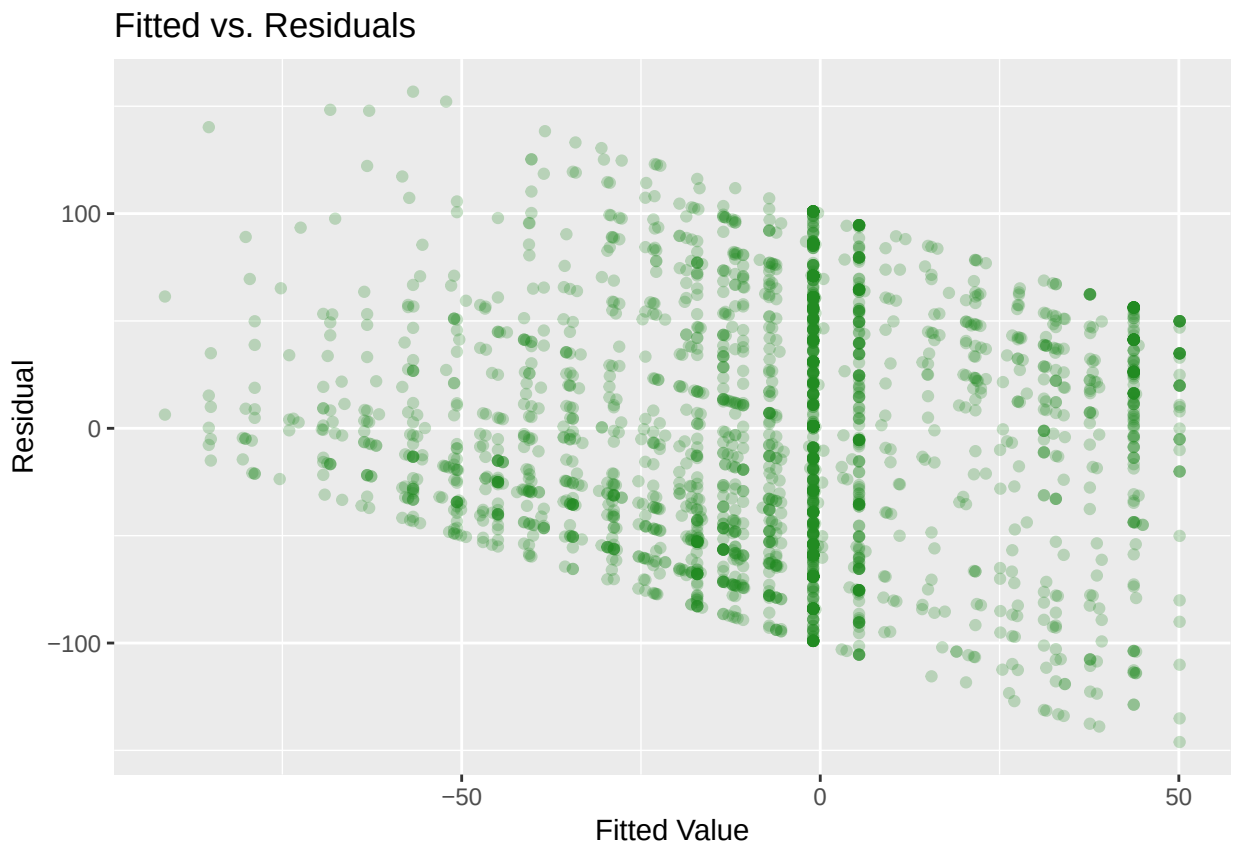
This type of linear model requires that some assumptions be met. Right off the bat, I know that the source of news variables and website variables are *not* independent. In fact, respondents can (and often did) choose more than one. So, there can be some collinearity.

Next, I exampined the residuals. They should be more or less normally distributed, which it appears that they



(somewhat) are from the graph below:

Next, I checked for uniform variability by plotting the residual values against the predicted values:



There is definitely a pattern in the above plot. So, there could be other variables in the survey that is causing the pattern and, if added to the model, may add more accuracy. With another 1,262 variables that is certainly possible.

Conclusion

In conclusion, there seems to be a pattern of some kind between the specific source of election-related news and a person's candidate preference. However, with the large number of variables, such a pattern may be difficult to detect with any degree of accuracy.

Also, such a pattern may not be linear in nature and may require more advanced techniques to properly quantify.