Russian Facebook Propaganda Efficacy Using Statistical Modelling

University of Virginia Department of Statistics
Donald Cooper, Jan Danel, Paul Franklin, and Tiger Hu
4/28/2020

Table of Contents

Abstract

The motivating question going into the analysis of the Russian Propaganda involved the efficacy of advertisements by the Russian "Troll" Farm that targeted the American public through social media platforms. Originally intended to observe the effectiveness of the ads as a political weapon leading up to the 2016 General Elections, the major factor considered in this study was to provide supporting evidence in the change of voting patterns due to the Russian propaganda. To analyze a change in voting patterns, a comparison was drawn from the election cycles during both 2016 and 2012 using the latter as the control group for the study. The underlying research question has since evolved in predicting the overall effectiveness of Russian ads given a combination of predictors like location, ad theme, and demographic information about the populations targeted among different states. The current analysis uses Bayesian methods applying a two-way fixed-effects linear regression model to measure the efficacy of ads with the response variable as a Click-through Rate among the designated state and year groups. Robust MM-Estimation modeling was used to improve original findings. Amid this research, a fascinating new data set was released, which broadened the analysis potential of this study. Specifically, political ad data gathered from ProPublica and NYU affiliates will be used to predict whether an advertisement is propaganda by comparing the original propaganda data with the ProPublica data.

~ ~ ~ ~ ~

"We will never know whether the Russian intervention was determinative in such a close election. ... What does matter is this: The Russians successfully meddled in our democracy and our intelligence agencies have concluded they will do so again."

- Ranking committee Democrat Representative Adam Schiff

~ ~ ~ ~ ~

I. Introduction

On February 16, 2018 Special Counsel Robert S. Mueller III indicted 13 Russian individuals and three Russian organizations for engaging in operations to interfere with U.S. political and electoral processes, including the 2016 presidential election. This was a significant step forward in exposing a surreptitious social media campaign and holding accountable those responsible for this attack. The indictment spells out in exhaustive detail the breadth and systematic nature of this conspiracy, dating back to 2014, as well as the multiple ways in which Russian actors misused online platforms to carry out their clandestine operations. In conjunction with this investigation, authorities ordered Facebook to locate and disclose data relating to the Russian Internet Research Agency sponsored Ads placed on their site.[1] This data is the basis of our analysis.

Our motivating question going into our own analysis of the Russian Propaganda involved the efficacy of advertisements by the Russian "Troll" Farm (The IRA) that targeted the American public through social media platforms. We originally intended to observe how effective the ads were as a political weapon leading up to the 2016 General Elections. The major factor we considered to help answer this question was the change in voting pattern due to the Russian propaganda. To analyze a change in voting patterns, we did a comparison from the election cycles during both 2016 and 2012 using the latter as the control group for the study. We also considered accounting for past presidential elections to build a

[1] "Social Media Advertisements." U.S. House of Representatives Permanent Select Committee on Intelligence. Accessed April 28, 2020. https://intelligence.house.gov/social-media-content/social-media-advertisements.htm.

time-series analysis. Considerations were also made as to whether data should be gathered on whether the IRA chose to create advertisements in prior years before the 2016 elections (i.e. our control year of 2012) as another baseline of comparison for ad effectiveness. However, our research question has since evolved as we found no significant evidence of a potential relationship between voter turnout and click through rate (i.e. effectiveness) across the country. This contradicts the findings of Spangher et al., which was an influence of ours going into this project[2(a)]. We instead decided to ask whether we can predict the effectiveness of Russian ads given a combination of predictors like location, ad theme, and demographic information about the populations targeted among different states. We plan to measure effectiveness with Click Through Rate.

Spangher et al.'s *Analysis of Strategy and Spread of Russia-sponsored Content in the US in 2017* provides evidence that Russians targeted unregistered U.S. Citizens after the election. However, the literature did not seem to have evidence to whether the advertisements had any effect on unregistered U.S. Citizens before the election. Our model is intended to give an estimate on how effective the advertisements were on unregistered citizens and whether the ads provoked a significant percentage change in registered voters from unregistered voters from 2012 to 2016[3]. In addition to testing the effectiveness of the advertisement, we wanted to measure the effectiveness by location, preferably at the city or state level. This will help us see if some locations are more susceptible or were targeted more by the Russian ads.[4(b)]

During this research, a fascinating new data set was released, which broadened the analysis potential of this study. Specifically, ProPublica published a survey of Facebook Advertisement data gleaned from the desktops of their users.[5] On top of this, the researchers learned that NYU affiliates scraped and compiled ad data from Facebook's ad Library back in 2018.[6] These sets appeared very compatible with the Facebook ad propaganda data released by the U.S. Government (The latter will be referred to as the HPSC data). Considering this, our analysis asks whether it is possible to accurately predict that a given Facebook ad is propaganda, given a combination of variables such as thematic content, syntax, observer location, number of factual references, etc. The results of this analysis are presented in section VII.

II. Data and Data Collection

i. Advertisement Efficacy Data

The data we selected was the same data collected during the House Intelligence Committee Minority Investigation by Special Counsel Robert S. Mueller who indicted 13 Russian individuals and three Russian organizations, including the IRA, for engaging in operations to interfere with U.S. political and electoral processes such as the 2016 presidential election. Facebook released "a total 3,519 total advertisements [that] were identified to have been purchased" with over "11.4 million American users exposed to those advertisements". The data has 25 different variables which include AdIDs, Adtext, Clicks, Impressions, Locations, CreationDate, EndDate, and AdSpend, all of which we use in our analysis[(c)]. To interpret and use most of the data we convert the variables in character or string format to categorical data. In addition to the 3,519 advertisements from Facebook made public by the House

---

[2] Spangher, Alexander, et al. "Analysis of Strategy and Spread of Russia-Sponsored Content in the US in 2017." Carnegie Mellon University, n.d. Accessed April 28, 2020.
[3] Ibid. p.8
[4] Dutt, Ritam, et al. "'Senator, We Sell Ads': Analysis of the 2016 Russian Facebook Ads Campaign." Indian Institute of Technology Kharagpur, India, n.d. Accessed April 28, 2020.
[5] ProPublica. "Political Advertisements from Facebook." ProPublica Data Store, March 19, 2019. https://www.ProPublica.org/datastore/dataset/political-advertisements-from-facebook.
[6] "FBPoliticalAds." Facebook Archive. shikhar394. Accessed April 28, 2020. https://github.com/online-pol-ads/FBPoliticalAds/blob/master/docs/Facebooks-archive.pdf.

Intelligence Committee Minority, we have also gathered data from the U.S. Census Bureau and the U.S. Bureau of Labor Statistics that mostly has to do with demographics, voting statistics, and median household income measured by state. This additional data was selected by states to combine this data with the Facebook Ad data which we decided to break down to state-level, as we will explain.

ii. Propaganda Detection Data

To answer the categorical research question (whether it is possible to accurately predict that a given Facebook ad is propaganda), researchers began by row-binding the HPSC data to the ProPublica data. Ad Ids were preserved for recognition, then both sets were subset to the following variables: target Gender, ad text, and target location. For exploratory data analysis, researchers compared the entire ProPublica political ad set (158117 ads) to the whole HPSC propaganda set (3516 ads). This way, researchers could look for noticeable differences in the proportion of ads with certain textual themes in each set. Unfortunately, only a handful of ads (about 10) in the ProPublica data were on Facebook at the same time as the propaganda data. This is a major potential objection to this analysis, so one solution researchers sought was to select themes based on their consistency over time. Specifically, the variables analyzed in section VII (Violence, race, voting) were chosen because they appeared consistently for the whole duration of Russia's 2015-2017 Facebook propaganda campaign, or in a large portion of the ProPublica ads. Boyd's findings[7] aided significantly in this regard. The assumption is that time-invariant themes were present in political ads in both time periods under consideration. The final step of data preparation randomly sampled 3516 ads so the model could analyze equal proportions of organic ads and propaganda, and so the testing set would not contain more ProPublica ads than HPSC ads.

iii. Data Dictionary for Analysis

Variable Name (R name)	Variable Type	Description
Click-Through-Rate (CTR)	Response Variable, E(y), y (Quantitative)	CTR is the proportion of Clicks to Impressions in each state. Note that CTR is denoted as a percentage.
AdWordCount	Quantitative Predictor	Number of words that appeared in an ad
AdDuration	Quantitative Predictor	The duration period for an ad, or the difference between when the ad was created and ended.
Ad_Freq_state	Qualitative (Categorical) Predictor	The frequency (i.e. count) of states that the ad appeared in. Count by state essentially forms a categorical explanatory variable.
ElectYear	Qualitative Predictor	A binary variable that indicates whether the ad appeared during the 2016 presidential election year (as 1) or not (as 0).
State (e.g. "Virginia", "Maryland", etc.)	Qualitative Predictors (fixed effects)	A binary variable that indicates the which state the ad appeared in by a 1 if present, 0 if not.

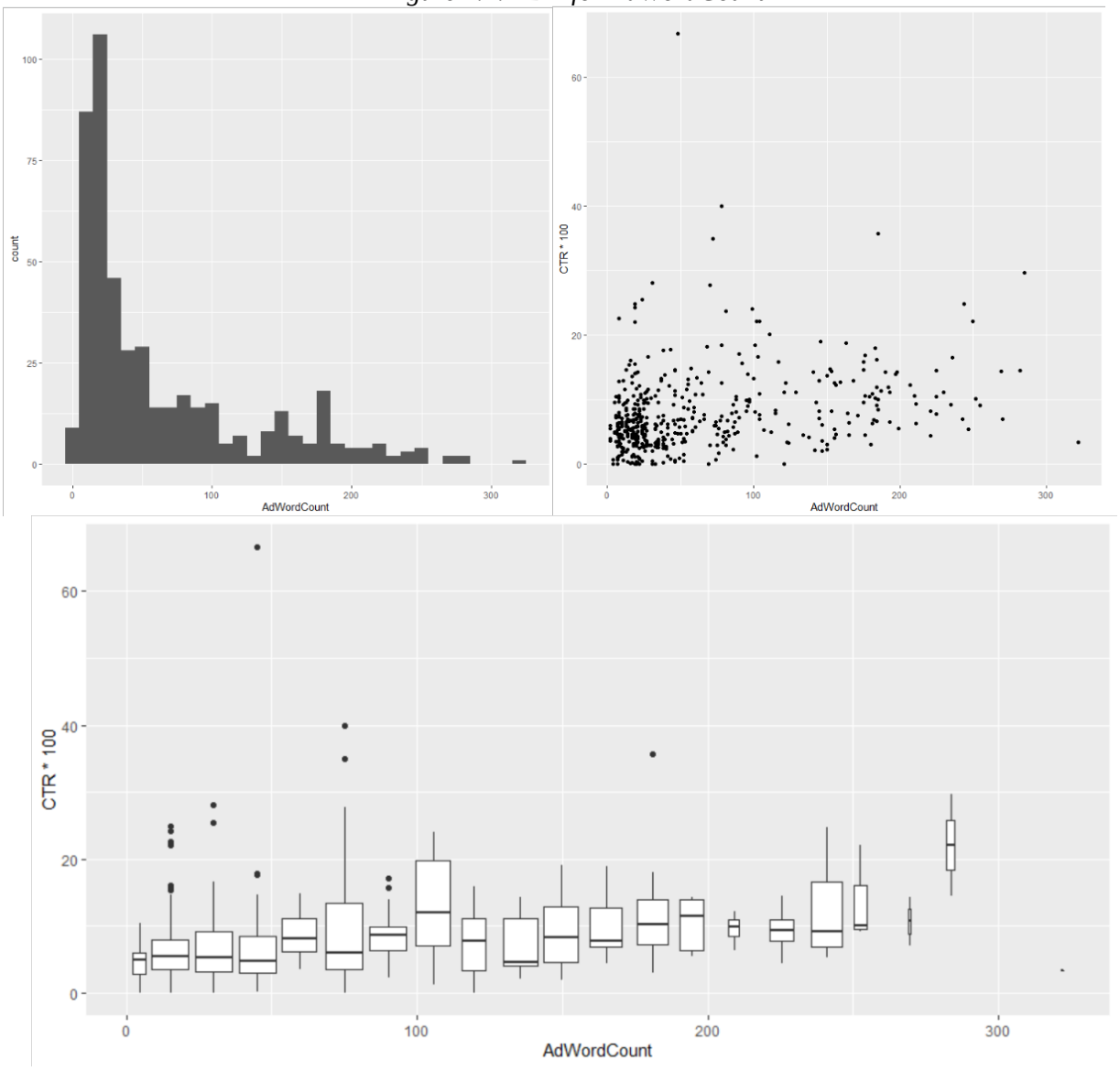
III. Exploratory Data Analysis (EDA):

While the data cleaning provided multiple variables as potential predictors, our Exploratory Data Analysis and intuition behind the origin of the data lead us to use only four variables for our first model in predicting the Click-Through-Rate (CTR) within each state: AdWordCount, AdDuration, Ad_Freq_state,

[7] Boyd, Ryan L, et al. "Characterizing the Internet Research Agency's Social Media Operations During the 2016 U.S. Presidential Election Using Linguistic Analyses." University of Texas at Austin, n.d. Accessed April 28, 2020. https://files.osf.io/v1/resources/ajh2q/providers/osfstorage/5bb21c61717460001650732d?action=download&version=2&direct&format=pdf, page 3.

ElectYear. When deciding between these variables and the others listed within the Method of Analysis section, much of our decisions were based on the fact that these four variables would directly affect both impressions and/or clicks by either a factor of higher count value (i.e. variables AdWordCount, AdDuration, and Ad_Freq_state) or by a binary factor of whether the ad appeared during a certain year or not (ElectYear). Additional evidence to support that the results from these variables would be more significant than other variables can be found in the supporting literature and studies conducted by Dutt and Boyd. These studies were conducted with similar variables, if not the exact same variables, as our predictors. So, using them in our model made sense and limited the time needed in creating entirely new predictors. One liberty, although not entirely without merit, we took when creating the binary variable ElectYear was figuring that more of the ads that received a higher CTR would be released during an election year. For reference, the average CTR of an ad running on Facebook in 2016 was .0149, so 1.49% of impressions resulted in a click.[8] We coded the variable ElectYear to mark when an ad was observed during the year of 2016. The results from our EDA begin with Figure 1.1 after the Data Dictionary.

Figure 1.1: EDA for AdWordCount



[8] https://www.statista.com/statistics/343220/facebook-advertising-metrics/

Figure 1.2: EDA for AdDuration

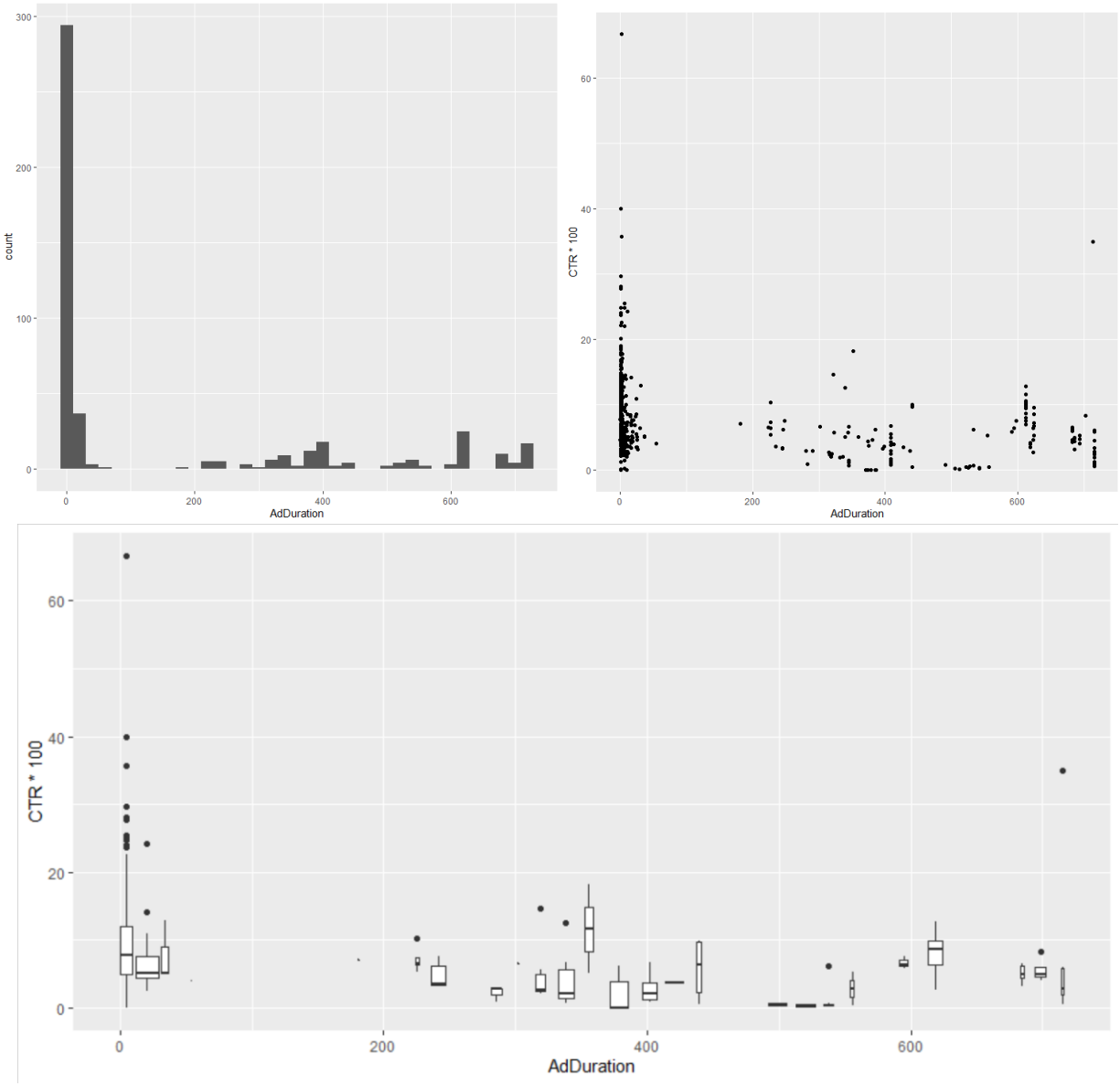
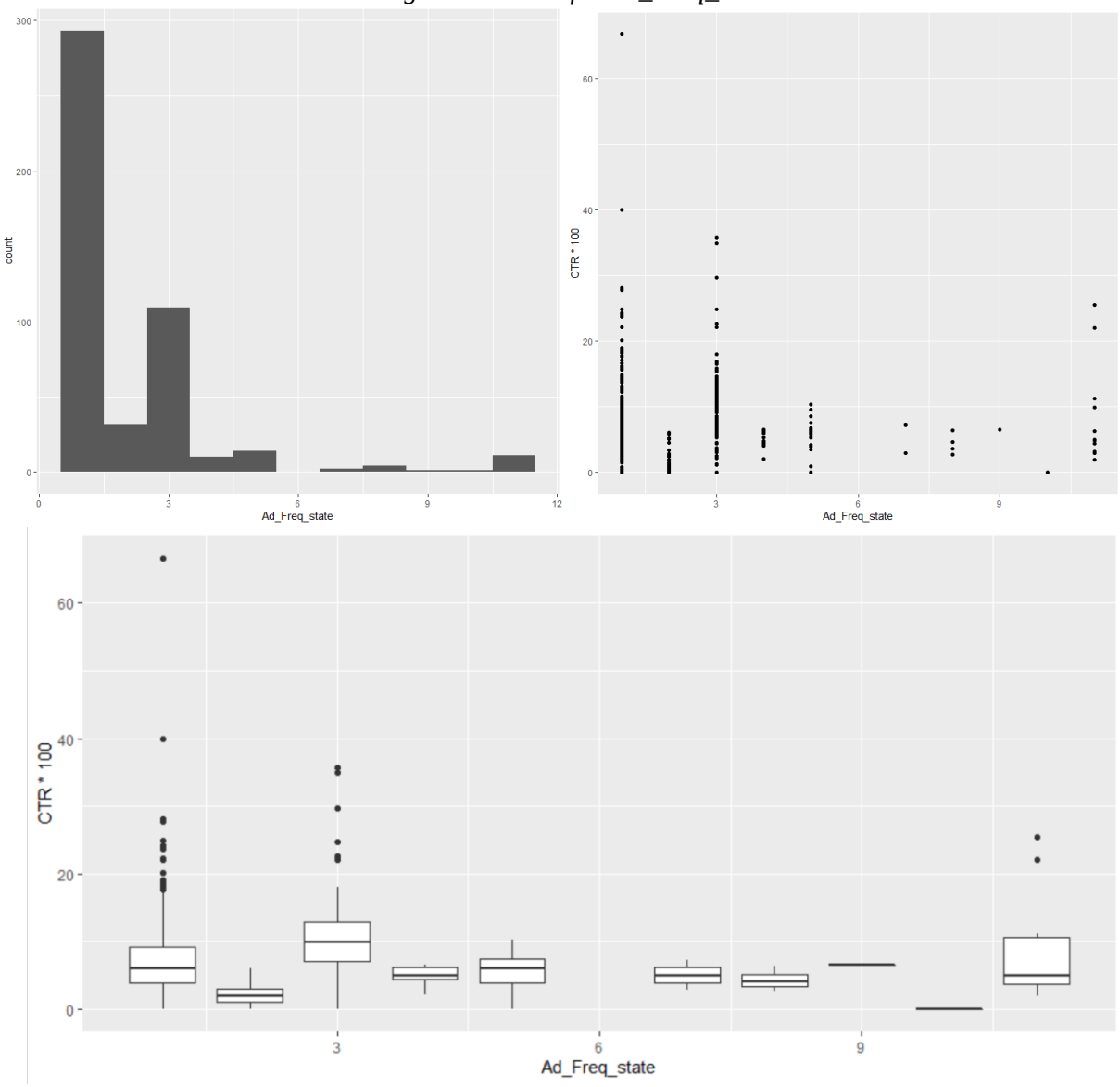


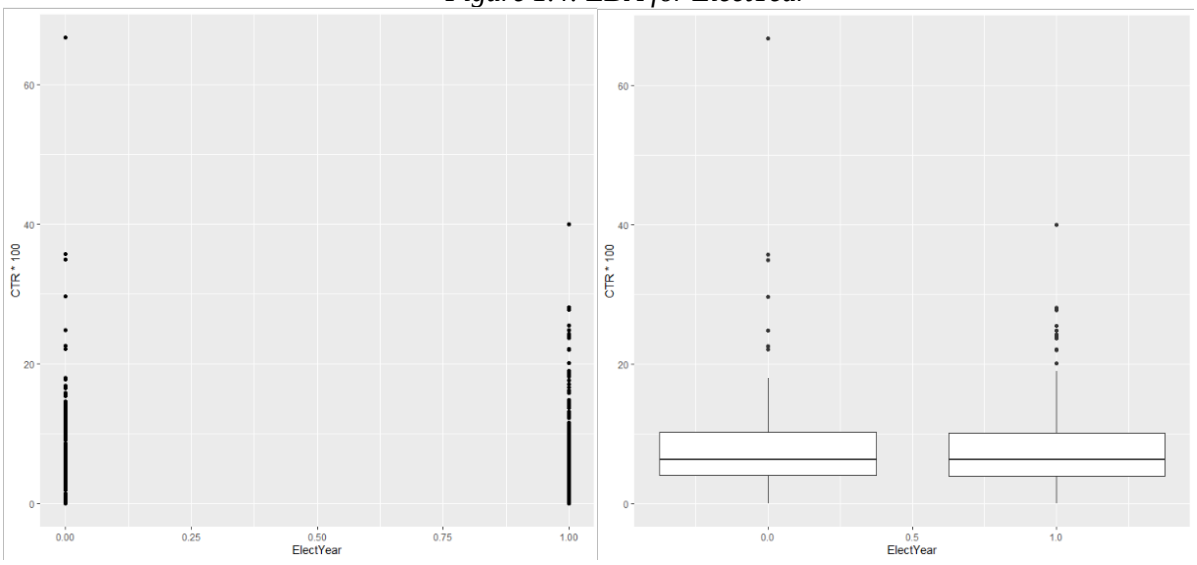
Figure 1.3: EDA for Ad_Freq_state



As we can see from the above graphs, specifically the histograms, that AdWordCount, AdDuration and Ad_Freq_state all tends to be skewed to the right with varying degrees of strength. AdWordCount seems to be the closest to normal, although far from it with most of the data concentration below 100 words per ad. Meanwhile, AdDuration is the complete opposite with almost all ads lasting less than 50 days with some clusters between 200 to 400, 600, and 800 days. This difference in distribution is also clearly seen when comparing both of their boxplot graphs. AdWordCount tends to have all the boxplot means at or below 20 CTR*100 with some clear difference in distribution (Q1 and Q3) but not something that is significantly different as most boxplots tend to overlap a good amount of their length. However, for AdDuration, all the boxplots are very different from each other with the means being almost all different and almost no overlap between the length of the boxplots themselves. Meanwhile, Ad_Freq_state is a little more difficult to check for any type of assumptions as the ads were specifically placed in US states by the IRA. In other words, these data points were not random, or normal, given the fact that people specifically chose where their ads were going to run. For this reason, we see that the most ads showed up in a single state, with the majority showing in at most three. This can most likely be

attributed to the difference in ad context that is more specific to certain states which the Russians most likely took advantage of to create specific division in specific places which they thought would be more effective for their cause. Something similar can be said about AdDuration, that the IRA can have some control for the amount of time an ad is online, but this is more random as the IRA isn't able to see how effective an ad is until being posted. For this reason, we see that AdDuration is more random and spread out.

Figure 1.4: EDA for ElectYear



Since election year is a binary variable, the ad either appeared during 2016 or not, then we see that there are only two boxplots. From these boxplots we can see that the distributions of CTR*100 between both election year and not election year are almost identical. The mean, Q1, Q3 and even the whiskers seem to be almost at the same level for both categories.

The following quantitative and graphical exploratory analysis pertains to the categorical research question.

Figure 1.5: Number of Russian Advertisement Targets by City and State

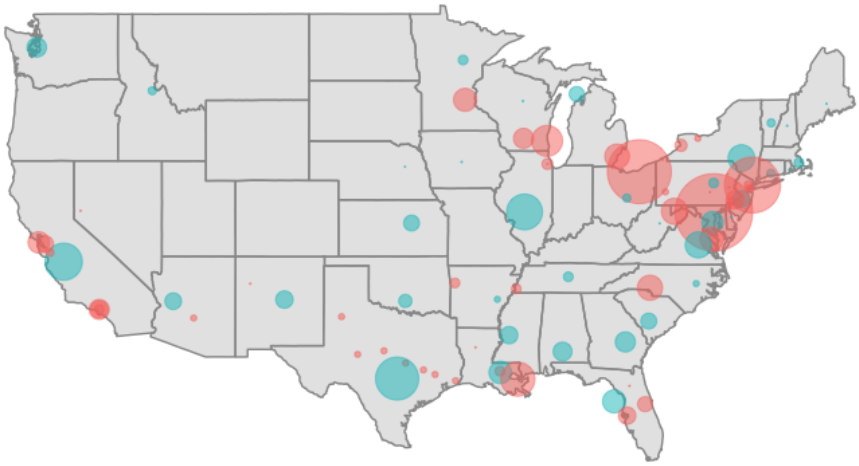
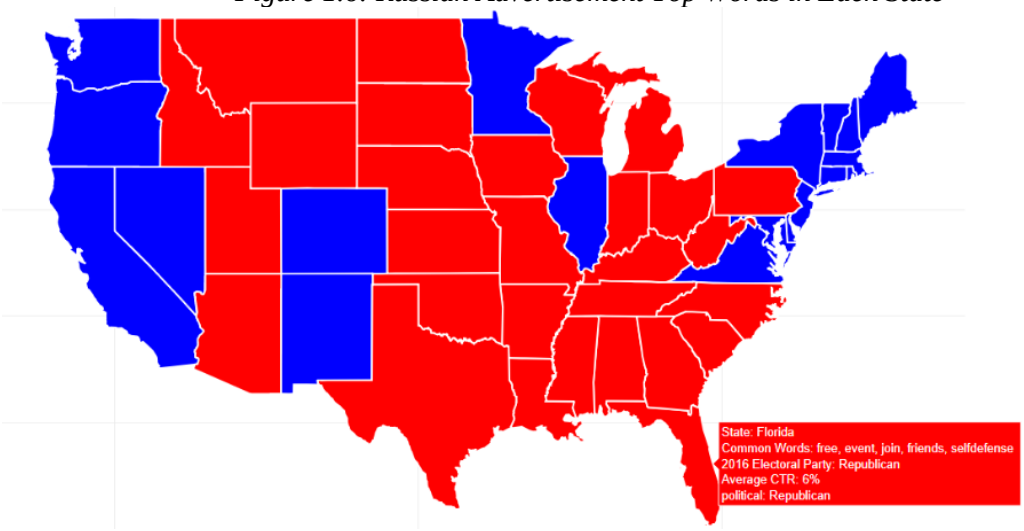


Figure 1.5: Each circle represents a city or state targeted by the Russian advertisements. Blue circles indicate the number of times that state was targeted, whereas red circles represent cities. The size of the circles are proportional to the number of times that location was targeted by an advertisement. One should note that while the largest bubble displayed here corresponds to Baltimore, MD with 244 targets,

while our data contained 2566 ads that only specified “United States” as the target location. The map provides another interesting dimension of the distribution of our location dummy variables. Notice the specificity of many Russian advertisements, which often targeted northeastern cities. This reveals a concerted effort on the part of Russian agents to influence specific groups of people, as opposed to a ‘land where it may’ strategy. This graphic informed the choices of the location predictor categories, since this study assumes that locations frequently targeted in the past are good predictors of future propaganda. To view this interactive graph click the following link: https://pbf2tp.github.io/.

Figure 1.6: Russian Advertisement Top Words in Each State



In figure 1.6, we see the 2016 electoral map which color is coded by the political party that won the electoral votes in the state: red for Republican Party and blue for the Democratic Party. As an example, we see that by hovering over Florida we can see the most frequently used words in Russian ads, when getting rid of stop words. These five common words were “free, event, join, friends, selfdefense” and they are all in lowercase and without hyphens and other punctuation, as we needed to make sure that all punctuation was the same to get the correct result. Additionally, we can see the average click through rate for all the Russian ads that were present in Florida. Other states such as Minnesota, were targeted by more Islamic and Muslin related content with its five more frequent words being “jamar, jsutice, mulsim, november, shot” which is of no surprise since Ilhan Omar, one of the first two Muslim women in Congress, is a Minnesota Representative. Another example regarding politics is Pennsylvania, where the top five words were “pennsylvania, trump, join, coal, jobs” which very much reflect the rallies that Trump held during his first presidential campaign. Lastly, there were other states that did not have as political ads such as Louisiana where the ads mostly revolved around race, with its top five words being “black, join, proud, south, lives”. To check out this interactive graph click on the following link: https://jandanel.github.io/Fall-2020-Capstone/.[(d)]

To look for an association between our variables and whether an ad was propaganda, we looked at the proportions of each class for each potential predictor. Starting with an indicator variable as to whether the thematic content was ‘violent’, the ProPublica political ads had violent themes 28.16% of the time, whereas the propaganda data mentioned a violent word 32.5% of the time. This was not the largest margin of change, but it was decided to include it in the model anyway since it was a consistent theme throughout the length of the IRA’s Facebook campaign. Researchers also noted a much different usage of exclamation points. Investigating further, the propaganda data used a lone exclamation point 46.07% of the time, whereas the political ads only used it 12.96% of the time. This was a considerable difference, especially compared to that of the last variable, so it was tested. Similar methods explored whether ads referenced law enforcement officers (propaganda referenced police 15.58% more than the political ads),

and whether the ad included racial words (this was the greatest difference, with propaganda mentioning race 28.725% more often than ProPublica ads).

Researchers also created a measure of the number of facts in each ad. Specifically, a variable counted the number of times each ad included a digit, a link, proper nouns, or parenthetical notation. The typical political ad in the ProPublica set had 11.34 'facts', whereas the typical propaganda ad had only 2.69 ‘facts’. Finally, researchers constructed a variable which indicated the presence of a location frequently targeted by propaganda. These were St. Louis, Cleveland, New York, Atlanta, Texas, California, and the United States. Encouraged by these results, the analysis was undertaken. The political ads targeted these locations 52.15505% of the time, whereas the propaganda targeted these places 98.89078% of the time. The margin was increased substantially upon the addition of a search for the “United States” as it seems that the Russians almost always must specify which country they want to target.

IV. Method of Analysis

We needed to measure ad effectiveness on different locations/demographics. In other words, we needed dummy variables to differentiate ad effectiveness by each unique location. While there are many different location subcategories, such as town or city, we opted to use state level location parameters because of the abundance of data available through government websites such as the Census Bureau, Bureau of Labor Statistics, etc. Advertisements would then be categorized by the states they targeted. Some ads included only one individual state while others needed to be split from other coded values, such as having a city coded in next to the state name which identified that the ad targeted a particular city within that state. There were 982 entries that included a state level characteristic remaining after the initial cleaning. Other ads that were either from a different country or did not contain any state level location were removed from the dataset. Some ads happened to have multiple states that they had targeted. To account for ads with two or more targeted state level locations, the states were separated, and advertisement IDs were counted the number of times it was included within a different state.

In addition to identifying which ad appeared in which state, a method needed to be developed to test the advertisement efficacy within each state individually for comparative purposes. For each state, a dummy variable was created to account for each location in which an ad had appeared. For the purposes of our model, the state dummy variable would account for when a specific ad had appeared or not for that respective state. The graphic below shows an example of what is occurring within the model. When the model is accounting for variables “AdID” = 1643, “AdCount” =124, “State” = Texas, the model will only attribute the state dummy created for Texas (i.e. TX in the model) with “AdID” = 1643, “AdCount” = 124, and “State” = Texas. This works for all other states.

Example of the role of the state dummy variable in model regression.

AdID	AdCount	State	TX	VA	MD
1643	124	*Texas*	1	0	0
23	25	*Virginia*	0	1	0
412	30	*Maryland*	0	0	1

Many of the variables that were needed were created through coding the original Russian Ad Dataset, many of the variables were also kept. The original variables we included were Clicks (Clicks), or the number of clicks that Ad had received, and Impressions: “an impression is counted as the number of

times an instance of an ad is on screen for the first time" (Impressions)[9]. The creation date (CreationDate) as well as the end date (EndDate) of the ad were kept in the dataset. For the advertisements that did not include an end date, we decided to add the end date of the ad as the date when the House Intelligence Committee Minority asked Facebook for the data. This happened sometime in January, therefore we opted to use the end date January 1st, 2017[(e)]. We decided to do this because we assumed that all the ads that Facebook provided to the House Intelligence Committee Minority were taken down no later than when Facebook turned the information over. The count of each advertisement (AdCount), as well as the cost for each ad (AdSpend), was included for each reported ad ID (AdID). The text (AdText) for each advertisement was also included. We created our response variable out of dividing the number of clicks over the number of impressions for a click-through-rate (CTR). The CTR is the number of clicks each ad had received in respect to the number of impressions it had garnered (i.e. number of clicks ad *i* received divided by the number of impressions ad *i* received). From both the creation and end date, we created a duration variable (AdDuration) that gave the time in which the advertisement was active in days. From the ad text variable, we created an ad word count (AdWordCount) variable that counts the number of words of that particular ad. Other than the location variables (Location), all the original data, including the created variables, were quantitative variables. In addition to the original data, other outside data was gathered to boost the effectiveness of the analysis. State level data regarding voting statistics, demographics, and population data were merged with the current Russian Ad data. Most of the useful predictors from this gathered data were quantitative in nature.

It was agreed upon that using a fixed-effects multiple regression model[(f)] was the best method for the analysis in examining the effectiveness of Russian political propaganda by accounting for the many differences between the states while simultaneously measuring the effectiveness of ads. As discussed before, the formula for a fixed-effects regression model is as follows

$$Y_{it} = \beta_0 + \beta_1 x_{1,it} + \cdots + \beta_k x_{k,it} + \gamma_2 D2_i + \gamma_3 D3_i + \cdots + \gamma_n Dn_i + u_{it}$$

and will account for all states for individual comparison to a baseline. Essentially, the model categorizes indicators for each subject, in this case each state, in the model. This makes each state a categorical variable to test the significance of our other predictors of the model within each state. For example, the categorical variable listed for Virginia is "VA", "MD" for Maryland, and continues through all the states included within the dataset including the District of Columbia (i.e. "DC"). Currently, we have discussed using Maryland as the baseline for our fixed-effects regression because Maryland was targeted by the Russian Ads the most. However, we have opted to use New York as our baseline because it is the median AdCount total of 85 different facebook ads. We concluded that using New York would offer a reasonable baseline for comparison because of the high concentration of ads as well the high concentration of population (StatePopulation). For example, a state could have been affected by ads more than New York or affected by ads less than New York based on the model prediction.

We will be conducting stepwise regression techniques to justify our choices in predictors. We will be calculating the Variance Inflation Factor (VIF) for all our predictors as well. If we see large VIFs within most of our variables, we may be inclined to use several augmentation methods like Ridge and Lasso regression to specialize to our data. This would introduce bias to our data and may not be needed as we have few predictors. Evidence of multicollinearity among our predictors will allow us to proceed with a multi-fold cross validation to pick our tuning parameters for Ridge and Lasso. However, if we see small VIF's throughout most of the predictors (i.e. only a few VIFs are large and little to know multicollinearity), this may not be a major concern and we would avoid Ridge and Lasso Regression. We also plan to conduct a Box-Cox transformation of the response if our variance is nonconstant. We also

[9] "Impressions." Facebook Business Help Center. Accessed April 28, 2020. https://www.facebook.com/business/help/675615482516035.

ruled out some nonparametric statistical methods like regression splines and smoothing splines. Other methods were not suitable given our objectives for the following reasons.

V. Supporting the Model

We believe a multiple regression fixed effects model to be the best algorithm for answering our research question. This is so because we are trying to account and control for the differences across states to make them as similar as possible. So, when we test for the efficacy of the ads, we will get the result of how individual ads affected states differently or similarly, and if they had a greater or lesser impact in specific states. Multiple regression is also most relevant here because we have a continuous, quantitative response variable. This conclusion lets us rule out classification techniques. Another advantage of multiple regression is that we can include categorical predictors. Including categorical variables (as dummies) will allow us to explain differences in CTR between locations. If we do not control for the differences between them, we would be facing Omitted Variable Bias which could cause the interpretation of our result to be invalid. This happens due to the fact that our model could be significant but only because the states are so different that they return significant differences, and if we say that ad efficacy was the only reason for the difference between states then we would be greatly mistaken. Outside research has identified multiple variables that correlate with the effectiveness of Russian political propaganda yet have not tried to predict effectiveness by combining all of these variables into a model. We believe that given new data, such a model could accurately predict the effectiveness of any newly detected advertisements (given some reliable predictors). This way, public officials and social media sites can narrow their search to the most effective types of advertisements or vulnerable populations, and hopefully be more successful at eliminating it in the future.

Since we had a response variable in mind from the beginning, we did not conduct Principal Component Analysis (PCA), which is unsupervised. This method would be informative if we wish to consider other response variables in the future but given time constraints and our desire to discern which ads were most effective, we skipped PCA. Regression decision trees were ruled out as they are generally less accurate in predicting quantitative response variables than multiple linear regression. Still, for the future a tree could let us compare the importance of one of our IVs to every other IV. A regression decision tree would also be advantageous if we wanted an algorithm that would produce an easily interpretable tree diagram. Such a diagram would be a good visual if we are fortunate enough to report our findings to the public. In the future, we might also consider a hierarchical or K-means clustering technique to identify subgroups of people most susceptible to ads. This could also be a valuable way to narrow the search for propaganda in the future, as the IRA will most likely try to exploit those subgroups it successfully influenced in the time frame of our data.

VI. State Based Click-Through-Rate (CTR) Model: Analysis & Results

i. Ordinary Least Squares (OLS) Fixed effects (FE) Model

First, we used R to perform the model and look at its diagnostic plots in hopes to improve it. Below are four diagnostic plots for the original linear (OLS) fixed effect model.

Figure 2.1: Residuals vs. Fitted for MLR Model

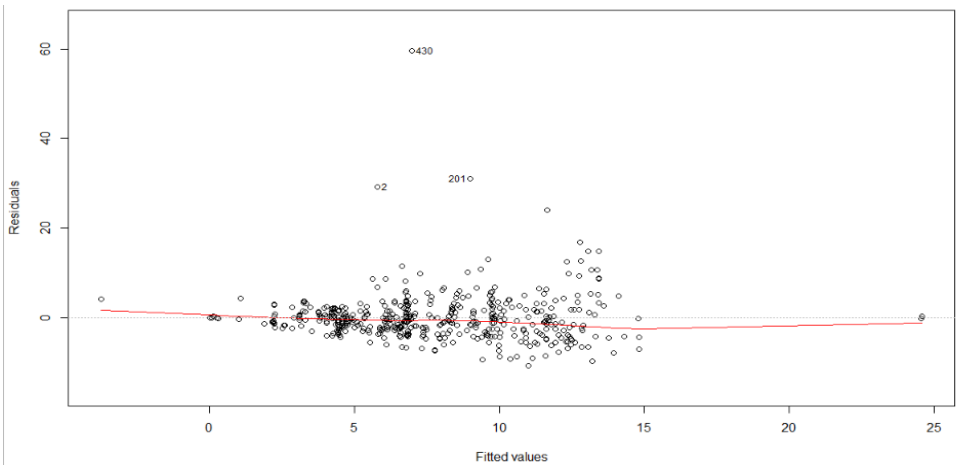


Figure 2.2: Normal Q-Q Plot for MLR Model

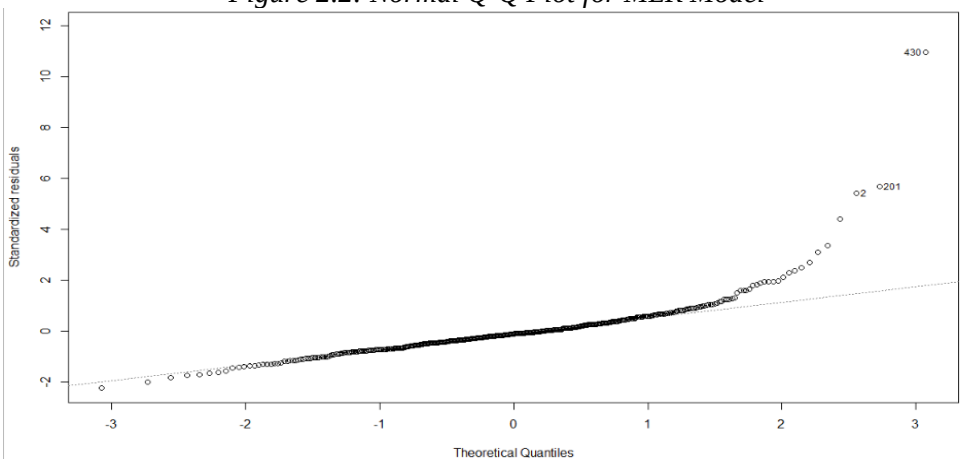


Figure 2.3: Scale Location for MLR Model

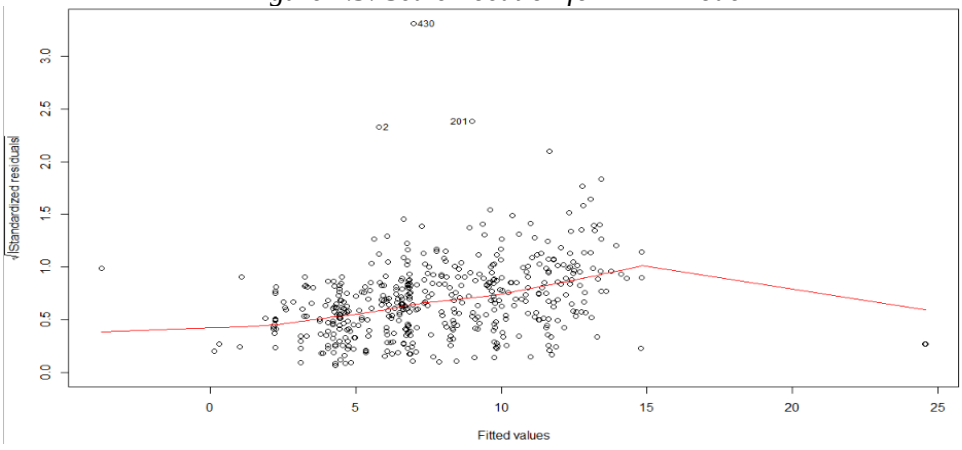
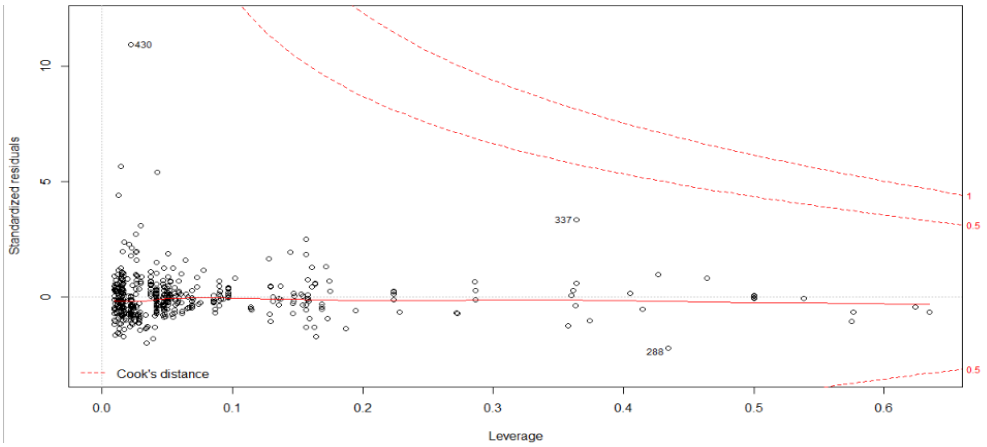


Figure 2.4: Residuals vs. Leverage for MLR Model



The diagnostic graphs showed clear indications that our intended model violates the assumptions of normality and homoscedasticity of residuals. Figure 2.1 indicates there is a fanning out pattern in the residuals, concluding that the errors do not have a constant variance for all levels of independent variables and are heteroscedastic. Viewing the results from Figure 2.3, we can prove further that the errors in the model do not follow constant variance when comparing the square root of the standardized residuals to the fitted values. This is likely due to the largely categorical data, as each category(state) has varied ranges from their recorded observations throughout each predictor, $\boldsymbol{\beta}_i$. Further, the results from Figure 2.1 also give us a hint of the non-linearity of the data by not directly falling on the origin. When observing Figure 2.2 there appears to be an exponential increase pattern in the standardized residuals of our model, or a distribution that is heavily skewed right. This is a clear indication that data does not follow a normal distribution. Finally, Figure 2.4 indicates that some of the observations were noted as outliers and not influential to the data. To account for this, we considered using a resampling procedure to limit the influence that outliers had on our data while keeping the leverage points that were pertinent to the model.

Following these models, we concluded that variable transformation was the most viable option. First, we conducted tests to transform the response variable. In determining the best transformation for our response variable, the Box-Cox method[10] of y-variable transformation was the next step.

Tukey Transformation to linearize the y-variable:

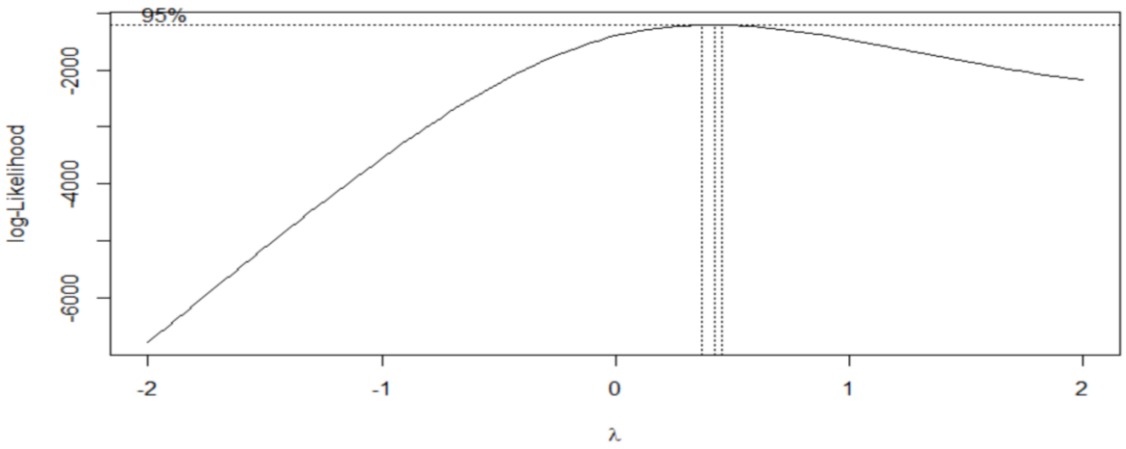
$$\hat{y} = \begin{cases} x^{\lambda} & \text{if } \lambda > 0 \\ \log x & \text{if } \lambda = 0 \\ -(x^{\lambda}) & \text{if } \lambda < 0 \end{cases}$$

where

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

Figure 2.5: Log-likelihood of the Box-Cox function

[10] "An Analysis of Transformations." Journal of the Royal Statistical Society. Accessed April 28, 2020. https://www.ime.usp.br/~abe/lista/pdfQWaCMboK68.pdf. pg. 214



We chose to focus on investigating the optimal value of lambda, $\boldsymbol{\lambda}$, in the Box-Cox transformation, denoted as $\boldsymbol{\lambda}^*$, using the algorithm of maximum likelihood estimator (MLE) by maximizing the log-likelihood of the Box-Cox function. This would give a more accurate estimate of the MLE than if we used Tukey's Ladder of Transformations[11], which essentially only considers the response variable's transformation of pre-calculated lambda values. One of the issues with using Tukey's Ladder was that both the $\boldsymbol{\lambda} = 0$, $\boldsymbol{\lambda} = 0.5$, and even $\boldsymbol{\lambda} = 1$, follow transformations of log(y), $\sqrt{y}$, and y^1 respectively and fall well outside of the calculated 95% Confidence Interval provided in Figure 2.5 above. Choosing a value between the confidence interval would provide the best possible maximum likelihood estimator for the box-cox transformation needed to account for the original model's non-normality and heteroskedasticity. Generally speaking, software packages often provide, or have the means of providing, an algorithm for calculating the MLE by maximizing the log-likelihood. Though an understanding of the algorithm mathematically clarifies what is best for statistical modelling.

The log-likelihood of our multiple linear regression model can be written as

$$\log L = -\frac{n}{2}\log(2\pi) - n\log\sigma$$

$$= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_{(i)}^{(\lambda)} - (\beta_0 + \beta_i X_i)\right]^2 + (\lambda - 1)\sum_{i=1}^{n}\log(Y_i) \qquad (1)$$

The function is rewritten as

$$L =* -\frac{1}{2}\log\left[\sum_{i=1}^{n}\left(\frac{(Y_i^{(\lambda)} - (\beta_0 + \beta_i X_i)}{\dot{Y}^{\lambda}}\right)^2\right]$$

where

$$\dot{Y}^{\lambda} = \exp\left[\frac{1}{n}\sum_{i=0}^{n}\log Y_i\right] \qquad (2)$$

[11] Scott, David. Tukey Ladder of Powers. Accessed April 28, 2020. http://onlinestatbook.com/2/transformations/tukey.html.

We can now calculate the optimal value of $\boldsymbol{\lambda}^*$ by maximizing the log-likelihood. While tediously calculated by hand, the 'MASS' Package[12] in R provides a quick and simple method for maximizing the log-likelihood and finding the optimal $\boldsymbol{\lambda}^*$ through the 'boxcox' function. Through running the automated test in R, the optimal value of $\boldsymbol{\lambda}^* = 0.383838\ldots$ for our proposed model. Being that $\boldsymbol{\lambda}^* \approx 0.38$, we utilize the transformation where $\boldsymbol{\lambda}^* \neq 0$, and apply y^{λ}, or $y^{0.38}$, to the model. We then once again computed the model graphically to observe normality, homoscedasticity, leverage, and outliers after transformation. The diagnostics are provided as follows.

Figure 2.5: Variance Inflation Factor (VIF)

Parameter	VIF	Increased Standard Error (SE)
AdWordCount	1.21	1.10
AdDuration	1.29	1.13
Ad_Freq_state	1.12	1.06
ElectYear	1.10	1.05

Before correcting the model with the y-variable transformation and the resampling, calculations were performed to obtain a Variance Inflation Factor (VIF) for the independent variables chosen for regression. Regarding the presence of multicollinearity, we observed low correlation between the variables selected for our model. Neither of the independent variables AdWordCount, AdDuration, Ad_Freq_state, or ElectYear showed a significant correlation to each other when using the VIF cutoff of 10, or that if a VIF > 10 is present we would have presence of severe multicollinearity. Additionally, the correlation matrix indicates very minimal autocorrelation between the independent variables. We saw no reason to remove any of the variables after concluding the model had low levels of correlation between the variables.

After determining an adequate VIF to proceed, the next step was to observe any added significance in our model when transforming the independent variables. In determining the best transformation for our independent variables, we utilized the Box-Tidwell method[13] of independent variable transformation. The Box-Tidwell method for common independent variables follows a graduating function in which we allow observations $\boldsymbol{y_1, y_2, \ldots, y_u, \ldots, y_n}$ to be present at $\boldsymbol{n}$ sets of conditions $\boldsymbol{x_1, x_2, \ldots, x_u, \ldots, x_n}$, where $\boldsymbol{x}$, is a $\boldsymbol{k \times 1}$ vector giving the levels of the $\boldsymbol{x}$'s for the $\boldsymbol{u}$th observation.

We assume that

$$E(y_u) = \eta_u$$

[12] Ripley, Brian, et al. "Support Functions and Datasets for Venables and Ripley's MASS." Package 'MASS.' R-Studio, April 26, 2020. https://cran.r-project.org/web/packages/MASS/MASS.pdf.
[13] Box, G. E. P. and Tidwell, Paul. "Transformation of the Independent Variables". *Technometrics*. Vol. 4, No. 4 (Nov., 1962), pg. 534 (20 pages). https://www.jstor.org/stable/1266288?seq=4#metadata_info_tab_contents.

$$E(y_u - \eta_u)(y_v - \eta_v) = \begin{cases} \sigma^2 & u = v \\ 0 & u \neq v \end{cases} \quad (3)$$

where the variance, σ^2, is unknown. Suppose that the response x can be closely represented over the region of interest by the function

$$\eta = f(\xi, \beta) \quad (4)$$

where the elements $\xi_1, \xi_2, \ldots, \xi_k$ of the vector ξ are the x's transformed in some way that

$$\xi_i = \xi_i(x_i ; \alpha_i) \quad (5)$$

with α_i a p_i-dimensional vector with elements $\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ij}, \ldots, \alpha_{ipi}$ the unknown constants of the transformation. The functional relationship also involves l unknown constants whose values of betas, $\beta_1, \beta_2, \ldots, \beta_t, \ldots, \beta_l$ elements of the vector β depend upon the particular transformations of the ξ's employed (i.e. they depend on the choice of α's for prediction).

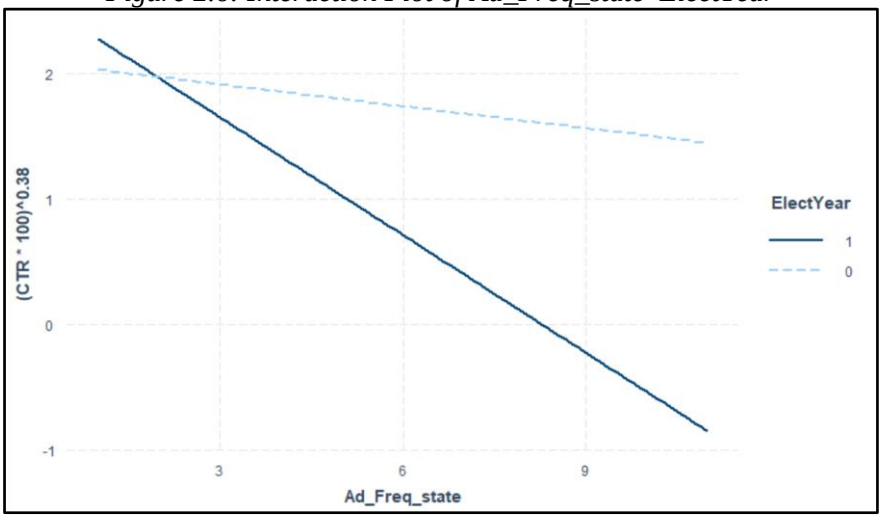
The calculations for determining the best ξ's depends on multiple vector calculations that are both too cumbersome and lengthy for this paper. The 'car' Package[14] in R-Studio saves us from calculating the guessed values of the transformation within each variable using the 'boxTidwell' function. The calculation determines the best α-value for each predictor, x. The function also provides the p-value for the estimate of each x transformation. From our model, we concluded the best variables to consider for transformation were the AdWordCount and Ad_Freq_state predictors. The Box-Tidwell method does not consider binary, dummy, or categorical independent variables, so these were left out. From the Box-Tidwell output, 26 iterations were calculated through the variable matrices (i.e. this appears to be the limit of 'boxTidwell' function provided in R) reporting a p-value of 0.099662 ($p \not< 0.05$) for AdWordCount and a p-value of 0.000301 ($p < 0.05$) for Ad_Freq_state. Like Box-Cox, Box-Tidwell provides the optimal MLE of λ. In the situation of independent variables, λ is the equivalent to the α_i. The optimal MLE of λ of our variable Ad_Freq_state transformation was 2.18365, or the approximation of a squared transformation when rounded to the nearest whole number. Therefore, we decided to include the transformation into our model, reporting no significant change to the assumptions of normality, linearity, or homoscedasticity.

Once we concluded to keep the higher-order term for Ad_Freq_state, the final step to model building was testing for significant interactions between our predictors. When considering which predictors should be interacted on each other, we carefully considered the data points of our predictors through the EDA procedure. First thoughts on interactions suggested AdWordCount against AdDuration, the idea that with more words in an ad meant more content being disseminated across state populations. However, the scatter plots seemed too inconsistent. AdWordCount observations become less consistent as CTR increases, and AdDuration has major inconsistencies for the time an ad could be accessed on Facebook. Another option was viewing the interaction between AdWordCount with ElectYear to see if there was a surge in content within the ad during the year of the presidential election. This also proved an unlikely interaction for much of the same reasons as AdWordCount against AdDuration. AdWordCount spikes simply occurred sporadically, meaning ads with extensive content (i.e. > ~50 words) were likely to have been seen regardless of it being an election year or not. The most sensical variables as an interaction term was Ad_Freq_state and ElectYear. Ad_Freq_state, or the frequency (i.e. count) of states that the ad appeared in, essentially forms a categorical explanatory variable due to the counts of ads appearing in different states. This helps us see how ads appeared across the country and whether they appeared more

[14] Fox, John, et al. "Companion to Applied Regression." Package 'car.' R-Studio, March 11, 2020. https://cran.r-project.org/web/packages/car/car.pdf.

than once in different states. Using the 'interaction' package in R, we were able to visually confirm the mean values of the predictors and whether their line of means crossed, indicating that an interaction would be possible. When interacted with the binary predictor ElectYear, the interaction plot in Figure 2.6 shows that not only is the interaction possible, but also that certain ads reappeared across the country during an election year. The idea for an interaction between these two predictors was to indicate whether Ad_Freq_state spiked during a very critical year in American Politics.

*Figure 2.6: Interaction Plot of Ad_Freq_state*ElectYear*



The interaction between Ad_Freq_state and ElectYear was significant ($p < 0.001496$) and was implemented to the model. It is worth noting that the optimal value of $\boldsymbol{\lambda}^*$ by maximizing the log-likelihood through Box-Cox remained $\boldsymbol{\lambda}^* \approx 0.38$ after both the variable transformation and the interaction were added. Thus, we utilized the original transformation where $\boldsymbol{y}^{\mathbf{0.38}}$, to the overall model. Now that our variables were selected, along with their higher-order terms and interactions, we moved on to the resampling procedure.

ii. Jackknife Resampling of OLS FE Model

Following the final steps of our model building, we considered the many outliers existing within our data as a possible issue to the overall end stage model. We had first discussed the possibility of the Bootstrapping method to resample the data because of the non-normality, but we eventually ruled out bootstrapping for several reasons. Considering the available data, we had (n = 476), bootstrapping was not necessarily needed as it generally works for smaller amounts of data than what our findings had gathered. Additionally, the extreme values, the outliers and/or leverage points, tend to not fare well when resampled with bootstrapping. While the bootstrapping method does offer a more consistent estimation of variance, the available data extracted for our modeling purposes were not sufficient to conduct the bootstrapping procedure. With this notion, we needed a sampling procedure that was more conservative with the presence of extreme data. Therefore, we opted for the more traditional Jackknife resampling procedure. Leave-out-out sampling, such as jackknifing, can be a simpler and more standardized procedure than resampling with replacement such as bootstrapping (i.e. the interpretability is more intuitive). Jackknifing requires observations to be deleted one at a time, each time refitting the regression model based only on the $\boldsymbol{n}$ - 1 observations, leaving out the $\boldsymbol{i}^{\text{th}}$ one.

With n jackknife statistics

$$\hat{\theta}^*_{(1)}, \ldots, \hat{\theta}^*_{(n)}$$

we can calculate the same measures as any other regression – MSE, bias, and variance – based on the jackknife samples where

$$\widehat{MSE}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta})^2 \quad (7)$$

and

$$\widehat{Var}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b^* - \overline{\hat{\theta}^*})^2 \quad (8)$$

and

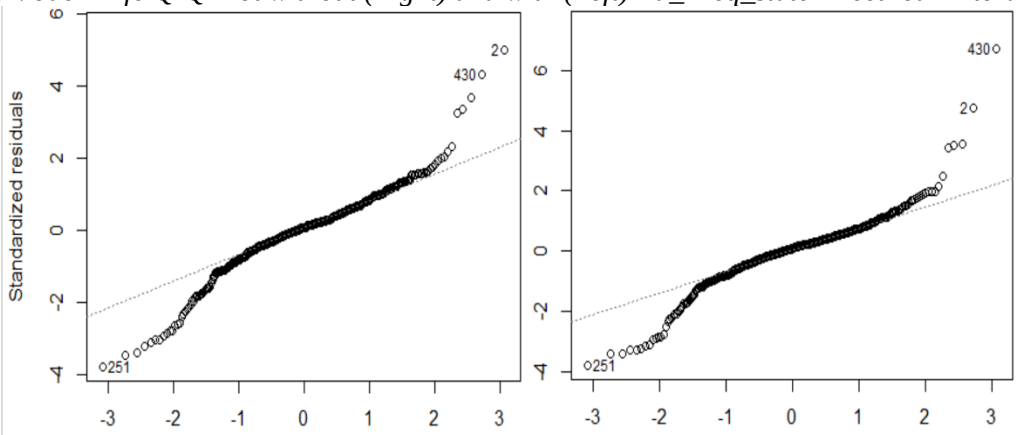
$$\widehat{bias}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^* - \hat{\theta} \quad (9)$$

where ***B*** represents each sample statistic.

We can estimate our parameters of interest ($\boldsymbol{\theta}$) using the appropriate sample statistic ($\hat{\theta}$). Jackknifing is computationally simpler than bootstrapping, and more orderly simply by the repetition of the same steps over and over for all ***n***. This means that, unlike bootstrapping, it can theoretically be performed by hand. However, for our data (***n*** = 476), it would be computationally intensive. The 'lmboot' Package in R-studio provides the function for re-sampling via the jackknife method. The model utilized in the jackknife model considers all the transformations of both ***y*** and ***x***, though, the fixed-effects for each state are left out simply for the fact that they only represent a controlled experiment within states utilizing our model predictors.

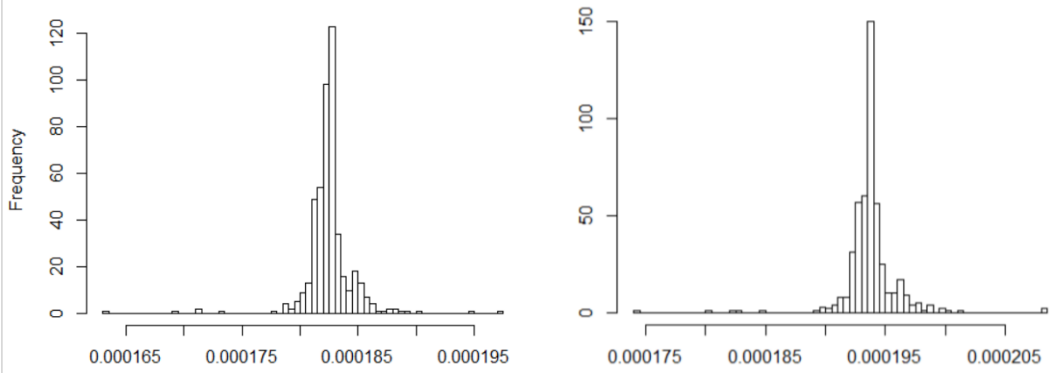
The results from our Jackknifed model were not the outcome we had hoped for, nor much better than the original model. AdWordCount remains significant in both models, yet only the interaction model retains significance in all variables. State fixed-effect dummies remain sparsely significant across both models with only 11 significant states. Despite the unpromising results, we conducted some additional diagnostics regarding the variable selection within the Jackknifed model. Our concern was that with a jackknifed model, there may be violation in the assumptions of normality, linearity, and constant variance. Figure 2.7 shows the linearity of our model with and without the Ad_Freq_state*ElectYear Interaction. The Q-Q Plot on the right indicates the normality of the model based on the standardized residuals and theoretical quantities (e.g. for the purposes of strictly graphing quantile against quantile) of the model without the interaction term, while the plot on the left indicates the model with the interaction.

*Figure 2.7: Jackknife Q-Q Plot without (Right) and with (Left) Ad_Freq_state*ElectYear Interaction*



The interaction and non-interaction models received adjusted R-square values of 0.3222 and 0.2661, respectively. While both plots are not from a true-normally distributed dataset (i.e. heavy tails), note that the both the interaction and non-interaction Q-Q Plots retains roughly the same amount of consistency in the tails. Other diagnostic plots between the interaction and non-interaction models showed similar results when observing the linearity and homoscedasticity of the two models. Figure 2.8 (below) indicates the jackknife sampling distribution of the model without the interaction. This histogram reveals a more conservative slope estimate (x-axis) than with the interaction model. Additionally, the non-interaction histogram appeared slightly more normal than the interaction model and more responsive to extreme values. Considering the data in question, we ascertained that the normality assumption was not too severely violated with the removal of the interaction, and therefore continued to move on without the interaction term for the prospects of better interpretability and lessened complications during robust regression. We continued with diagnostics regarding the non-interaction model to help us determine which robust modeling techniques would be best for our data.

Figure 2.8: Jackknife Sampling Distribution with (Left) and without (Right) Interaction



Notably, states such as Arizona (+22.23%), Idaho (+30.53%), New York (+5.36%), Ohio (+7.88%), Pennsylvania (+8.25%), Texas (+11.52%), and Washington (+17.09%) had estimates of greater than five-percent above the base category (i.e. Maryland). This indicated that Russian ads had a significant effect on these states likely due to the combination of higher CTR's$^{(g)}$, frequency in which ads appeared in these states, higher content, and both the release and duration of the content. Although not strictly a predictor, it is also likely that these states may have been affected more than the baseline due to their population size. In other words, the Jackknifed model showed that the average effect of our predictors on CTR percentage after controlling for state fixed effects was greater than the baseline. The state Missouri (+3.38%) showed the average effect of our predictors on CTR percentage after controlling for state fixed effects was greater than the baseline by only a modest percentage. Many states, while significant, showed that the average effect of the predictors on CTR after controlling for state fixed effects was less than the baseline by a given percentage. These states include Arkansas (-24.54%), Massachusetts (-20.70%), New Jersey (-15.46%), New Mexico (-25.07%), Oklahoma (-21.41%), and West Virginia (-20.30%). The other states that have insignificant coefficients showed that the Russian ads have the same effect on them as the baseline (Maryland).

Overall, the global F-test, where F (37, 438) = 5.6556 revealed a significant model where the model p-value (~ 0.0001) was much less than $\boldsymbol{\alpha} = 0.05$ but only returned an adjusted R-squared of ~ 0.27. Thus, after adjusting for sample size and degrees of freedom ($n = 476$, $d.f. = 37$), there is evidence to support that the predictors included within the Jackknifed model account for approximately 27% of the sample variance of the percentage change in Click-Through-Rate (CTR). We concluded that the model had a less than acceptable fit (generally speaking, an acceptable fit for data of this nature would be between 30-60% sample variance explained). More robust methods would be needed to account for the many outliers and leverage points contaminating the model and causing a less than satisfactory model fit.

Figure 2.9: Residuals vs. Fitted for Transformed Model using Jackknife

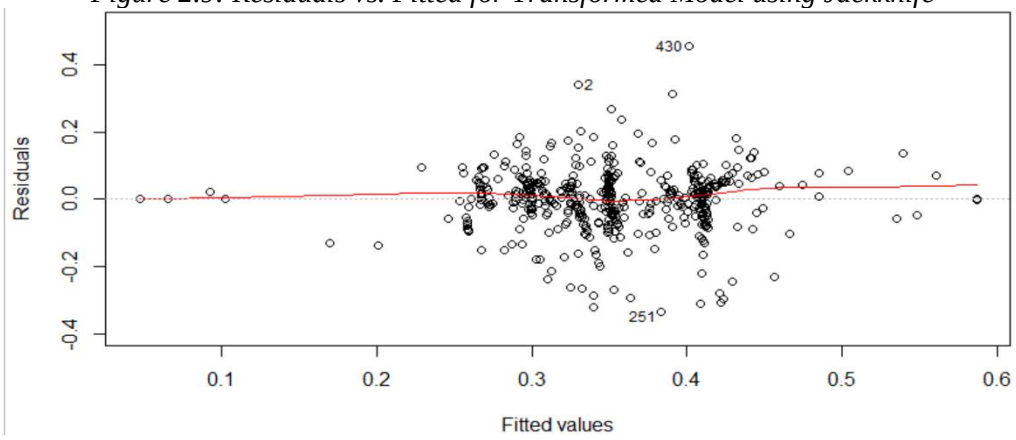
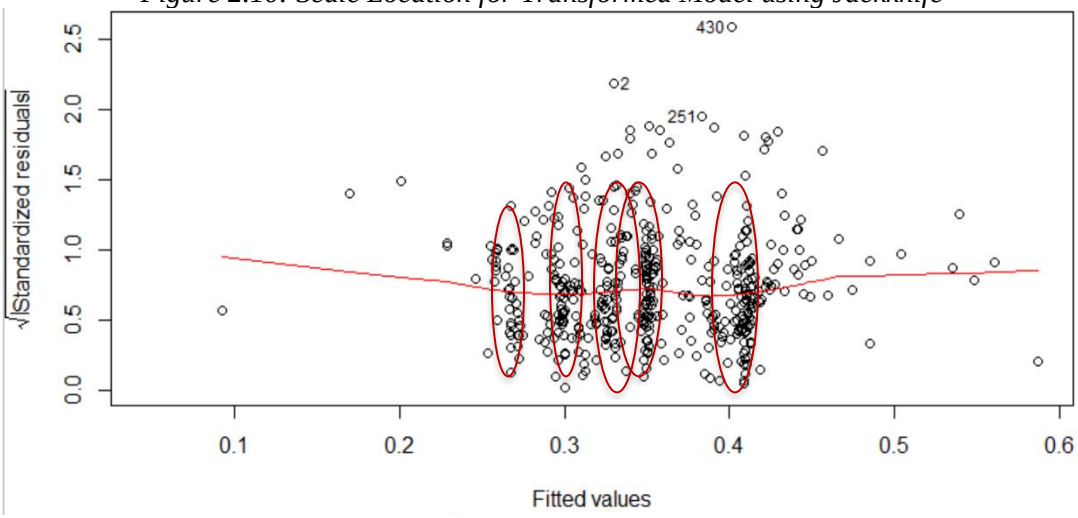


Figure 2.9 (above) indicates the linearity of the model after the jackknife procedure. While there is a noticeable change from the original model, the jackknife model has a more consistent residual vs. fitted value spread. Further, the resampling attempts for adjustment in the original fanning out pattern observed in Figure 2.1. We have concluded that the points in Figure 2.9, while not perfectly linear, have not severely violated the functional form enough to raise significant issues. While running these tests through regression packages in R throughout our model building stages, there have been specific observations (#2, 201, 249, 251, 288, 292, 322, 327, 347, and 430) that have consistently interfered with the regression assumptions of linearity, as well as normality and homoscedasticity. These are discussed later where more appropriate (Figures 2.12 & 2.13). Finally, there is a noticeable oval-shaped pattern in the residuals to fitted values, suggesting the possible violation in constant variance mentioned before.

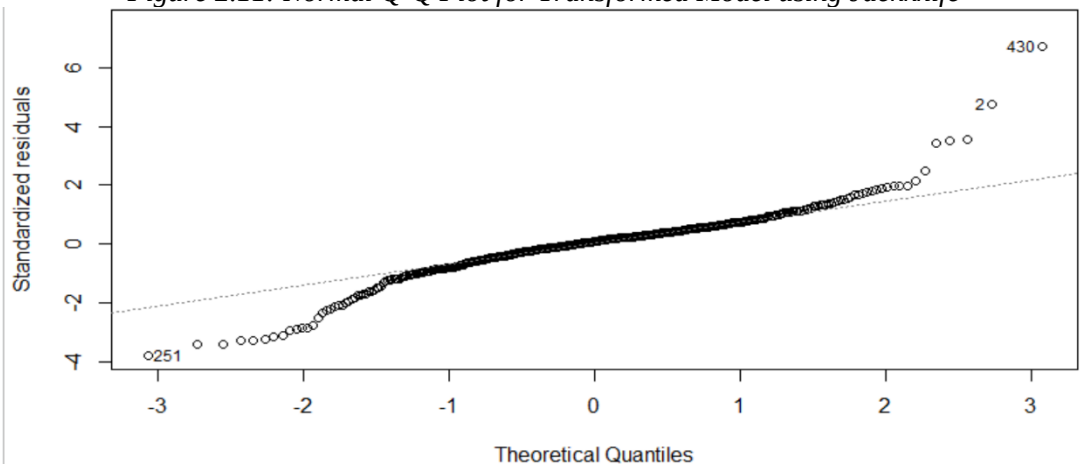
Figure 2.10: Scale Location for Transformed Model using Jackknife



Unusually high or low CTR observations tend to be the culprit in violating the assumptions needed for OLS regression. Such observations include those in New York City, which are exceptional in terms of connectivity, population, and other variables. Others states often simply include exceptionally different CTR than other reported observations. In addition, the errors do not appear to have constant variance, especially between the fitted values of 0.3 and 0.4 where a slight fanning out pattern is observed. While Figure 2.9 and 2.10 have residuals scattered somewhat evenly above and below the mean line, the points do not always 'fill a rectangular box' so to speak (see the tails), most points appear randomly scattered. For this reason, we do not fully conclude that the model has met the homoscedasticity assumption. We also noted that there is an odd clustering of standardized residuals circled in red between the fitted values of ~ 0.25 and ~ 0.41 likely attributed by the AdDuration variable. Recall that the

AdDuration variable contains our assigned date of May 26th, 2017 for ads which had no end date. This has likely attributed to many of the errors related to extreme observations and leverage points. Further action needed to be taken to account for the AdDuration variable during robust modelling because of the outlier and leverage points attributed from its use.

Figure 2.11: Normal Q-Q Plot for Transformed Model using Jackknife



Judging by Figure 2.11 (previously the right directional plot in Figure 2.8), one might question whether the residual normality assumption has been met. The points typically fall along a straight line on the Q-Q plot, which gives evidence that our data came from a relatively normal population apart from extreme observations in the tails. Still, the model was trained on many observations (***n*** = 476), so the regression should be robust to one or two violations of normality. Judging that the normality assumption has not been severely violated with any indication of exponential or logarithmic trends, we have confidence that the model has a linear interpretation apart from heavy tails.

Figure 2.12: Residuals vs. Leverage for Transformed Model using Jackknife

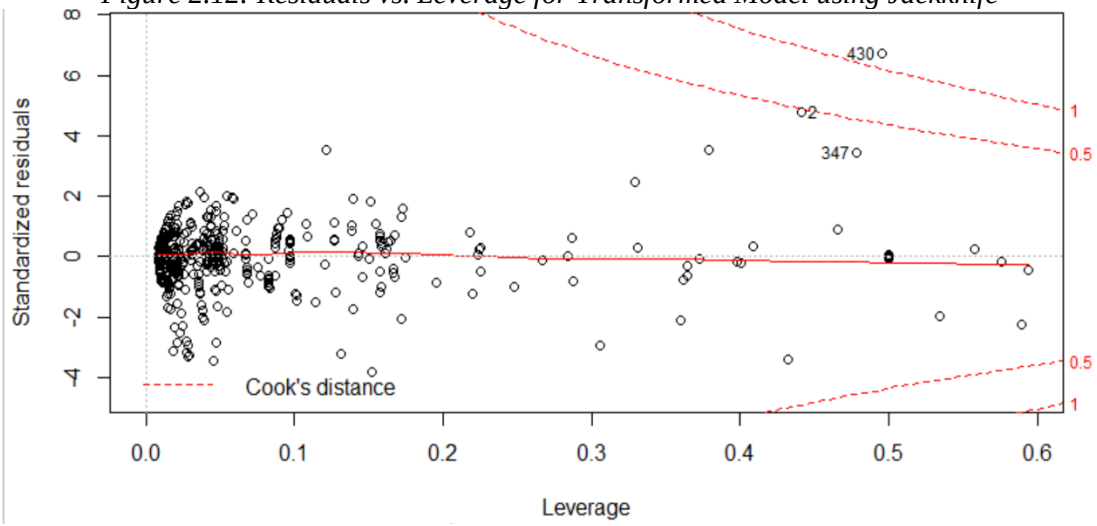
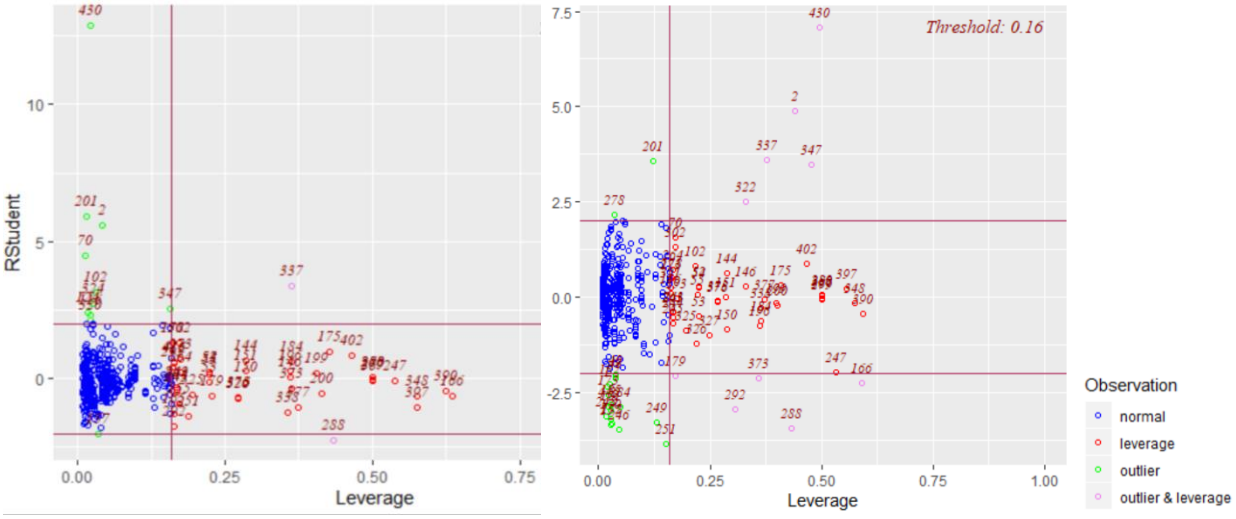


Figure 2.12 reveals several observations with high leverages. These points also threaten to violate the normality assumption despite jackknifing. Note that observation #430 appears exceeds the Cook's distance outlier threshold along the dotted lines; observations #2 and #347 are located very close to the 0.5 threshold. We also note that there were several observations that had a leverage of or greater than one (#176, 250, 285, 366). The high leverage values are a concern in that they had an extremely low CTR across all state fixed effect parameters. This means that these points, associated with an ad, often occurred less frequently than all other ads while also having a significantly lower CTR than the rest of the

observations across all states. An issue such as this often occurs when there is an error in the data entry or a typo. We accounted for the data entry for these high leverage points; all had the correct entry of data in terms of AdWordCount and CTR[15]. Again, issues arise we considered was that certain ads have large counts in the AdDuration variable, meaning these ads were to the public for long periods of time. As mentioned earlier in the methodology, we assumed for ads which had no end date (often left blank in the original data source) to have an assigned date of May 26th, 2017. Ads that may have received low CTRs while also having a very high number of days active may have created additional leverage points as well as outliers. Data containing many high leverage points suggests further analysis using other robust techniques since these points exert strong influence on the current model. Considering the observations that happen to exceed Cook's Distance, we can conclude that these are no doubt influential points. The inclusion of observations #2, #347, and #430 results in a weak relationship between CTR and the other variables. Using robust techniques with or without excluding these points could also result in a stronger relationship between our response and predictors.

Figure 2.13: Comparison Between Original Model (Left) & Jackknife Model (Right) in Studentized Residuals and Leverage



Once the data was resampled, we noticed a difference between the first model's studentized residuals and leverage plot compared with the Jackknifed models. Resampling with ***n*** - 1 provided a model seen in the original model plot of Figure 2.13 (left) in the top-left corner of the plot. As the jackknifed model (right) was replotted with the same 'jtools' Package[16] in R with the scale reformatted to account for the ***n*** – 1 resampling. Note, however, that outliers remain within our data, though more spread out through resampling. We concluded that since the outliers lie relatively close to the origin, the outliers present no real threat to the validity of our model. What was more disconcerting were the amount of points that happened to be both outliers and leverage points. Real-world data, such as the data gathered and available for use in this study, often has extreme values. This can occur when the data processing was either not completed or simply that there was data missing and could not be accounted for. The Jackknife model was not a successful attempt in model fit improvement. However, we do recognize that the jackknife provided a resampling necessary to reveal observations that contaminate the overall data. Many of the high leverage points in our data also reveal themselves as influential points, data points which have

[15] "Social Media Advertisements." U.S. House of Representatives Permanent Select Committee on Intelligence. Accessed April 28, 2020. https://intelligence.house.gov/social-media-content/social-media-advertisements.htm.
[16] Long, Jacob A., et al. "Analysis and Presentation of Social Scientific Data." Package 'jtools'. R-Studio, April 21, 2020. https://cran.r-project.org/web/packages/jtools/jtools.pdf.

unduly influenced our model. We opted to continue with a jackknifed resampled model while utilizing robust regression techniques to improve fit and meet model assumptions.

iii. Robust FE MM-Estimation Model

Before recomputing the model using robust techniques, several adjustments were implemented to the base formula of the model once the diagnostics of the jackknife fixed effects regression were concluded. The diagnostics revealed that the predictor AdDuration was not a reliable variable within our model due to the sporadic range in dates associated with the ads. First, AdDuration had questionable measurements of center and spread. The predictor had a median value of five (i.e. five days an ad was active) and a mean of approximately 151.12 indicating a high vulnerability to outliers. AdDuration also had a standard deviation of approximately 241.28 and an interquartile range (IQR) of 332 which also indicates that the spread of observations is considerably different from the mean value of the predictor. As indicated in the section concerning results from the jackknife regression, this had a considerable influence when regressed on the CTR response variable through the presence of high leverage points. Second, the density plot in Figure 3.1 indicates a probable discrepancy in the assumption of ads ending on May 26th, 2017. Note the red data points clustered to the far right of the plot: this clustering indicates that approximately 68 points with creation dates during 2015 have an active range of over 500 days. This is concerning since most of the data points (around 110) from 2015 have an active range of at most six days. A similar scenario exists for data points during 2016; approximately 69 points with creation dates during 2016 have an active range of between 150 and 500 days. Once again, most of the 2016 data points (around 162) have an active range of at most 54 days, the median active range being three.

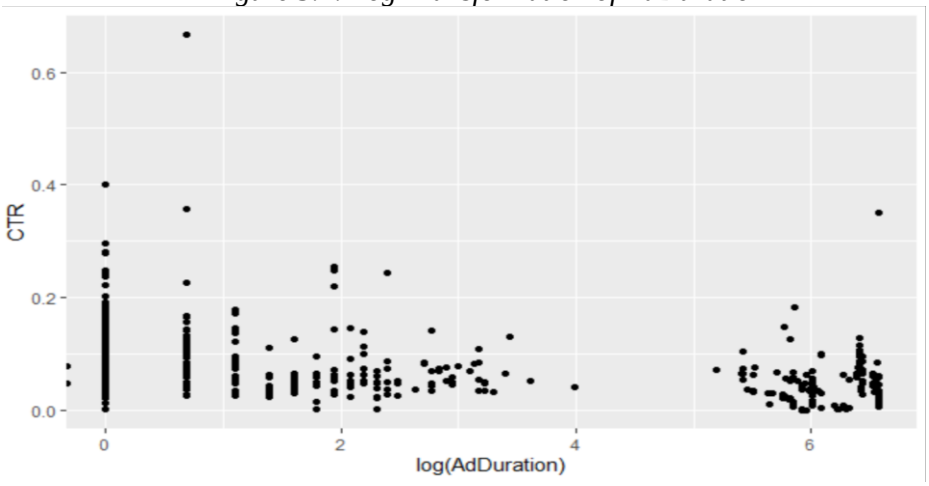
Figure 3.1: Density Plot of Duration of Ad and CTR

This new information provided the evidence in support of either removing or transforming the AdDuration variable. While a square transformation of the AdDuration was discussed because the transformation provided a more even spread of the data points, it was eventually decided that a square transformation of AdDuration was not the best approach in that it further harmed the assumptions of linearity and normality for a better Adjusted R-squared value. Being that the data of AdDuration had a wide range, a log transformation seemed the most logical decision to compress the data, potentially giving the model a better linear estimate, reducing skewness, and improving overall fit in the model without the compromise of regression assumptions[17]. It was important to use caution when applying the log transformation. While the log transformation is one of the most popular, it is prone to misuse. In this case,

[17] Feng, Changyong et al. "Log-transformation and its implications for data analysis." Shanghai archives of psychiatry vol. 26,2 (2014): 105-9. doi:10.3969/j.issn.1002-0829.2014.02.009

we wanted to reduce variability in the data by applying the transformation which could potentially do more harm than good. Fortunately, for the purposes of implementing a log transformation, the magnitude of the mean of the observations was relatively high, suggesting a promising transformation and an adequate compression of the data[18]. Figure 3.2 shows the log transformation of AdDuration with the response CTR, note that the data points have been compressed and are closer together than in EDA Figure 1.2 on page 7. Although the transformation has compressed the data, revealing a more distinct, downward facing pattern, there is still a discrepancy to the far right of the plot where the log(AdDuration) value range is around ~5 through ~6.5. This was unfortunately an issue that could not be resolved for two reasons. First, there was no logical transformation that could account for these extreme values that would also be easy to interpret. Second, the decision to use these values rather than removing them comes back to the original assumptions that ads with no end date were eventually taken down by around the same time the investigation took place. Additionally, the transformation also calculated two data points with a value of zero; log(0), or undefined values. Ultimately, the transformation did not provide a significant linear relationship with CTR, but it did provide visual evidence of which data were repeated offenders to the overall model fit. To proceed with the analysis, many of these data points would need to be removed. Once the transformation had been agreed upon, an additional analysis of the model was conducted with the removal of the variable for comparative purposes.

Figure 3.2: Log Transformation of AdDuration



The transformation of AdDuration would require us to revisit the Box-Cox transformation of our response variable $\boldsymbol{y}$ for the best lambda value for linearity. The computation provided a slight adjustment in the optimal value of lambda where $\boldsymbol{\lambda}^* = 0.424242\ldots$ for our proposed model. Therefore, being that $\boldsymbol{\lambda}^* \approx 0.42$, we utilized the transformation where $\boldsymbol{\lambda}^* \neq 0$, and applied $\boldsymbol{y}^{\boldsymbol{\lambda}}$, or $\boldsymbol{y}^{0.42}$, to the robust model. Once the response had the proper lambda, it was time to utilize robust regression methods.

We had originally intended to use OLS fixed effect linear regression because of the universal acceptance and computational simplicity. However, this method depends on many restrictive (and often, unrealistic) assumptions in the data. These assumptions include normality of error distribution, independency of the explanatory variables and error terms being homoscedastic. Seeing that we have outliers, the assumption of independent and identically distributed (*i.i.d*) errors for linear regression completely violates the dataset, resulting in bias and unreliable estimates of the model parameters[19][20]. Even through jackknifing, the model is simply resampled by $\boldsymbol{n} - 1$, keeping high leverage points and

[18] Ibid
[19] Simpson, J. R. and Montgomery, D. C. (1998). "The Development and Evaluation of Alternative Generalized M-Estimation Techniques". *Communications in Statistics - Simulation and Computation*, 27, 999–1018.
[20] Chatterjee, S. and Hadi, A.S. (2006). *Regression Analysis by Example.* 4th edition. New York: Wiley.

outliers in the dataset. The purpose in using robust regression methods was to provide an alternative to least squares regression by requiring less restrictive assumptions in our data. These methods attempt to suppress the influence of outlying cases to provide a better fit to the majority of the data. The outliers noted earlier within our data tended to tug the least squares fit too far in their direction, receiving much more "weight" than due credit and leading to distorted estimates of the model coefficients. In addition, this distortion resulted in these outliers to be very difficult to identify since their residuals are much smaller than they would otherwise be without the extreme points. The effects of outliers can be crucial for fixed effect model especially when multiple ***x***-outliers or high leverage points occur concentrated in the model[21]. Thus, the data becomes highly contaminated due to techniques in data transformation by non-robust centering procedure. We would opt for a Robust Within Group (fixed-effect) MM-estimator under the MM-centering procedure while simultaneously utilizing the resampling procedure through jackknifing. The utilization of the MM-centering method is introduced to bring back linearity into the transformed data and to provide high efficiency in estimates within the overall model. We provide a brief aside into the theoretical framework of the outlined procedure so the reader may understand the mathematics behind the MM-centering method and why it was the method chosen for our purposes.

MM-estimation is a common robust regression method falling into a class of estimators called M-estimators. Before exploring MM-estimation, it is important to understand the theory and application of M-estimators (M being “maximum likelihood”). M-estimators attempt to minimize the sum of a chosen function $\boldsymbol{\rho(\cdot)}$ which acts on the residuals since $\boldsymbol{\rho(\cdot)}$ is related to the likelihood function for a suitable assumed residual distribution. The original model follows

$$\mathbf{Y} = \mathbf{X}\beta_i + \epsilon^*,$$

where ϵ^* is assumed to be (multivariate) normally distributed with mean vector 0 and nonconstant variance-covariance matrix and define the reciprocal of each variance, σ_i^2, as the weight, $\omega_i = 1/\sigma_i^2$, then let matrix **W** be a diagonal matrix containing these weights:

$$\begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_i^2 \end{pmatrix} = \mathbf{W} = \begin{pmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_n \end{pmatrix}$$

Since each weight is inversely proportional to the error variance, it reflects the information in that observation. So, an observation with small error variance has a large weight since it contains relatively more information than an observation with large error variance which would follow a small weight. The weights in this data are known through the residual plot against the predictors exhibiting a megaphone shape. The regressing of absolute values of residuals against the predictors results in the fitted values of the regression being estimates of σ_i^2. As such, M-estimators are influenced by the scale of the residuals, so a scale-invariant version of the M-estimator is used:

$$\hat{\beta}_{\mathbf{M}} = \arg\min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{\epsilon_i(\beta)}{\tau}\right)$$

where τ is a measure of the scale. An estimate of τ is given by

[21] Bramati, M.C and Croux, C. (2007). Robust Estimators for the Fixed Effects Panel Data Model. *Econometrics Journal*, 10(3), 521–540.

$$\hat{\tau} = \frac{\text{med}_i|r_i - \tilde{r}|}{0.6745}$$

where $\tilde{\boldsymbol{r}}$ is the median of the residuals. The above is then minimized through an influence function and then through a numerical method called iteratively reweighted least squares (IRLS), which is used to iteratively estimate the weighted least squares estimate until a stopping criterion is met. Specifically, for iterations $\boldsymbol{t}$ = 0, 1, in the weighted least squares matrix…

$$\hat{\beta}^{(t+1)} = \left(\mathbf{X}^{\mathbf{T}}(\mathbf{W}^{-1})^{t}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}(\mathbf{W}^{-1})^{(t)}\mathbf{y},$$

where $(\mathbf{W}^{-1})^{t} = \text{diag}(\omega_1^{(t)}, \dots, \omega_n^{(t)})$ such that

$$\omega_i^{(t)} = \begin{cases} \dfrac{\psi\left(\dfrac{y_i - \mathbf{x}_i^t\beta^t}{\hat{\tau}^{(t)}}\right)}{\dfrac{y_i - \mathbf{x}_i^t\beta^t}{\hat{\tau}^{(t)}}}, & \text{if } y_i \neq \mathbf{x}_i^{\mathbf{T}}\beta^t \\ 1, & \text{if } y_i \neq \mathbf{x}_i^{\mathbf{T}}\beta^t \end{cases}$$

The usual mean centering procedures from M-estimation are often highly sensitive to high-leverage points (HPLs)[22]. As an alternative, the centering using the median (hence MM-estimation) is utilized due to the median's resistance to HPLs and outliers. Using the median centering procedure tends to be most effective when the data is contaminated with many HPLs and outliers resulting in a linearity to the transformed data after calculations[23(h)]. Just like within IRLS for M-estimation in weighting residuals by an average, the common functions used to weight MM-estimation are either the Tukey Bi-square or the Huber Method. The R-package "lmrob" calculates the complex algorithms that make up M-estimation briefly discussed earlier, automatically choosing the best of the two functions according to the makeup of the data for the creation of robust weighted residuals. The best way to view the method used is through plotting the "lmrob" weights by the actual residuals of the robust model, the same practice performed earlier to provide absolute values of residuals against the predictors resulting in the fitted values of the robust regression estimates σ_i^2, as the weights, $\omega_i = 1/\sigma_i^2$. Given the shape of the plot in Figure 3.3, the resampled robust model uses the Huber method to assign weights to the MM-estimate. The Huber function follows:

$$\rho(z) = \begin{cases} z^2, & \text{if } |z| < c; \\ |2z|c - c^2, & \text{if } |z| \geq c, \end{cases}$$

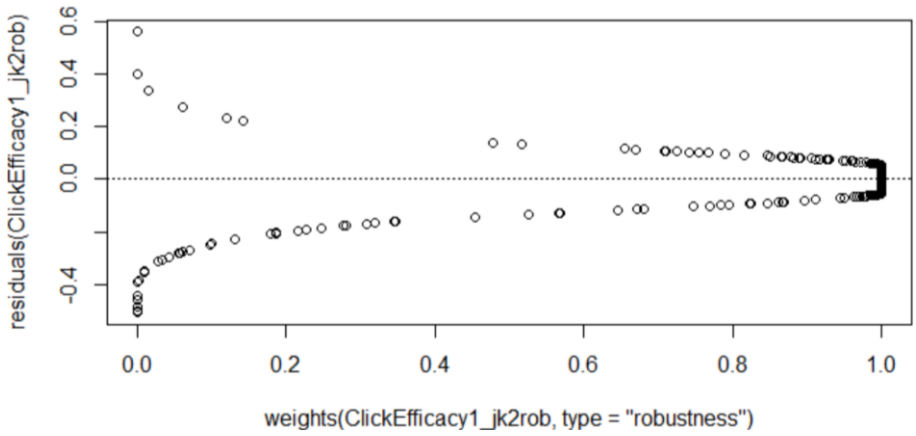
$$\psi(z) = \begin{cases} z, & \text{if } |z| < c; \\ c[sgn(z)], & \text{if } |z| \geq c, \end{cases} \quad \text{where } c \approx 1.345$$

$$\omega(z) = \begin{cases} 1, & \text{if } |z| < c; \\ \dfrac{c}{|z|}, & \text{if } |z| \geq c, \end{cases}$$

Figure 3.3: Weights of Robust Resampled Within Group Model

[22] Bramati, M. C. and Croux, C. (2007). Robust estimators for the fixed effects panel data model. *The econometrics journal*, 10(3):521–540.

[23] Maronna, R., Martin, R. D., and Yohai, V. (2006). *Robust statistics*, volume 1. John Wiley & Sons, Chichester. ISBN.



For the completion of MM-estimation, the algorithm "lmrob" computes S-estimates[24] and covariance to obtain the preliminary scale estimate, $\hat{\sigma}_n$. The scale estimate is then obtained by minimizing an M-estimate of scale. By fixing the scale estimate, the shape and location are re-estimated by a highly efficient M-estimator, down weighting the HLPs and outliers, and providing a major efficiency boost to the central model. Unfortunately, there were still problematic observations within the data, though this was to be expected. Robust MM-estimation pinpoints the problem data that happen to be contaminating the remaining data. These observations were 2, 52, 53, 54, 55, 145, 149, 166, 183, 199, 200, 293, 321, 322, 325, 326, 347, and 429. Removing these observations allowed for a more accurate assessment in the overall data. While there were remaining HLP and outliers in the data, the robust model accounted for these through the reweighting procedure described in the previous paragraphs. The algorithm in R calculates the robust regressions until the iteratively reweighted least squares (IRWLS) converges to an optimal weighting of the residuals. This often takes a few tries because the IRWLS may take multiple iterations, often decreasing iterations and eventually remaining fixed after a certain number of attempts (i.e. about 4-5 attempts). The model chosen for this analysis converges after 32 IRWLS iterations. While the robust modeling technique alone would prove sufficient through an improved model fit and highly significant predictors, resampling (through jackknifing) the data was used to further aid the overall model.

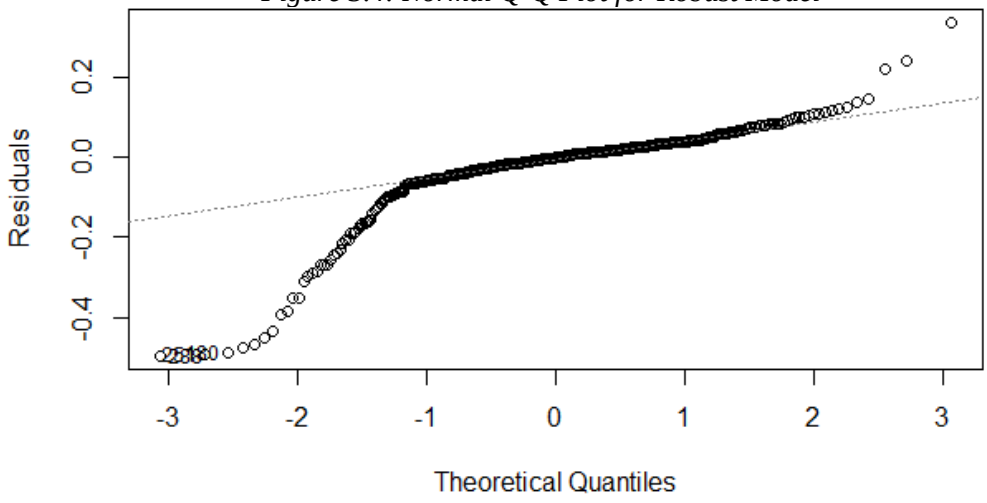
Notably, states such as Arizona (+22.23%), Idaho (+30.53%), New York (+5.36%), Ohio (+7.88%), Pennsylvania (+8.25%), Texas (+11.52%), and Washington (+17.09%) had estimates of greater than five-percent above the base category (i.e. Maryland). This indicated that Russian ads had a significant effect on these states likely due to the combination of higher CTR's[25], frequency in which ads appeared in these states, higher content, and both the release and duration of the content. Although not strictly a predictor, it is also likely that these states may have been affected more than the baseline due to their population size. In other words, the robust jackknifed model showed that the average effect of our predictors on CTR percentage after controlling for state fixed effects was greater than the baseline. The state Missouri (+3.38%) showed the average effect of our predictors on CTR percentage after controlling for state fixed effects was greater than the baseline by only a modest percentage. Many states, while significant, showed that the average effect of the predictors on CTR after controlling for state fixed effects was less than the baseline by a given percentage. These states include Arkansas (-24.54%), Massachusetts (-20.70%), New Jersey (-15.46%), New Mexico (-25.07%), Oklahoma (-21.41%), and West Virginia (-20.30%). The other states that have insignificant coefficients showed that the Russian ads have the same effect on them as the baseline (Maryland).

[24] S-estimates share the flexibility and nice asymptotic properties of M-estimators. The name "S-estimators" was chosen as they are based on estimators of scale.

[25] Note that CTR remains a change in terms of percentage. This is calculated within all model regressions as CTR*100 to account for the response already being in percentage terms when it comes to interpreting the model.

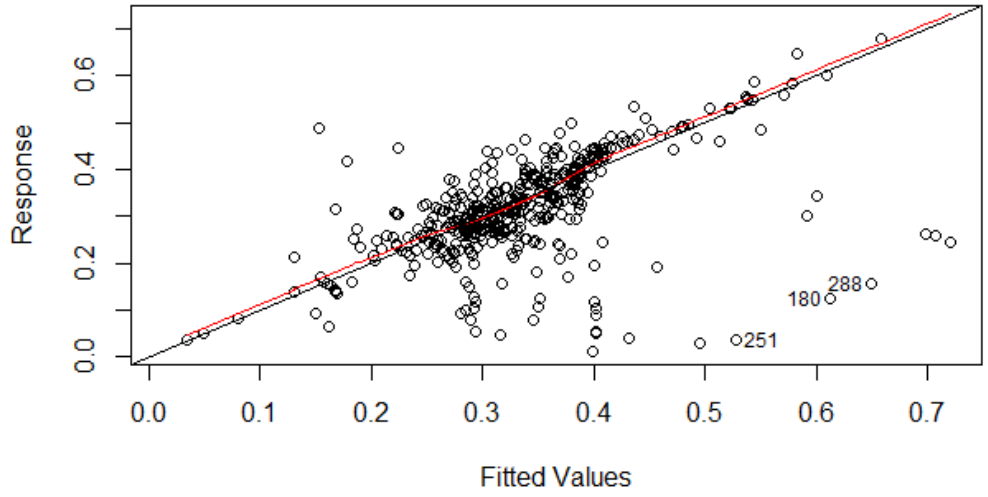
The new model revealed to be significant with observed standard errors remaining consistent across all states; the model's robust residual standard error approximately 0.0532. Additionally, the model returned a Multiple R-squared of ~ 0.71 and an Adjusted R-squared of ~ 0.68. Thus, after adjusting for sample size, degrees of freedom, and robust weighting, there is sufficient evidence to support that the predictors included within the robust jackknifed model account for approximately 68% of the sample variance of the percentage change in Click-Through-Rate (CTR). Considering the data, the model is highly significant given some of the restrictions observed throughout the study. A final analysis of the model was necessary to confirm that these findings are sound and did not violate any assumptions.

Figure 3.4: Normal Q-Q Plot for Robust Model



From observing Figure 3.4, the reader may again question whether the residual normality assumption has been met (similar issues within Figure 2.11 from the previous model). Many of the observational points fall along a straight line on the Q-Q plot apart from the many extreme observations in the lower tail region, suggesting our data is still skewed left. Again, it is likely that the predictor log_AdDuration is influencing the model due to its extreme values discussed earlier. Due to the model now operating under robust weighting of the residual values, it is normal in datasets such as these to see an increase of extreme residuals in either tail region because of the reweighting. Not all points find themselves close to the mean residual value. Thus, the model may still be interpreted linearly apart from the heavy tail.

Figure 3.5: Response vs. Fitted Values for Robust Model using Jackknife



The response vs. fitted visualization (Figure 3.5) is simply a scatter plot of the robust model's response variable, CTR, versus the all the predicted fitted values for the robust model. The ideal shape for this plot is having all points on as close to the black line as possible with an intercept of 0 and a slope of 1 (the line being about 45 degrees). The robust model indicates that the points are scattered in a diagonal band around the (0,1) line, the red line indicates that the robust model is linear. The points deviating greatly from this band indicate the outliers and HLP remaining after reweighting.

Figure 3.6: Standardized Residuals vs. Robust Distances for Robust Model using Jackknife

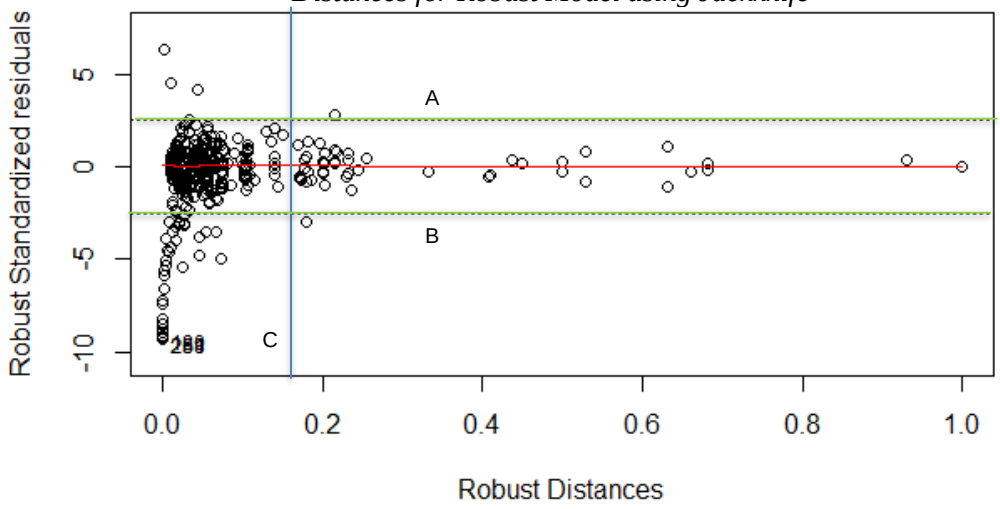


Figure 3.6 depicts points in which are influential, or which points influence the model. The reader will observe that the points that appear outside of the two dotted lines (Cook's distance) are like the ones found in Figure 3.4. These points should be excluded from model estimation with care, as they do influence the regression result. Recall that other observations have already been removed to decontaminate the data for a better overall picture, removing any more points ultimately removes data pertinent to the analysis. Interpretation of the standardized residuals vs. robust distance plot divides the value range in four regions marked by the green and blue lines. Points within the two green lines and to the left of the blue lines are regular points. Outliers reside to the left of the blue line and outside of the

green lines. The leverage points for the model happen to be within the two green lines and to the right of the blue line. Finally, the points that are of major importance are the outlier and leverage points which occur to the right of the blue line and above (line A) and below (line B) the green lines within the plot. Fortunately, the robust estimation procedure removed most of these points from the model. With the residuals reweighted through MM-estimation, the model provides a more representative description of the overall observations within the data with little influence by extreme observations unlike the non-robust model[26].

Figure 3.7: Residuals (Top) and Square root of Residuals (Bottom) vs. Fitted Values for Robust Model using Jackknife

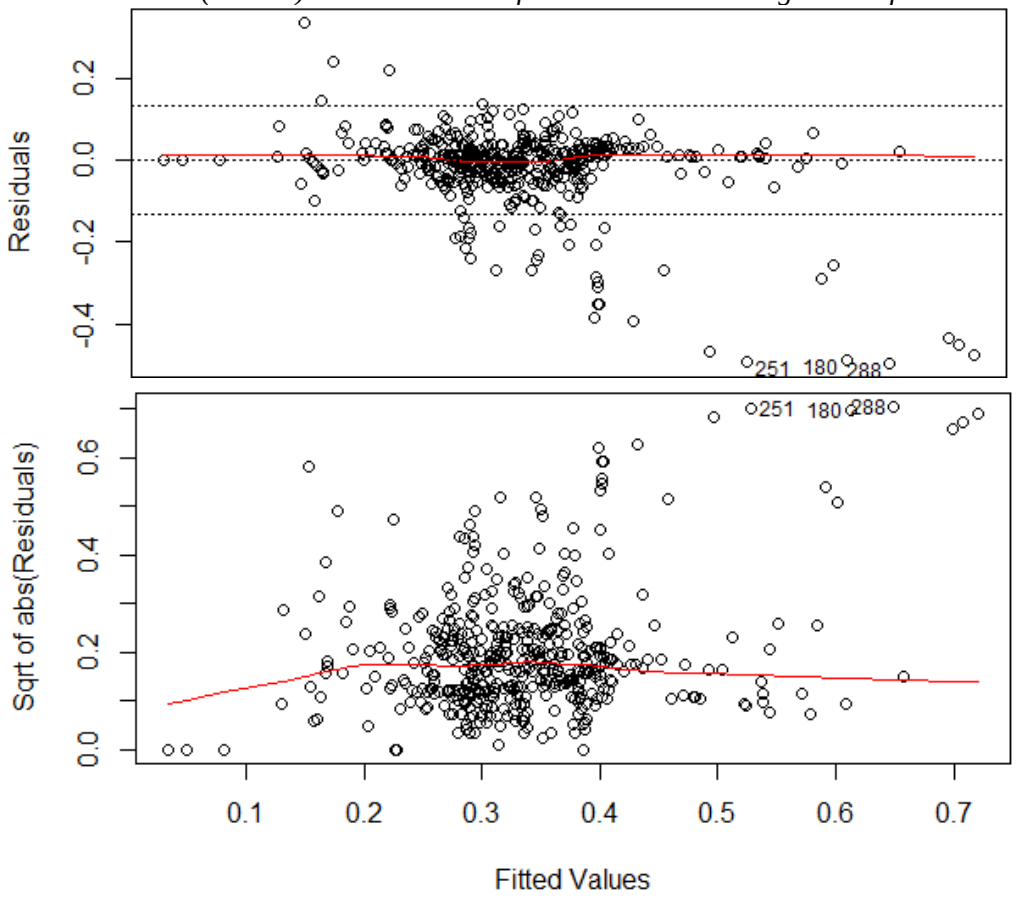


Figure 3.7 (top graph above) indicates the linearity of the model after the robust weighting of the residuals. The robust resampled model has a more consistent residual vs. fitted value spread than the jackknife model and has adjusted for the fanning out pattern observed in Figure 2.9. The red line also indicates the observations follow a linear pattern. Therefore, the points in Figure 3.7 have not severely violated the functional form enough to raise significant issues in the linearity assumption, so a linear interpretation is acceptable for the robust resampled model. The robust model still reveals specific observations (#180, 251, and 288) that have consistently interfered with the regression assumptions of linearity, as well as normality and homoscedasticity since the first two model iterations. The following graph (bottom) in Figure 3.7 indicates the level of constant variance, or homoscedasticity by scale location. The scale location plot does suggest some non-linearity in the square root of the standardized residuals plotted against the fitted values. The plot also indicates that the spread of magnitudes seems to be lowest in the fitted values close to 0.1 and 0.5 – 0.7, the highest in the fitted values between 0.2 and

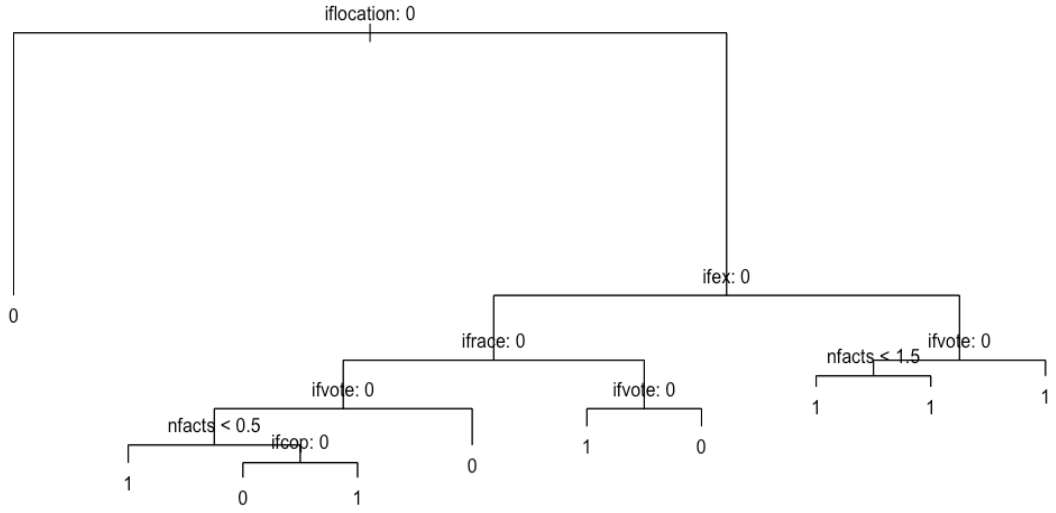
[26] Brenner, Johannes et al. (2019). SWCalibrateR: Interactive, Web – Based Calibration of Soil Moisture Sensors. *Journal of Open Research Software*. 7. 10.5334/jors.254.

0.4. This would suggest some level of heteroskedasticity in the reweighted data. Formally tested with the Breusch-Pagen (BP) Test for homoskedasticity. For the robust reweighted model, the null hypothesis of homoskedasticity is rejected with a p-value of approximately 0.0006. In comparison to the Jackknifed model (p-value = 0.0002), the robust model performs slightly better. For this reason, the robust model cannot be fully considered to have met the homoscedasticity assumption. Although, due to the inconsistency of the data in question, the level of heteroscedasticity could be considered acceptable. Recall in Figure 2.10 that there was an odd clustering of standardized residuals circled in red between the fitted values of ~ 0.25 and ~ 0.41 likely attributed by the AdDuration variable. Both the log transformation of AdDuration and the robust weighting of the residuals has removed this clustering.

VII. Propaganda Detection with Classification Models: Analysis & Results

It was decided that a classification tree model would provide an adequate prediction accuracy and allow for interpretation. Other classification models have been found to show greater test data accuracy, but the ability to interpret the relative importance of predictors was a priority for this study. This model grew a tree by binary recursive partitioning using the indicator variable = 1 if propaganda, 0 otherwise. Weights and all other settings were default according to the 'tree' package in R. The split criterion was Gini Index, as it is more sensitive to node purity than the classification error rate. Accordingly, the following tree diagram was generated. Propaganda was coded as 1, so the first node reads "if the ad did not target a location, then the ad was classified as an organic/ProPublica post (the first terminal node to the left).

Figure 4: Propaganda Detection Classification Tree Model



The first predictor to stand out was location. It appears that location is the most important variable for setting node purity, followed by the presence of exclamation points, and others. One interesting subcluster generated contains ads that referenced race (the only internal node labelled ifrace:0), then mentioned voting. This split reveals that among ads discussing race, voting is the next most important predictor of propaganda. Also, ads with fewer facts were always predicted to be propaganda. Finally, all ads with lone exclamation points were classified as propaganda. This could reflect the IRA's penchant for commands like "Follow!" or "Subscribe!".

Regarding the model's prediction accuracy, overall, it correctly predicted whether a test-set ad was propaganda 85.12514% of the time. Of greatest interest, however, is how often it successfully detected that an ad was indeed propaganda: *the model was 87.99776% accurate when the ad was propaganda*. In other words, if you showed the classification tree algorithm 100 pieces of propaganda, we would expect it to correctly label 87 of them as 'propaganda'. The true positive rate decreased slightly

upon the application of a random forest model, to *86.26098%*. The random forest randomly sampled $\sqrt{7}$ = 2.64575 variables per split since we had 7 predictors. Based on node sizes 1-20, we found that the best size was 15. Resampling from the ProPublica data 40 times, the random forest model returned the following distribution of true positive rates:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.8305	0.8487	0.8571	0.8626	0.8778	0.9072

In every model attempted, the true positive rate drove up the overall error rate. Given a ProPublica ad, the random forest and classification tree correctly identified it 87.90723% and 82.16965% of the time, respectively. In all, the classification tree excels at identifying a propaganda advertisement, whereas a random forest model is better for predicting legitimate political ads.

VIII. Potential Objections

A possible objection to how we manipulated our data would involve the end date of the advertisements that were unknown, and our assumption that they were taken down on the day that Facebook gave the House Committee the report. In other words, we assumed that the end date for all missing "end date" values was the last day Facebook compiled the ads before giving it to the House Committee.

We could not assume normality of individual ads concerning their distribution of clicks and impressions across different states which is why we created individual rows for a single ad which had multiple locations in order to have ad information by individual states. This could be thought of as manufacturing points since we could be counting multiple variables more than once. The problem with this is that we have a certain number of clicks for a specific ad and we cannot assume homogeneity across the states; citizens of one state may react differently when seeing the same ad that another citizen from a different state causes clicks to also be different between the states. The way we addressed this problem was by creating the Click Through Rate of the ad which is a rate and thus should be the same across heterogeneous states.

Not all our regressions to date include ads that appeared nationally. These ads could have appeared in any of the states for which we have dummy variables, but we are unable to train the dummies on those clicks since we do not know where clicks of a national-level ad occurred locally. In this way, our dummy variables may not represent the true change in effectiveness when a state is targeted by a national ad.

For comparing the effectiveness of the Russian Facebook ads, we wanted to gather either not political or simply generic ad data from Facebook between the years of 2015 and 2017. Unfortunately, Facebook does not publish ad statistics from before May of 2018. In fact, data available without special request only has a range of 90 days prior from the request date. Thus, we could not get an accurate date range for ads during this period that were not from the IRA. We have decided to use ad statistics from May 2018 to May 2019 and calculate the effectiveness of these ads to compare with the 2015 and 2017. This comparison may not be as significant due to various changes in the confounding factors we are measuring between the 2015-2017 period and the period of 2018-2019. This happens to be the only date range of non-political ads that we could use to compare with the Russian ads are the 2018-2019 ads. So, we will use these ads as a control group.

States are very heterogeneous which makes running regression between them difficult due to multiple variables that could be omitted. The reason why we are conducting a fixed effects model is so we can account for this and control for the major differences between states such as politics, demographics, income, and population size.

One glaring objection to the classification model is the nature of the ProPublica data. These ads are very much a convenience sample, as ProPublica solicited its users to provide information on ads they had purchased on Facebook. This means the ProPublica data is not representative of the population of all non-propaganda political advertisements which ran on Facebook at the time. Therefore, one should not say that the results of the classification model reflect the true nature of all political ads on Facebook. For reference, in the third quarter of 2019, Facebook announced that more than 7 million advertisers used their platform.[27] Each advertiser surely had more than one ad running. Our classification model saw only 3516 ProPublica advertisements, sampled from the set of 158,117: a small fraction of the ads probably running on Facebook at the time. In the future, researchers might consider web scraping a random sample of ads from Facebook's Ad Library, using code like that written by NYU researchers.

This study's classification techniques could be questioned based on time as well. Specifically, almost all the organic political advertisements (not propaganda) ran on Facebook 1-2 years after propaganda ads had concluded. The first propaganda ad started in 2015, and the last ended in August 2017, while most of the ProPublica ads ran from fall 2017 to 2019. Considering the fleeting nature of political issues, the topics discussed by organic political data in the span of the IRA's campaign may have been much different than the topics discussed by the ProPublica data only a year later. This could damage the validity of our thematic predictors, such as whether the ad contained 'racial' themes. It is very possible that real political ads running from 2015-2017 discussed race just as often as the propaganda. In this case, said predictor would be invalid. We would need Facebook to disclose earlier organic political advertisements if we are to prove such a predictor valid.

IX. Applications

The most important application of this research is to allow committees that are designed to enforce a private, safe, and most importantly, free election, to understand which states are most likely to be affected by divisive ads that countries like Russia used in our 2016 election. By knowing which locations are the ones that are most vulnerable, election committees, or intelligence committees, will be able to focus down on states that are most susceptible. This will not only maximize resources such as time and effort, by increasing the probability of finding the fake ads and have them be taken down, but it will also make sure that the states which are the most susceptible to the ads get the protection that they deserve in order to allow for fair election progress where the citizens make up their own mind about who will be best for them and the country, and not have fake information fill up their social media feed. The jackknife regression model has applications in showing how information through advertising is spread. More importantly, this algorithm provides detail on which ads are targeting and spreading within specific locations based on the content of the ad. Companies involved in social media or marketing can take this information to understand the differences between how propaganda and actual sponsored advertisements affect different populations. Companies can then measure the outcome of such ad campaigns regarding the social actions of these populations and whether they are ethically sound. The model also has applications in many other industries such as business analytics, political research, and economics in that it provides significant detail regarding how ads, propaganda or legitimate, could potentially affect certain populations differently by the transmission of content. Our model can be further improved by companies who have the resources to gather more data regarding target demographics of these locations that were affected by Russian Propaganda. Having more concrete data with less inconsistencies would allow our model to be used for predictive purposes as well.

The classification model is especially practical in that anyone can pull up a Facebook ad, enter their location and race, copy and paste the ad's text and it will tell you whether it thinks the ad is propaganda. This may be advised because of its low false positive rate, but we hope others will build upon this progress and build a model that can even more reliably identify disingenuous Facebook

[27] Clement, J. "Facebook Active Advertisers 2019." Statista, January 30, 2020. https://www.statista.com/statistics/778191/active-facebook-advertisers/.

advertisements. Facebook and government officials (Such as the U.S. Cyber Command or the NSA[28]) could potentially employ this algorithm with a greater degree of accuracy since they have access to ad creation hour time stamps and can verify sponsors. In the past, they have easily identified propaganda by looking at receipts, since the Russians bought ads under their actual agency name and paid in rubles. After being exposed in 2017, however, they probably shifted their tactics and have begun funneling money through shell companies pretending to be legitimate U.S. political interest groups. In addition, the number of advertisers on Facebook has increased dramatically in recent years (from 3 million in 2016 to 7 million in late 2019[29]), so it is increasingly difficult to verify the legitimacy of each firm advertising on Facebook. This means Facebook and others will have to rely more heavily on propaganda characteristics and traits (thematic content, syntax, target demographics, etc.) if they are to successfully root out propaganda. The IRA will also likely expand its campaigns to include other social media platforms (They actually recently upgraded their office building), so unless these platforms gather detailed metrics on each advertisement, they will have to rely on an advertisement's facade to combat disinformation.

In the future, researchers might improve true positive prediction rates by adding predictors like post hour created (see Boyd p. 2), run time of the ad, number of totally capitalized words, or by increasing the number of target locations to detect. Future research should also select locations in which the Russians have historically invested much time and money (this could be measured in a rubles per minute online variable). The target location predictor could greatly benefit from such a change. Russian agents have historically masqueraded as individuals promoting Southern identity/nationalistic themes, so a predictor indicating the presence of these themes could increase node purity. This research only examined ad text (the text you normally see above a picture, typed by the poster), but page post name and image text should also be searched for thematic content data. Additional data could be gathered from NYU's scraped set of political advertisements. These were not included in the present study due to merging difficulties and lack of message data. Finally, it could be valuable to build a model stating the probability that a given ad is propaganda. This way, Facebook users can decide for themselves whether a statement like "there is an 87.99776% chance that this ad is Russian Propaganda" is cause for concern.

X. Conclusion

The results from the Jackknifed model indicated that Russian ads, given a specific set of predictors, affected different parts of the country more effectively than others. Through our analysis, this was no doubt a strategic endeavor conducted by the Russian Internet Research Agency (IRA) to create upheaval in the United States political arena and to also create chaos in the diverse social dynamics scattered across different populations across the United States. Our regression analysis using transformation and jackknifed methods gathers that the IRA purposely targeted states with higher populations, often including major cities, as focal points for the dissemination of propaganda generating controversial topics or falsified facts pertaining to current affairs within the United States during a range of three years which included a presidential election year. Additionally, states were targeted with specific ads focused on specific content, whether false or true, that were relatable to the political and/or sociological ideologies of those individuals within that state. While our analysis raised potential objections primarily concerning the data available for us to use, the analysis provides an important secondary argument in data collection. An important thing for any statistician, or individuals using the applications of statistics, is that an analysis is only as good as the data it uses. It seems as if certain data simply cannot be collected from Facebook's website. This creates a major problem for both guaranteeing the personal data security of users and overall advertising transparency provided by the social media

[28] Nakashima, Ellen. "U.S. Cyber Command Operation Disrupted Internet Access of Russian Troll Factory on Day of 2018 Midterms." The Washington Post. WP Company, February 27, 2019. https://www.washingtonpost.com/world/national-security/us-cyber-command-operation-disrupted-internet-access-of-russian-troll-factory-on-day-of-2018-midterms/2019/02/26/1827fc9e-36d6-11e9-af5b-b51b7ff322e9_story.html.

[29] Clement, J. "Facebook Active Advertisers 2019." Statista, January 30, 2020. https://www.statista.com/statistics/778191/active-facebook-advertisers/.

conglomerate. Although our model provides a general conclusion about how ads were viewed and spread throughout the states, the lack of more specific quantitative data gives our analysis limitations in how the propaganda affected the political outcome of 2016, which was our original intention of this study. We can also only speculate with moderate confidence that the content of the ads, whether political or socially driven, entices audiences to click more on these ads and thus being influenced by propaganda from a foreign actor.

While the first model does well in explaining the way IRA Propaganda spreads and where, we found that classification models can accurately predict whether an ad is propaganda. This study also concluded that predictors like target location, number of exclamation points, and thematic predictors (racial topics) are very important for separating observations into pure groups. Although the best classification model is only 37.99% better than random guessing, it appears that propaganda detection rate can be improved upon the addition of informative predictors and given augmented classification models. Finally, true positive rates could improve with the application of a boosted model or linear support vector machine. Either way, a new model should employ a cross-validation style technique to study whether prediction rates change given different subsamples of the ProPublica data.

In terms of our original political inclination going into this study, we were not able to address this given the limited number of ads that each state had and the complexity of politics. Since politics varies widely across states and even within states, we were not able to have sufficient and appropriate data to analyze the effects of fake Russian ads within individual states. The closest we were able to get was finding the most common words within states but there were not sufficient ads for almost all states to establish a theme for each individual state. These common words did show that there are states that were targeted with specific themes such as Minnesota with Islamic and Muslim related ads, or Pennsylvania with Trump and coal related ads. As the coming elections approach, it is almost certain that the Russians will once again try to interfere in our elections in one way or another. So, it is possible that, with Facebook's willingness to help, there will be many more ads to analyze nationally but more importantly also by state and local levels which will help further studies try and tackle this question. By having the specific ad and its information in more local, or state, levels the data will be better suited to try and tackle the relationship between the type of ads that the Russian created to interfere in our election and the relationship that they have with politics. This could answer questions such as if the Russians ads focus mostly on creating division within or between states, if they take on both sides of politics (Republicans and Democrats) or if they mostly back one side, etc.

Notes:

a. It should be noted that while we did not find any significance in voter turnout based on the data. Others, such as Spangher et al., may have found significance using different predictors or other statistical modeling techniques that we have not considered, nor have the time to implement. This could be a follow-up task within the parameters of our analysis if given the opportunity to extend our period of data analysis.

b. Others, notably Dutt, have measured effectiveness at the state level using ad language. However, we hope to measure it at the city, state, and national level. Like Dutt, we considered conducting Wilcoxon tests with Bonferroni adjustments or bagging. While not performed in this study, it is something we could explore in future studies.

c. Note that variables may have been renamed from the original dataset to fit our needs in R-Studio.

d. Not all states had sufficient information to be included in this map, so some states have NA and 0% CTR. Similarly, Delaware only has “marchprotests” as its ads that were used in our analysis only included that word.

e. We believe that choosing this date in January would not be significantly different from choosing a random date within this month when running this variable in our model. Many of the ads that we chose to have this end date already exceed triple digits in days, so adding at most 30 days will not significantly change the predictive model.

f. Having individual specific intercepts $\boldsymbol{\alpha_i}$, $\boldsymbol{i = 1, \dots, n}$, where each of these can be understood as the fixed effect of entity $\boldsymbol{i}$, the model will have fixed effects on the response variable. For the formula above, $\boldsymbol{\alpha_i}$, are entity-specific intercepts that capture heterogeneities across entities and are denoted as $\boldsymbol{\gamma_n D_{ni}}$ (Econometrics with R, 10.3).

g. Note that CTR remains a change in terms of percentage. This is calculated within all model regressions as CTR*100 to account for the response already being in percentage terms when it comes to interpreting the model.

h. Note that the following example provided by Maronna, R. et al and Bramati, M. C. et al is within the context of time-series modelling. However, similar approaches provided by MM-centering, such as the one we have done, may be used without time-series parameters to equal effect.

XI. References

"An Analysis of Transformations." Journal of the Royal Statistical Society. Accessed April 28, 2020. https://www.ime.usp.br/~abe/lista/pdfQWaCMboK68.pdf.

Boyd, Ryan L, et al. "Characterizing the Internet Research Agency's Social Media Operations During the 2016 U.S. Presidential Election Using Linguistic Analyses." University of Texas at Austin, n.d. Accessed April 28, 2020.

Box, G. E. P. and Tidwell, Paul. "Transformation of the Independent Variables". *Technometrics*. Vol. 4, No. 4 (Nov., 1962), pg. 534 (20 pages). https://www.jstor.org/stable/1266288?seq=4#metadata_info_tab_contents

Bramati, M.C and Croux, C. (2007). Robust Estimators for the Fixed Effects Panel Data Model. *Econometrics Journal*, 10(3), 521–540.

Brenner, Johannes et al. (2019). SWCalibrateR: Interactive, Web – Based Calibration of Soil Moisture Sensors. Journal of Open Research Software. 7. 10.5334/jors.254.

Chatterjee, S. and Hadi, A.S. (2006). *Regression Analysis by Example.* 4th edition. New York: Wiley.

Clement, J. "Facebook Active Advertisers 2019." Statista, January 30, 2020. https://www.statista.com/statistics/778191/active-facebook-advertisers/.

Dutt, Ritam, et al. "'Senator, We Sell Ads': Analysis of the 2016 Russian Facebook Ads Campaign." Indian Institute of Technology Kharagpur, India, n.d. Accessed April 28, 2020.

Feng, Changyong et al. "Log-transformation and its implications for data analysis." Shanghai archives of psychiatry vol. 26,2 (2014): 105-9. doi:10.3969/j.issn.1002-0829.2014.02.009

"FBPoliticalAds." Facebook Archive. shikhar394. Accessed April 28, 2020. https://github.com/online-pol-ads/FBPoliticalAds/blob/master/docs/Facebooks-archive.pdf.

Fox, John, et al. "Companion to Applied Regression." Package 'car.' R-Studio, March 11, 2020. https://cran.r-project.org/web/packages/car/car.pdf.

Howard, Philip N, et al. "The IRA, Social Media and Political Polarization in the United States, 2012-2018," n.d. Accessed April 28, 2020.

"Impressions." Facebook Business Help Center. Accessed April 28, 2020. https://www.facebook.com/business/help/675615482516035.

Long, Jacob A., et al. "Analysis and Presentation of Social Scientific Data." Package 'jtools'. R-Studio, April 21, 2020. https://cran.r-project.org/web/packages/jtools/jtools.pdf.

Maronna, R., Martin, R. D., and Yohai, V. (2006). Robust statistics, volume 1. John Wiley & Sons, Chichester. ISBN.

Nakashima, Ellen. "U.S. Cyber Command Operation Disrupted Internet Access of Russian Troll Factory on Day of 2018 Midterms." The Washington Post. WP Company, February 27, 2019.

https://www.washingtonpost.com/world/national-security/us-cyber-command-operation-disrupted-internet-access-of-russian-troll-factory-on-day-of-2018-midterms/2019/02/26/1827fc9e-36d6-11e9-af5b-b51b7ff322e9_story.html.

ProPublica. “Political Advertisements from Facebook.” ProPublica Data Store, March 19, 2019. https://www.propublica.org/datastore/dataset/political-advertisements-from-facebook.

Ripley, Brian, et al.“Support Functions and Datasets for Venables and Ripley's MASS.” Package ‘MASS.’ R-Studio, April 26, 2020. https://cran.r-project.org/web/packages/MASS/MASS.pdf.

Scott, David. Tukey Ladder of Powers. Accessed April 28, 2020. http://onlinestatbook.com/2/transformations/tukey.html.

Simpson, J. R. and Montgomery, D. C. (1998). “The Development and Evaluation of Alternative Generalized M-Estimation Techniques”. *Communications in Statistics - Simulation and Computation*, 27, 999–1018.

“Social Media Advertisements.” U.S. House of Representatives Permanent Select Committee on Intelligence. Accessed April 28, 2020. https://intelligence.house.gov/social-media-content/social-media-advertisements.htm.

Spangher, Alexander, et al. “Analysis of Strategy and Spread of Russia-Sponsored Content in the US in 2017.” Carnegie Mellon University, n.d. Accessed April 28, 2020.

XII. Appendix

A.1: OLS Jackknife Regression Results

Predictors	**Click-through Rate (CTR) Estimate w/ 95% CI**
bootEstParamAdWordCount	-9,194.3090***…….(-13,852.9000, -4,535.7140)
bootEstParamAdDuration	26,131.9600***…….(7,508.1080, 44,755.8100)
bootEstParamsq_Ad_Freq_state	83.2108……..(-69.0512, 235.4729)
bootEstParamElectYear	5.0707……(-32.9705, 43.1120)
Alabama	0.0671……(-0.0922, 0.2263)
Arkansas	-0.2454**…….. (-0.4331, -0.0577)
Arizona	0.2223***……..(0.1075, 0.3371)
California	-0.0282……...(-0.0742, 0.0178)
`District of Columbia`	0.0318…….……...(-0.0236, 0.0873)
Delaware	-0.0126…….…..(-0.1476, 0.1225)
Florida	0.0011……...…..(-0.0450, 0.0472)
Georgia	0.0094…….(-0.0355, 0.0543)
Iowa	-0.0287……(-0.3306, 0.2732)
Idaho	0.3053***……..….(0.1653, 0.4453)
Illinois	-0.0422……...(-0.1103, 0.0260)
Kansa	0.0538……(-0.0269, 0.1345)
Louisiana	-0.0147……...(-0.0707, 0.0413)
Massachusetts	-0.2070**……...(-0.3962, -0.0178)
`New York`	0.0536***……..(0.0216, 0.0855)
Michigan	0.0291……(-0.0162, 0.0744)
Minnesota	0.0293……(-0.0421, 0.1006)
Missouri	0.0338*……..(-0.0050, 0.0726)
Mississippi	-0.0282……..(-0.1450, 0.0887)
`North Carolina`	0.0407……(-0.0340, 0.1154)
`New Jersey`	-0.1546**……..(-0.3025, -0.0067)
`New Mexico`	-0.2507***……(-0.3691, -0.1324)
Ohio	0.0788***…….(0.0418, 0.1158)
Oklahoma	-0.2141***……(-0.3588, -0.0693)
Pennsylvania	0.0825*…….(-0.0143, 0.1793)
`South Carolina`	-0.0335……..(-0.1115, 0.0445)
Tennessee	0.0627……(-0.0535, 0.1788)
Texas	0.1152***……..(0.0771, 0.1534)
Virginia	-0.0252……..(-0.0957, 0.0453)
Vermont	0.1701…….(-0.1223, 0.4626)
Washington	0.1709***…….(0.0949, 0.2468)
Wisconsin	-0.0247……..(-0.0841, 0.0348)
`West Virginia`	-0.2030*……(-0.4115, 0.0055)
Constant	3.3178***……..(1.6834, 4.9522)
Observations	476
R2	0.3233
Adjusted R2	0.2611
Residual Std. Error	0.0955 (df = 438)
F Statistic	5.6556*** (df = 37; 438)
	Notes: ***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level. *CTR*100 for percentage and transformed with lambda of .38*

A.2: Robust Jackknife Regression Results

Predictors	**Click-through Rate (CTR) Estimate w/ 95% CI**
bootEstParamAdWordCount	-18,767.72***……..…(-21,428.84, -16,106.60)
bootEstParamlog_logAdDuration	1,484.5230***……..(1,325.5080, 1,643.5380)
bootEstParamAd_Freq_state	753.9083***…….....(584.5852, 923.2314)
bootEstParamElectYear	81.2453***……..….(53.0445, 109.4460)
Alabama	0.2216***……...…..……(0.0876, 0.3556)
Arkansas	-0.7383***……..……..(-0.9132, -0.5634)
Arizona	0.1510***……..……...(0.0868, 0.2153)
California	0.0061……...(-0.0232, 0.0354)
`District of Columbia`	0.0157……..... (-0.0176, 0.0490)
Delaware	-0.0243……..…… (-0.0962, 0.0475)
Florida	0.0392**……...……... (0.0088, 0.0695)
Georgia	0.0166……..…….... (-0.0165, 0.0497)
Iowa	0.1767*……..…….. (-0.0226, 0.3759)
Idaho	0.2685***……..…….. (0.1937, 0.3434)
Illinois	-0.0127……...…..…….. (-0.0569, 0.0316)
Kansas	0.0316…….. (-0.0132, 0.0765)
Louisiana	0.3211***……..…….. (0.2586, 0.3836)
Massachusetts	-0.2135***……..…….. (-0.3135, -0.1136)
`New York`	0.0539***……..…….. (0.0313, 0.0765)
Michigan	0.0486***……..…….. (0.0207, 0.0765)
Minnesota	0.0819***……...…….. (0.0312, 0.1327)
Missouri	-0.0180……...…….. (-0.0440, 0.0081)
Mississippi	-0.6458***……..…….. (-0.8191, -0.4725)
`North Carolina`	0.0373…….. (-0.0080, 0.0827)
`New Jersey`	-0.4985***……..…….. (-0.6001, -0.3969)
`New Mexico`	-0.1818***……..…….. (-0.2602, -0.1034)
Ohio	0.1139***……...…….. (0.0899, 0.1379)
Oklahoma	-0.3767***……..…….. (-0.5597, -0.1938)
Pennsylvania	0.0636**…….. (0.0057, 0.1215)
`South Carolina`	0.0006…….. (-0.0472, 0.0484)
Tennessee	0.0807**…….. (0.0070, 0.1543)
Texas	0.1081***……...…….. (0.0831, 0.1331)
Virginia	0.0637***……...…….. (0.0177, 0.1096)
Vermont	0.3532***……...…….. (0.1910, 0.5153)
Washington	0.0894***……..…….. (0.0442, 0.1346)
Wisconsin	-0.2027***……..…….. (-0.2430, -0.1624)
`West Virginia`	-0.4415***……..…….. (-0.5591, -0.3239)
Constant	12.7685***……..…….. (11.5051, 14.0320)
Observations	456
R2	0.7060
Adjusted R2	0.6799
Residual Std. Error	0.0532 (df = 418)
	Notes: ***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level. *CTR*100 for percentage and transformed with lambda of .42*