The 2016 Presidential Election: Finding the Relationship Between Public Interest and Controversy

Kristen Bystrom, Tessa Ramburn, Barinder Thind

301252960, 301205956, 301193363

Abstract

The 2016 presidential election was unlike any other. We had two of the most controversial candidates in American history running simultaneously which made for a never-ending news cycle filled with scandals that only encouraged more and more public opprobrium. In this paper, we explore the magnitude of these controversial news stories in terms of the curiosity they sparked among people, by looking into a data set containing information about Google searches over the past year. This data set contains relative numbers on how often a particular term ('Hillary' or 'Trump') was searched. We log-transformed the data and took the first and second differences to reach stationary processes and estimate parameters. We also used a timeline of events[1] to make inferences about trends in the data. We selected an ARIMA (0,1,1) for the search on Hillary Clinton and an ARIMA (0,2,1) for the one on Donald Trump.

Keywords: Time Series, Log Transformation, Election, Trump, Hillary

I. Introduction

The US 2016 election, in addition to being one of the most controversial, also became associated with inadequate projections and failed statistical models. In the future, the use of alternative data sources, and the inferences that we can pull from these, may help to produce more accurate predictions. Throughout this project, we hoped to do two things:

- After collecting data, fit and evaluate an optimal time series model and estimate parameters
- Analyse trends to find a connection between news stories (regarding the election) and public interest reflected in the popularity of Google search terms related to the two main candidates, Hillary Clinton and Donald Trump.

[1] A timeline of events corresponding to important stories regarding the election

This process allowed us to observe the effect that specific events had on the Google search popularity of the candidates' names. It would also be useful to make predictions in future elections based on the type of stories that emerge about each candidate. In addition, the resulting model we analysed would allow us to infer how different this election was with respect to other ones in terms of the search popularity of the candidates' names.

II. Methodology

Gathering the Data

We developed a data scraper in Python that draws data from Google trends on specific search terms in different countries. The efficiency of this scraper allowed us to quickly acquire a broad data set with several variables to choose from, in a manner that would have been impractical to do by hand. However, we decided to limit our data to the United States, as searches in that country were thought to be more affected by the timeline of campaigning events than searches in other countries. Additionally, the analysis of searches in multiple countries would be beyond the scope of this report.

The search terms Hillary and Trump were selected in order to analyse changes in the number of times people would research the two main presidential candidates (Hillary Clinton and Donald Trump) on Google. Hillary and Trump were chosen as the candidates were commonly referred to by these respective names, rather than by their full names.

By specifying the two search terms and including only the data pertaining to the United States, we created our final data set.[2]

The information collected spanned from November 2015 to November 2016, with one data point per week and the last data point representing the week after the elections. The data is based on a scale that measures the relative popularity of a search term compared to the most popular term in all the other weeks. For example, a term with a score of a 100 is the most popular,

[2]While we went through the effort of making a data scraper, it wasn't required for the purposes of this paper because the data was readily available via Google trends. The scraper only helped in the preliminary analysis.

whereas one with a score of 50 is half as popular. The two terms chosen (Hillary and Trump) were compared on the same scale.

Analyzing the Data

(i) Identifying stationary processes

Having gathered the data, we examined it and analysed it in R.[3] First, the raw data for each search term was plotted versus time to identify possible trends. Then the first and second differences were taken but these did not yield stationary processes (Figures 2 and 3). Since the raw data seemed to be following an exponential trend, a log-transformation was used to get a roughly linear time trend (Hogg et al., 2013). Afterwards, the first and second differences of the transformed series were taken. The first difference of the Hillary data and the second difference of the Trump data resulted in nearly stationary processes. Neither was perfectly stationary as the respective variances seemed to increase, but these processes were the closest we could get to stationarity. The first difference for Hillary was chosen to avoid over-differencing the data through the second difference (Please see the figures and details in the Results section).

(ii) Parameter Estimation

The log-transformed Hillary data was identified as an ARIMA (0, 1, 1) while the log-transformed Trump data was found to be an ARIMA (0, 2, 1) (Please see the figures details in the Results section). The function arima() was therefore used to calculate the Maximum Likelihood Estimator and the Conditional Sum of Squares for each of the MA(1) models found through differencing.

(iii) Residual analysis

Residuals for the two models were calculated and were plotted against time. Normal probability plots of the residuals were also analysed to identify possible departures from normality.

[3]The code used to make the graphs are attached separately in a .R file.

Furthermore, the sample autocorrelation function of each process was examined to determine whether the models chosen were appropriate (figure 12 and 13 in the appendix).

Finally, a timeline of all the important events that took place during the election campaign period was created based on information from. The events were organized by week and aligned with the raw time series plot to identify which controversies sparked the largest public interest during the time frame analyzed.

III. Results & Discussion

The plots for the raw time series data for both the search terms are:

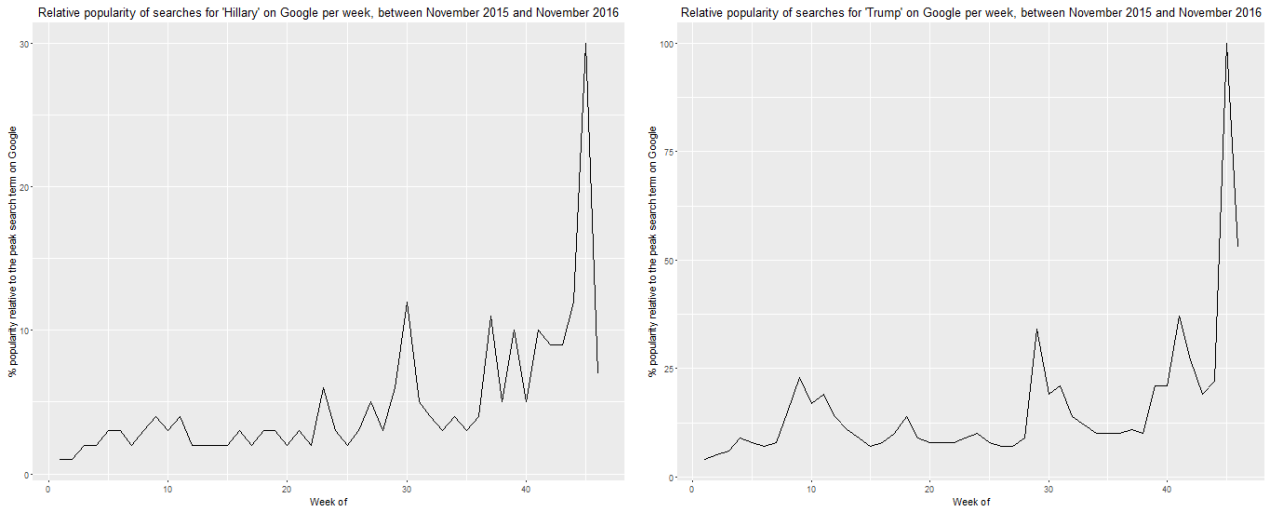

Figure 1: Time Series Plots - without log transformation

For this data, there is a lot of background knowledge available that can help interpret some of the rises and falls of each candidates' popularity. Donald trump has his first small peak during Super Tuesday, a day on which many US states hold primary elections. Hillary also has a small peak at this time. Clinton's popularity experiences its first big peak when enough caucuses are secured to win the democratic presidential nomination. At this time, Bernie Sanders officially endorses her. The local maximum of her popularity occurs a week or so after this event, during the week of June 24^{th}. Donald trump then has a huge peak during the republican national convention. Even after the peak, his popularity permanently stays higher than it was before. The republican national convention occurs on June 18^{th}. On September 11^{th}, Hillary Clinton's

popularity spikes, in contrast to Donald Trump's which remains constant. Perhaps this is because due to the articles that reported her leaving the memorial ceremony early or perhaps it is due to her calling Trump supporters 'deplorable' the day before.

Both candidates experience a peak during the first presidential debate on September 26^{th}, after which the popularity of their search terms continue to rise until the election. They both experience a peak around October 9^{th}, during which two major events occured. The first of which was when the FBI released their official statement saying that the investigation warranted no new action against Hillary Clinton: a conclusion to one of the biggest scandals that followed her during the entire election. The second of which is when the tape of Donald Trump discussing his sexual encounters was released followed by accusations of sexual assault from four different women. Finally, both candidates experienced a huge climb in popularity in the week leading up to the election and a subsequent fall afterwards.

Figure 1 shows that the raw data for either search term is not stationary. The first difference of both series resulted in the plots in figure 2.[4]

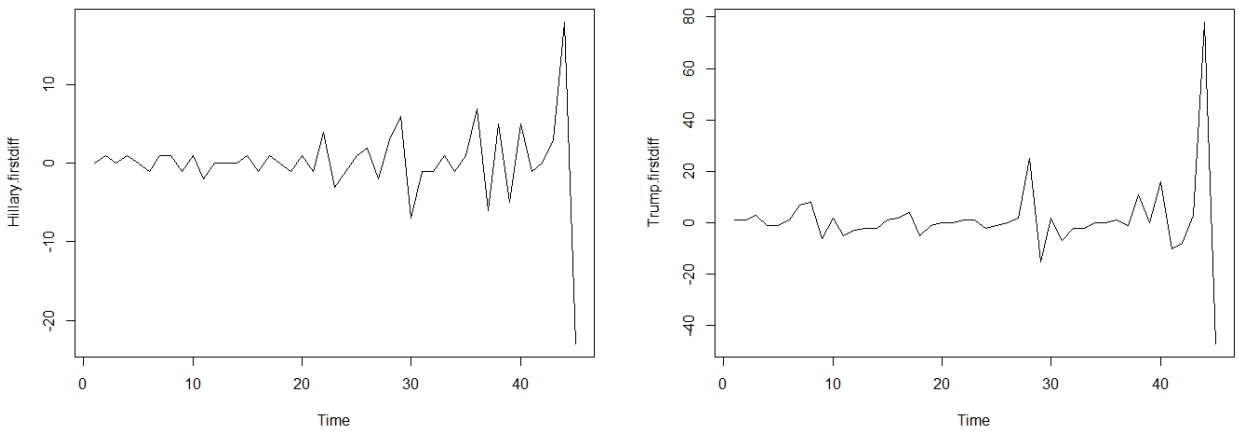

Figure 2: First Difference Plots: Hillary [left], Trump [right] - without log transformation

The second difference was also performed and the results are presented in figure 3 below. These graphs indicate that the first and second difference for both data sets are not stationary.

[4]The autocorellation functions of all of the time series analysed are available in the appendix

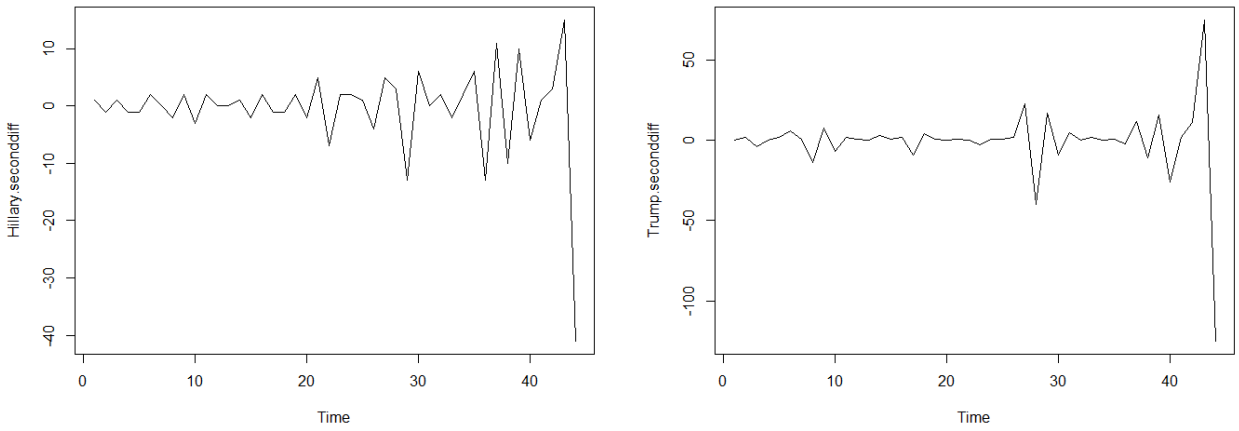

Figure 3: Second Difference Plots: Hillary [left], Trump [right] - without log transformation

The log transformation of this data is presented in Figure 4. Figure 5 shows the first and second differences of the log transformed data for each candidate. The first difference for 'Hillary' shows stationarity while the second difference brings us close to stationarity for 'Trump'. However, the second difference for 'Trump' indicates increasing variance. The autocorellation function of the first difference for 'Hillary' indicated that it was an MA(1) (Figure 8), and the autocorellation function of the second difference for 'Trump' indicated that it was an MA(1) (Figure 9). Parameter estimates for the ARIMA (0, 1, 1) model (for the log-transformed data of 'Hillary') and ARIMA (0, 2, 1) (for the log-transformed data of 'Trump') are displayed in Figure 11, in the appendix. For the 'Hillary' model, based on the plot of residuals versus time and the autocorrelation function, the residuals looked like white noise. The normal probability plot indicated that the distribution might have had a slight heavy tail. The histogram of residuals showed that they generally met the normality assumption. The model selected for 'Hillary' was therefore appropriate. For the 'Trump' model, based on the autocorrelation function, the residuals looked like white noise. However, the plot of residuals versus time indicated that the variance increased with time. The normal probability plot and the histogram of residuals indicated showed that the normal distribution had a positive skew. The model selected for 'Trump' may therefore not have been appropriate.

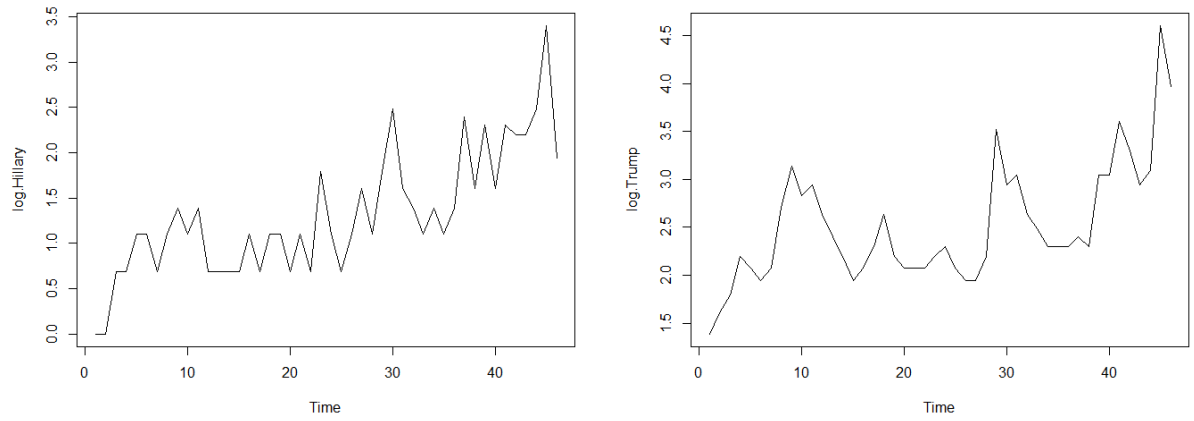

Figure 4: Time Series Plots - with log transformation

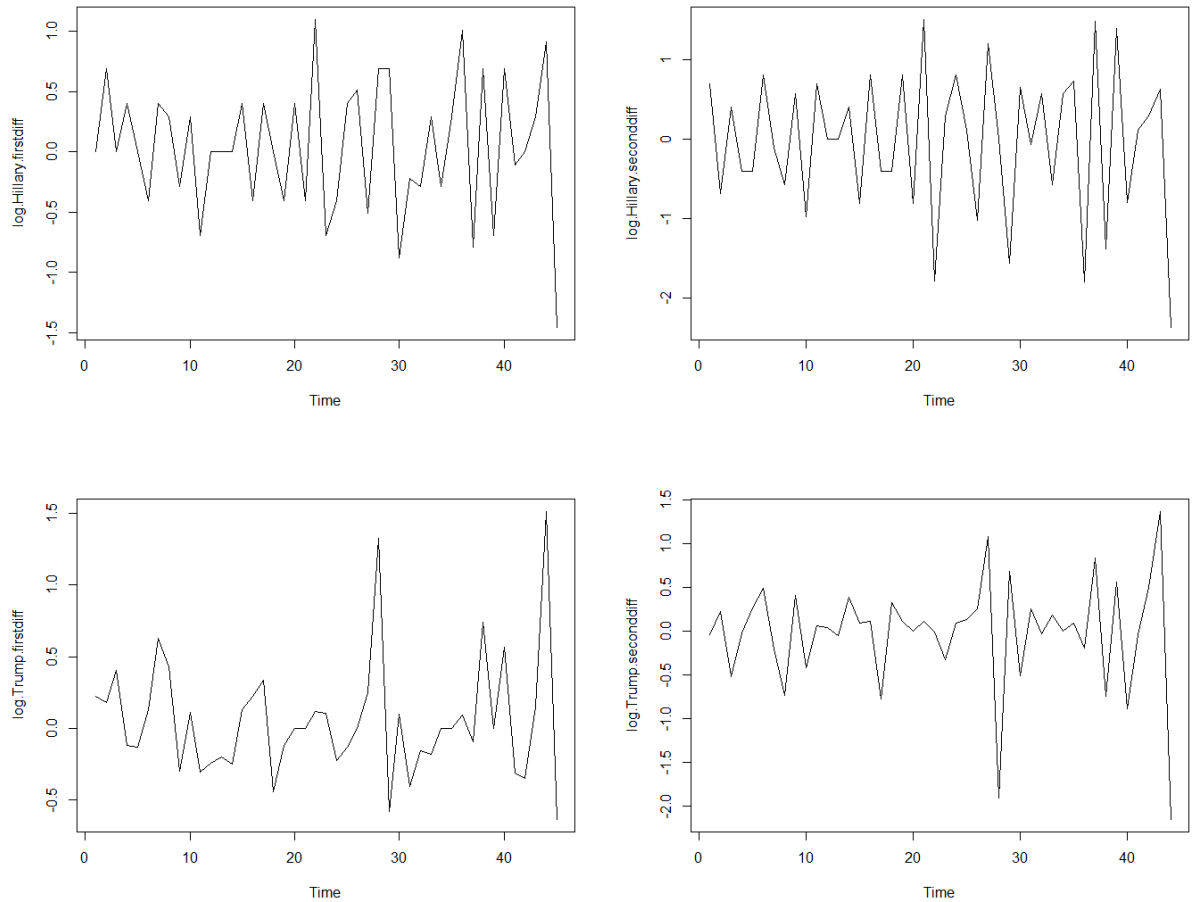

Figure 5: Difference Plots With Log Transformation: Hillary [top], Trump [bottom]

IV. Conclusion

As far as our conclusions are concerned, we noticed that the Democratic National Convention, September 11th memorial, and the FBI email investigation results are the events that had the most impact on Hillary and Super Tuesday, the Republic National Convention, and the tape of himself and Jeb Bush are the ones that affected Trump the most. Both candidates were affected by the first presidential debate and the final election date on November 8^{th}. This can be seen by looking at the time line presented in the appendix. It is difficult to infer what exactly these impacts imply when we consider future results.

Regarding the time series model, we decided to go with an ARIMA (0, 1, 1) for Hillary and an ARIMA (0, 2, 1) for Trump[5]. However, in the future, it might be fruitful to look into a poisson distribution for the Trump data. While our analysis seemed sufficient for the scope of this class, it was incomplete, as the second difference of the log-transformed Trump data did not look completely stationary. It could be useful to consider a poisson distribution for such data.

V. References

- A. (n.d.). Election Day. Retrieved November 23, 2016, from http://www.aol.com/2016-election/timeline/
- Cryer, J. D., & Chan, K. (2008). Time series analysis: With applications in R. New York: Springer.
- "United States Presidential Election, 2016 Timeline." Wikipedia. Wikimedia Foundation, n.d. Web. 23 Nov. 2016.

VI. Appendix

Included here are the autocorrelation functions for each of the graphs plotted:

[5]As mentioned in section II

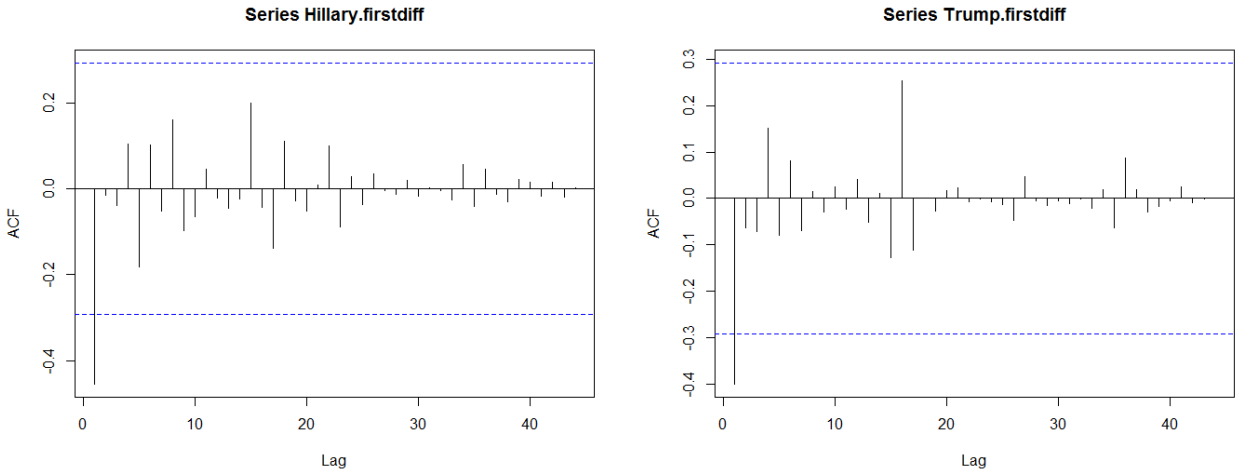

Figure 6: First Difference ACF: Hillary [left], Trump [right] - without log transformation

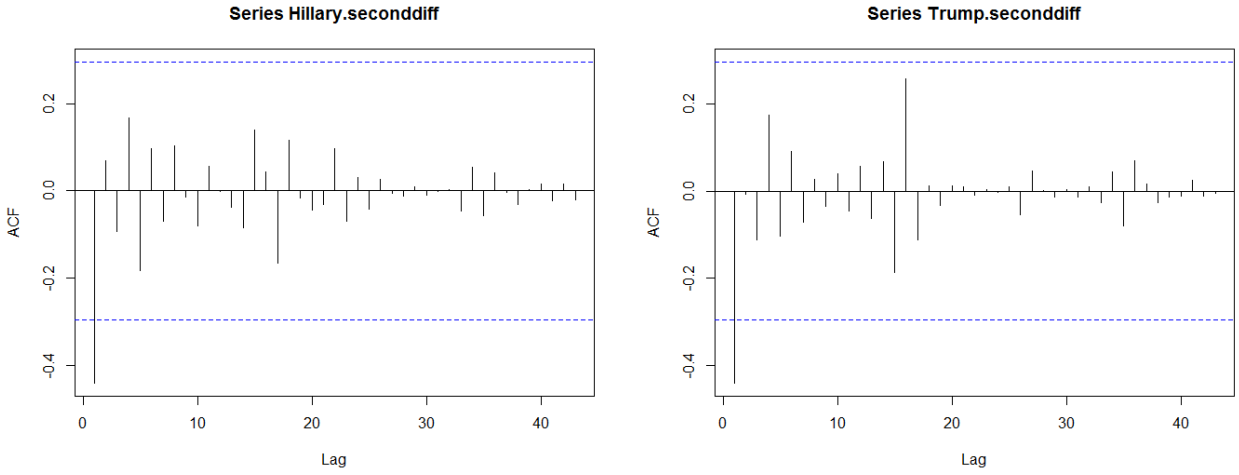

Figure 7: Second Difference ACF: Hillary [left], Trump [right] - without log transformation

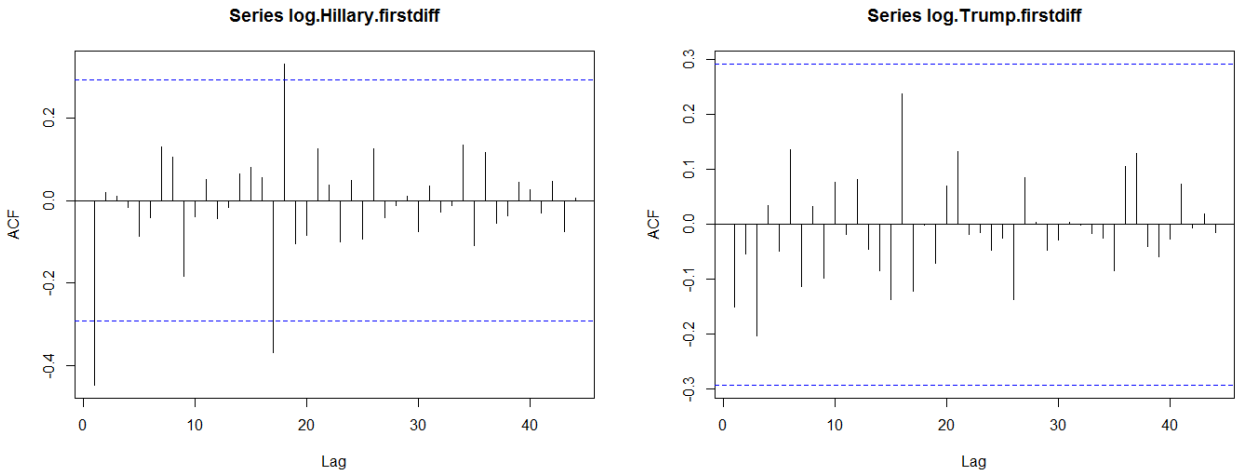

Figure 8: First Difference ACF: Hillary [left], Trump [right] - With log transformation

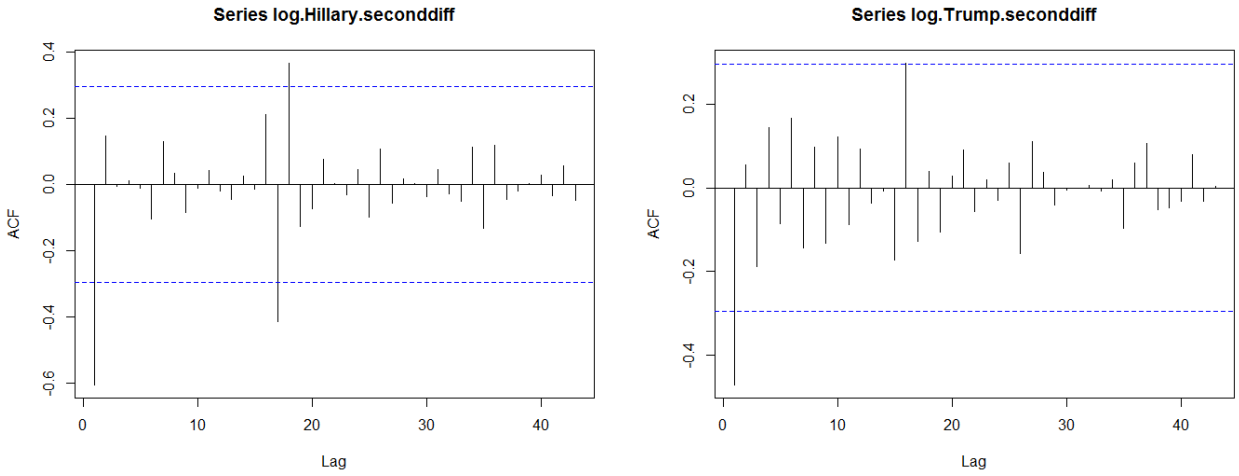

Figure 9: Second Difference ACF: Hillary [left], Trump [right] - With log transformation

Jan	Feb	Mar	Apr	May	Jun	Jul				Aug	Sep	
		1	19	26	6	5	12	18	25	10	10	11
Republican and Democratic debates continue	Hillary Clinton and Donald Trump begin to win Democratic Caucuses and Republic Primaries, respectively	Super Tuesday	New York Primaries won by Hillary Clinton and Donald Trump	Donald Trump secures the Republican presidential nomination	Hillary Clinton secures the Democratic presidential nomination	FBI ending Clinton email probe, will not recommend prosecution	Sanders endorses Clinton	Republican National Convention	Democratic National Convention, DNC chair resigns after email leak	Newly released emails show State Department ties with Clinton Foundation	Hillary Clinton calls Trump supporters 'deplorables'	Hillary Clinton leaves 9/11 memorial early

Sep		Oct								Nov	
18	26	2	4	7	9	12	19	28	31	6	8
Hillary Clinton and Donald Trump react to the New York City bombing	The first general presidential election debate takes place in	Donald Trump's 1995 tax records suggest no federal taxes for years	The vice presidential debate takes place	Tapes show Donald Trump talking about sexual exploits in 2005	The second general presidential debate takes place in Missouri	Four women accuse Donald Trump of inapropriate touching	The third general presidential debate takes place in Nevada	The FBI announces its plans to restart the investigation into Hillary Clinton's emails	Hacked emails reveal that CNN employee provided debate questions to Hillary Clinton	FBI director says emails warrant no new action against Hilary Clinton	Donald Trump is elected president of the United States

Figure 10: Weekly time line of all events leading up to the election - Corresponds to the data.

Hillary		
	Method	
	Maximum Likelihood	Conditional Sum of Squares
theta1	-0.5938	-0.5699
standard error	0.127	0.1171
Trump		
	Method	
	Maximum Likelihood	Conditional Sum of Squares
theta1	-1.000	-0.9552
standard error	0.0661	0.0419

Figure 11: Parameter estimates found using MLE and Conditional Sum of Squares

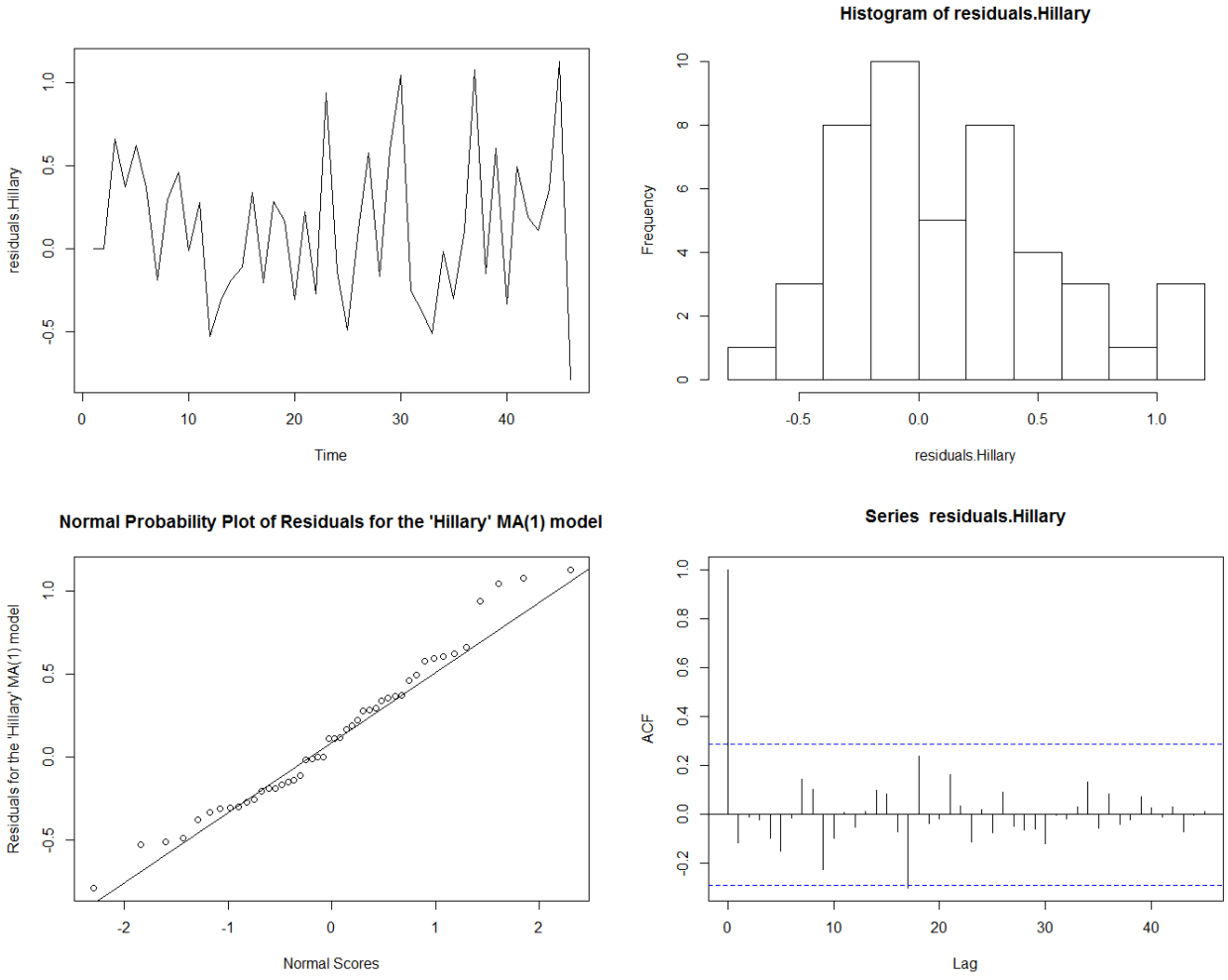

Figure 12: Model Diagnostics for Hillary

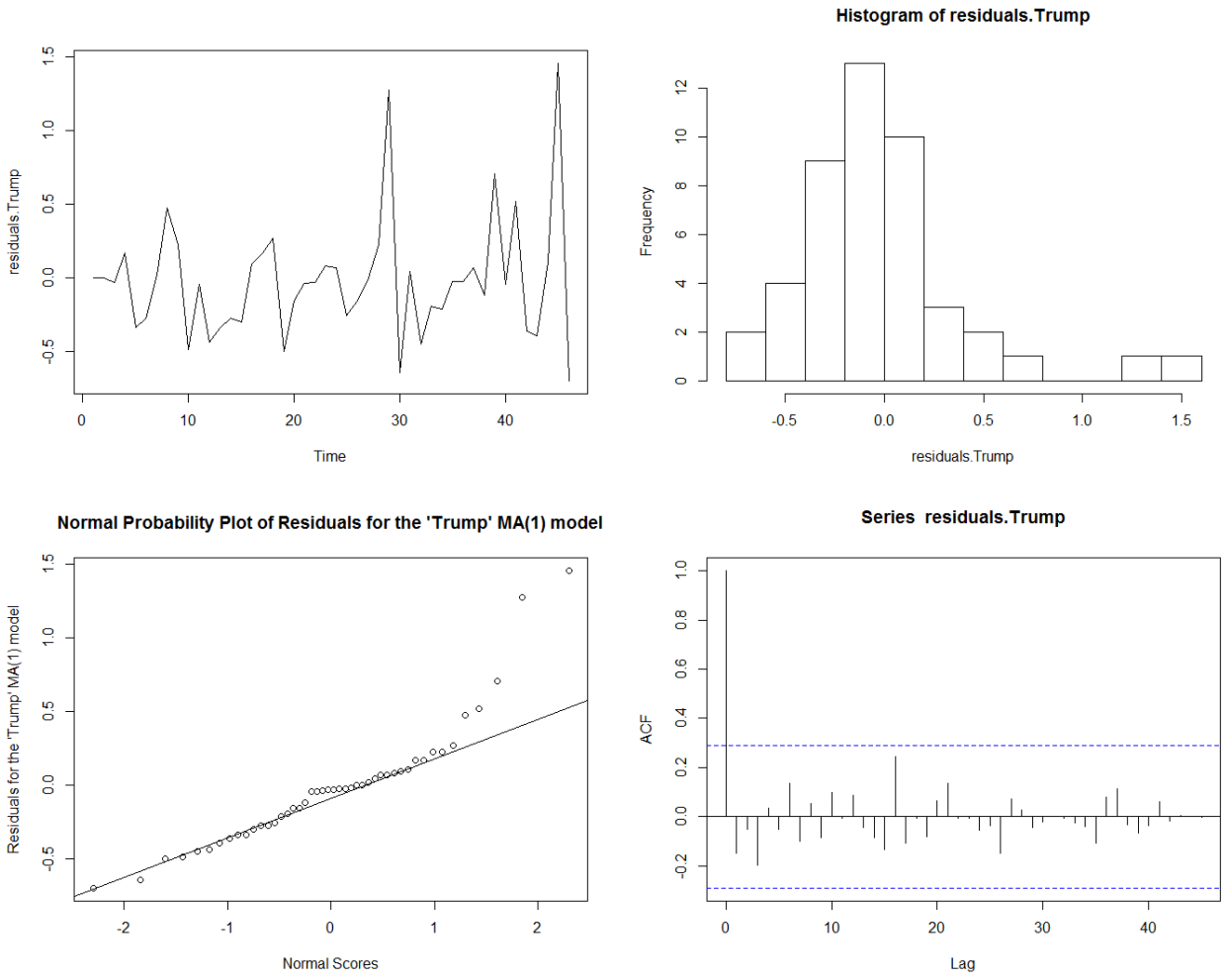

Figure 13: Model Diagnostics for Trump