

Time Series Analysis:

U.S. Military Casualties in World War Two

Rachael Becker, University of Central Florida

Abstract

This paper aims to show how statistical analysis can be used in the field of History. The primary focus of this paper is show how SAS® can be utilized to obtain a Time Series Analysis of data regarding World War II. The hope of this analysis is to test whether Truman's justification for the use of atomic weapons was valid. Truman believed that by using the atomic weapons he would be preventing unacceptable levels of U.S. casualties that would be incurred in the course of a conventional invasion of the Japanese home islands.

Introduction

Following the end of World War One, Germany was deeply penalized for its actions in the War. Many see this as a main cause for World War Two. Economic hardship and political problems lead to the rise of the Nazis party. Adolf Hitler used the German people's hatred toward the Treaty of Versailles as a springboard to his seizure of power. By mid-1934, Hitler had successfully turned Germany into a one-party state with himself the sole ruler of Germany. He convinced the people of Germany that their land had been stolen from them and that if Germany was going to flourish once again then they would need to regain the land that had been lost and to conquer for themselves vast new territory in Eastern Europe. In 1938 Hitler began invading neighboring countries in an effort to forge "Greater Germany."

Following the invasion, prominent physicists attempted to inform President Franklin Delano Roosevelt of the danger of nuclear research currently occurring in Germany and evident by the halt of the sale of Uranium from mines in German-occupied Czechoslovakia. On September 1st 1939, the German army invaded Poland. This event marks the beginning of World War Two. In 1941, the United States of America was plunged into World War Two following the bombing of Pearl Harbor by the Japanese. These lead to one of the most controversial issues in history: the decision to use the atomic weapons.

Harry Truman's main reason for dropping the atomic bomb was to bring the war to an end quickly in an effort to obviate the need for an invasion of the Japanese home islands and avoid the extreme U.S. losses that were predicted. Although probably not one of his main concerns, Japanese citizens were dying at a rapid rate. This was a result of the propaganda that had been spread by the Japanese military about the U.S. soldiers.

This caused an extremely high rate of suicide among the Japanese citizen because they feared what would happen when the U.S. soldiers invaded the island they were residing on.

However, this paper will not examine the effect a prolonged war could have had on Japanese civilians, but rather, the purpose of this paper is to examine the reasons behind the dropping of the bombs. It was estimated that 500,000 U.S. soldiers would lose their lives in order to take the Japanese home islands. With advances in statistical computing since 1945, I want to examine whether or not the estimates made in 1945 were reasonable.

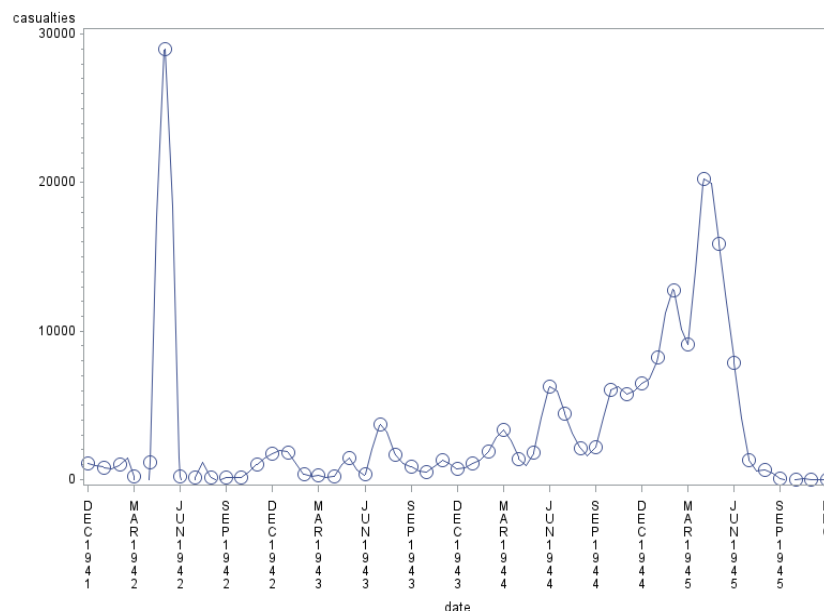
For this particular data set, the death rates of U.S. soldiers are available by month for the entirety of U.S. military involvement in the Pacific Theater. I have decided to analyze the data up until, but not including, August of 1945 because the first atomic bomb was dropped in early August. The dropping of two atomic bombs caused the unconditional surrender of Japan. This signifies the end of the war, so it does not make sense to include the data after the atomic bombs' use when trying to understand why the bombs were dropped.

Models

The following graphs show the number of casualties sustained by the U.S. Military in the Pacific theater by month.

This following code produced this graph:

```
proc gplot data = PacTheater;
  symbol i = spline v = circle h = 2;
  format date MONYY7.;
  plot casualties*date /
    haxis = "1dec1941"d to "1dec1945"d by qtr;
run;
```

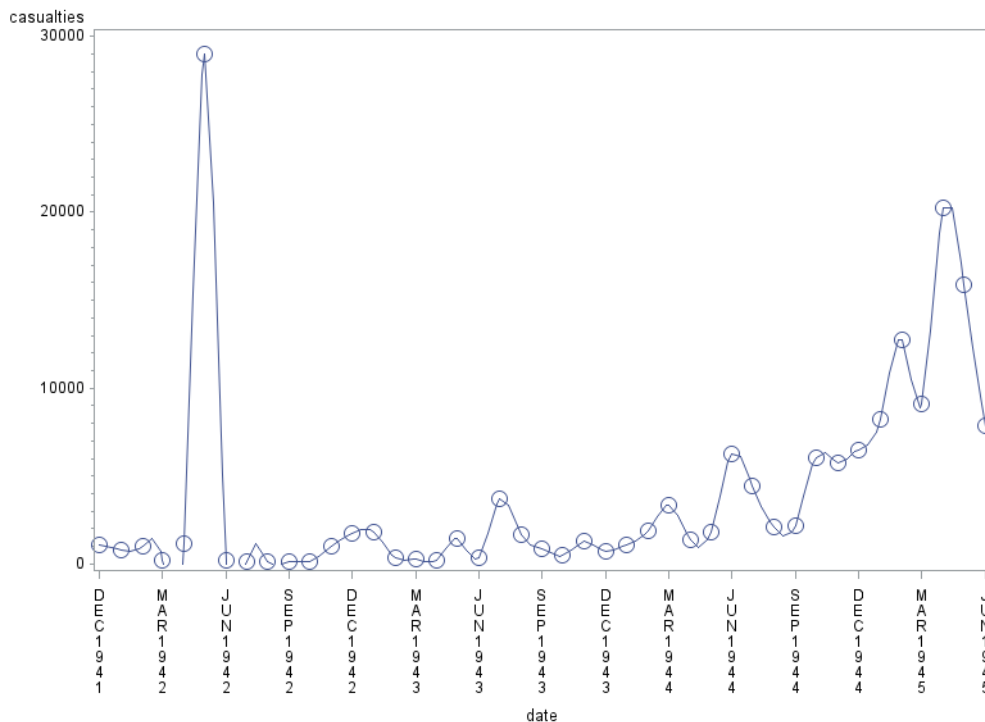


In order to exclude the data gathered after the first atomic bomb was dropped, the following code was utilized and produced this graph:

```
proc gplot data = PacTheater;
    symbol i = spline v = circle h = 2;
    format date MONYY7.;
    plot casualties*date /
        haxis = "1dec1941"d to "1jul1945"d by qtr;
run;
```

This code results in a dataset that excludes observations past July of 1945:

```
data PacJul; set PacTheater; if date>-5267 then delete; Format date MONYY7.;
run;
Proc print data = PacJul;
run;
```



A first order differencing was done in order to make the data stationary.

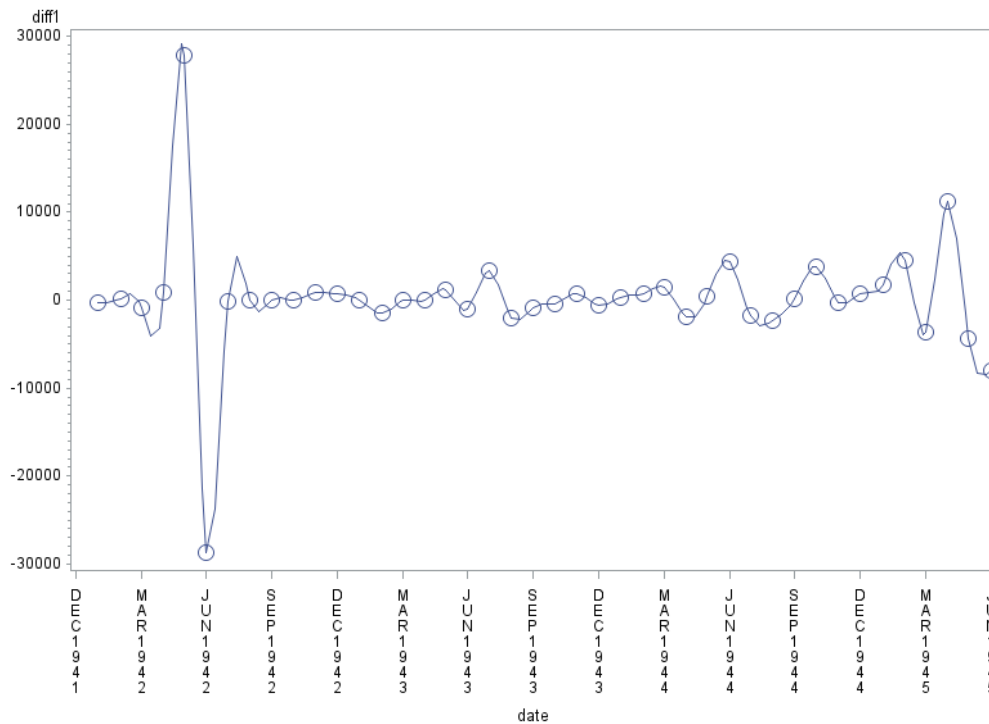
$$z_t = y_t - y_{t-1}$$

where $t = 2, \dots, n$

Data is stationary when the mean and variance of the data are relatively constant through time. The following code and graph are a result of the differencing.

```
Data FirstOrder;
set PacJul;
diff1=DIF(casualties);
run;
proc gplot data = FirstOrder;
    symbol i = spline v = circle h = 2;
    format date MONYY7.;
    plot diff1*date /
        haxis = "1dec1941"d to "1jul1945"d by qtr;
run;
```

```
format date MONYY7.;
plot diff1*date /
      haxis = "1dec1941"d to "1jul1945"d by qtr;
run;
```



The first order differencing indicates that the integrated term in the ARIMA model should be set to one. The ARIMA model can be coded in SAS using PROC ARIMA. The previous work shows that a first order differencing of the variable “casualties” is necessary. It is necessary to look at the autocorrelation and partial autocorrelation to determine what the p and q should be set to. The following is a graph of the autocorrelation that can be described by this model:

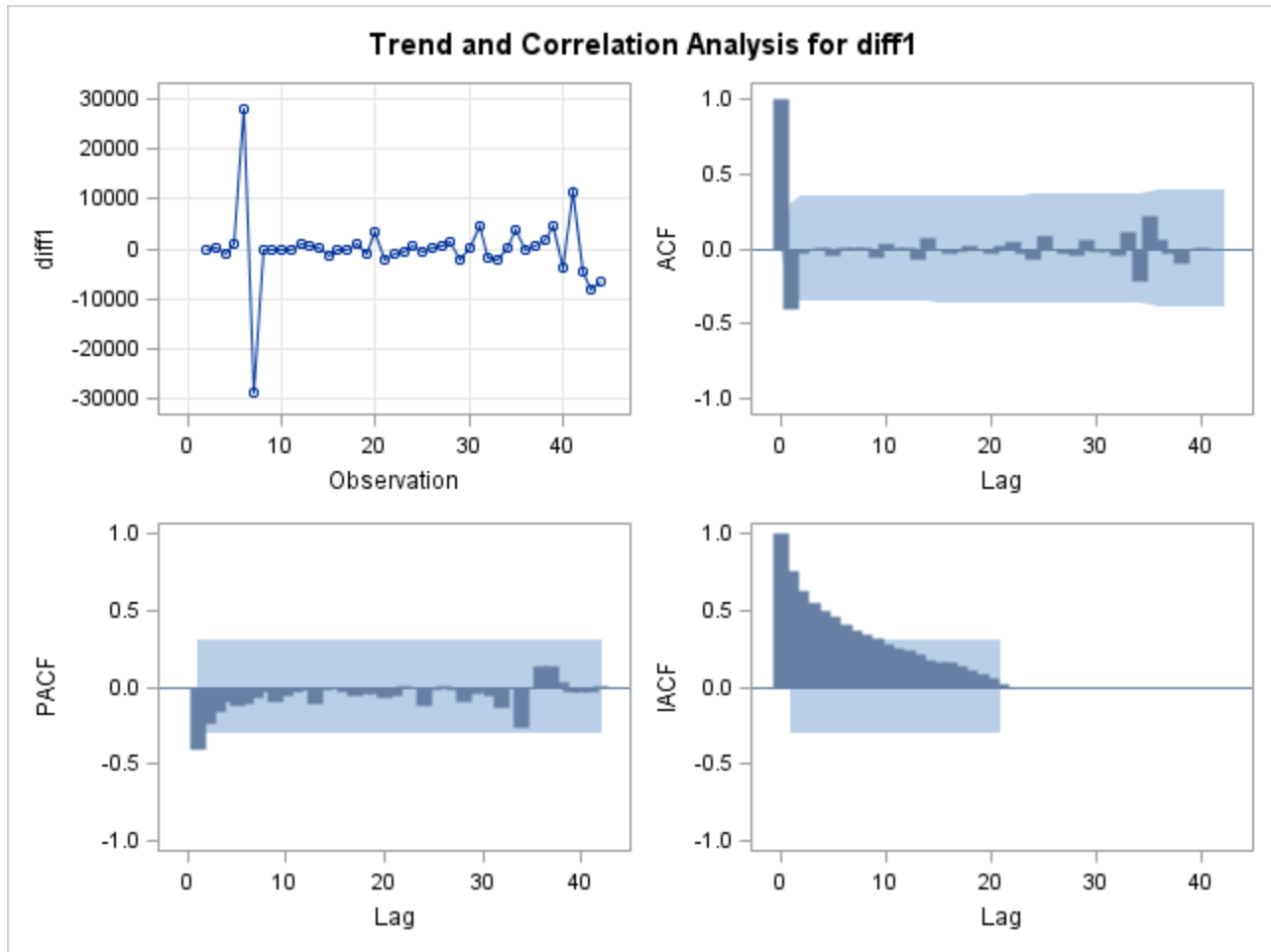
$$z_t = a_t - \theta_1 a_{t-1}$$

The graph of partial autocorrelation can be obtained from this equation:

$$z_t = \Phi_1 z_{t-1} + a_t$$

The graph was obtained in SAS with the following code:

```
Proc Arima data= FirstOrder;
  Identify Var=diff1 Nlag=47 OutCov=corr;
run;
```



The ACF graph indicates that a moving average with model order one should be used because the ACF drops off after lag one.

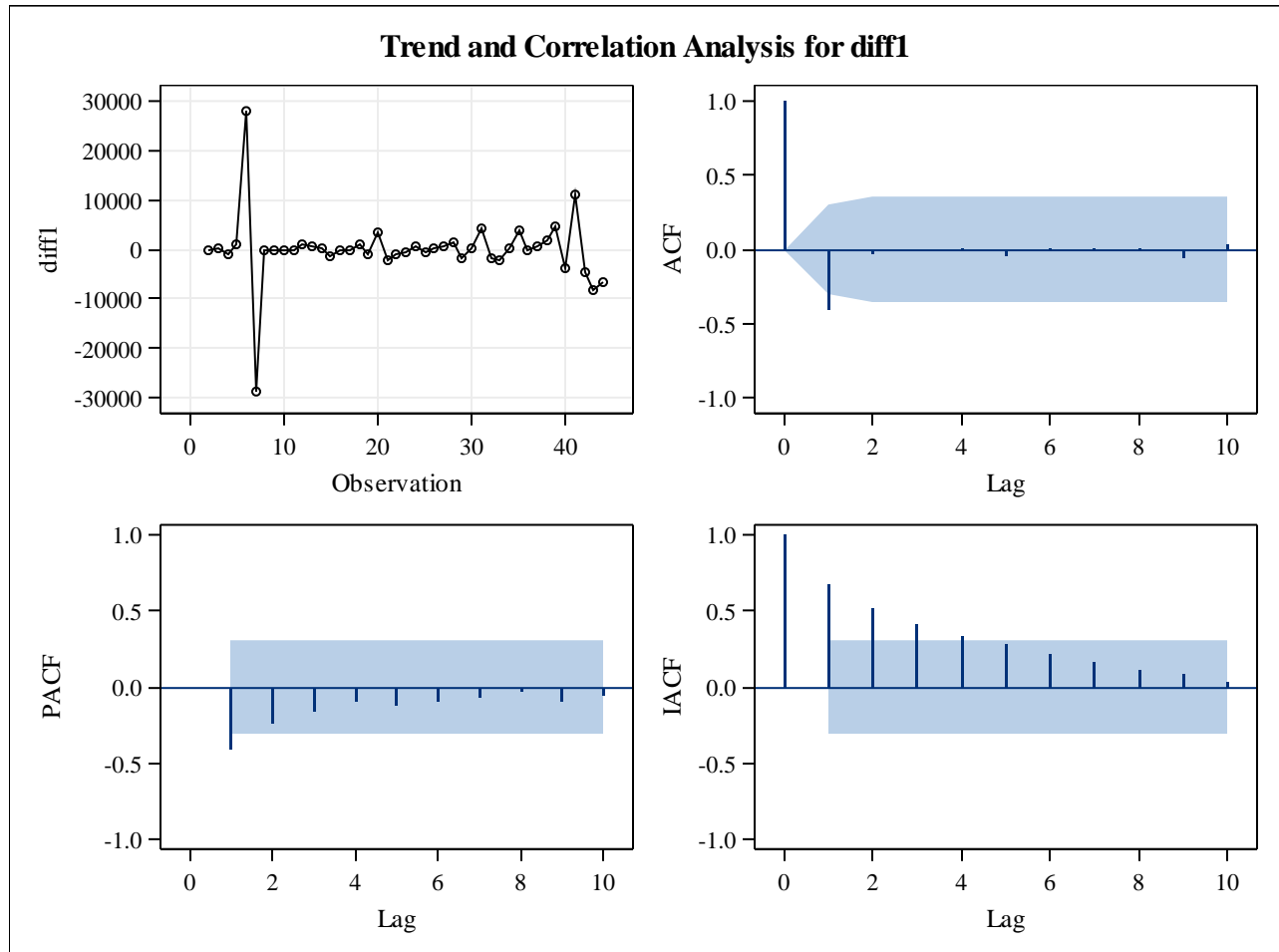
The PACF graph indicates that an autoregressive model of order one should be used because the PACF drops off after lag one.

The code below produced the following graphs and estimates:

```
Proc Arima data = FirstOrder;
  identify var=diff1;
  estimate q=(1) (12) noint method=ml;
  forecast id=date interval=month printall out=ForcastedPac;
run;
```

Name of Variable = diff1	
Mean of Working Series	5.860465
Standard Deviation	6754.598
Number of Observations	43

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	7.80	6	0.2532	-0.406	-0.038	-0.009	0.011	-0.051	0.006

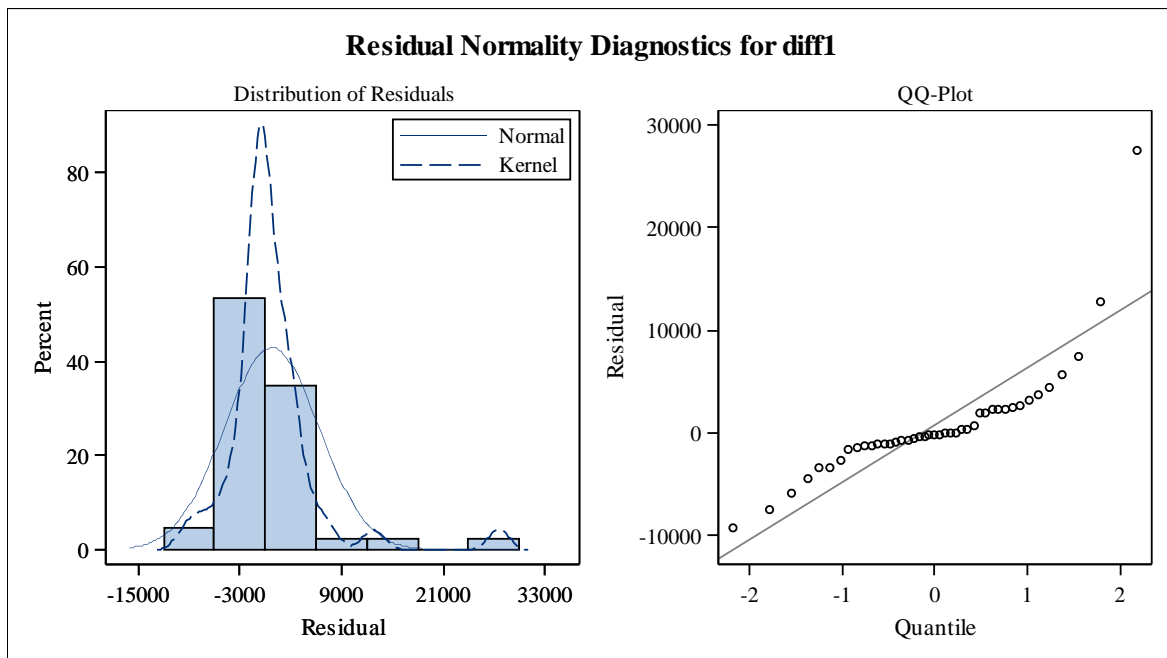
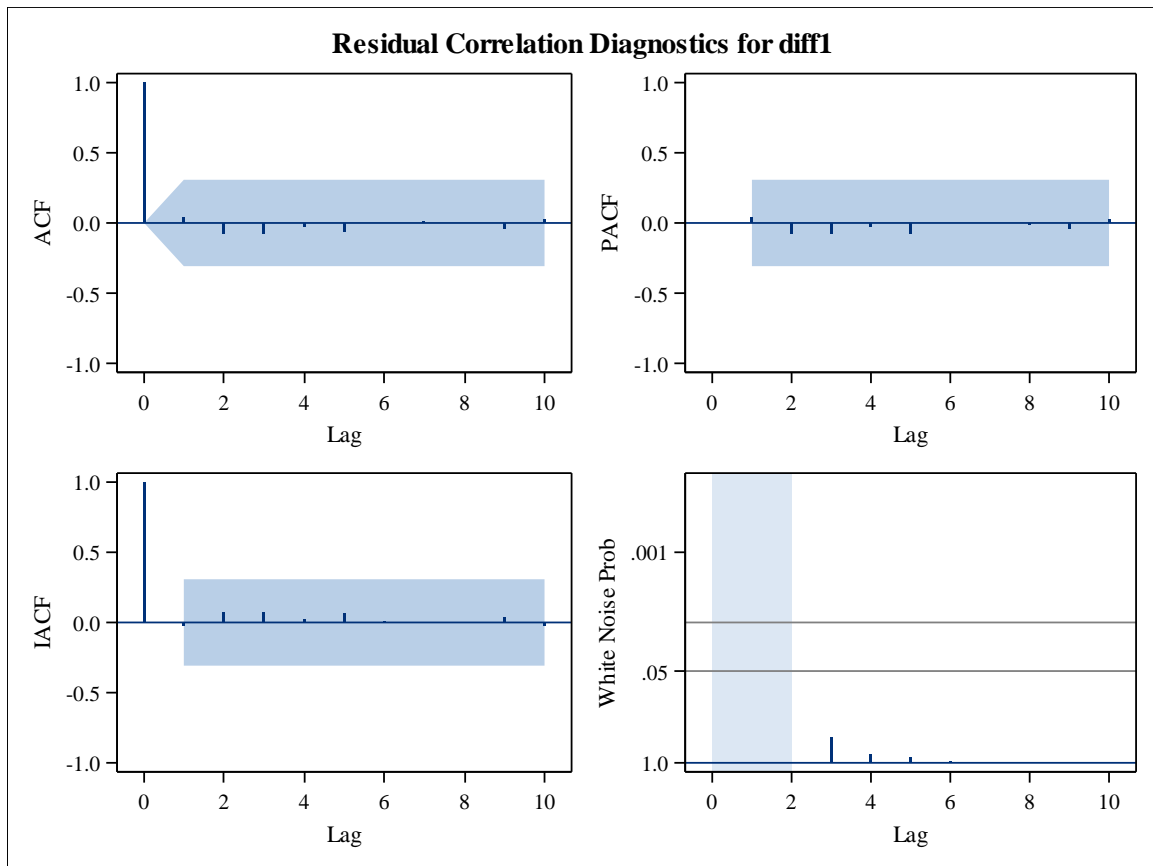


Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag
MA1,1	0.77776	0.10578	7.35	<.0001	1
MA2,1	0.08240	0.18440	0.45	0.6550	12

Variance Estimate	32406100
Std Error Estimate	5692.636
AIC	868.635
SBC	872.1574
Number of Residuals	43

Correlations of Parameter Estimates		
Parameter	MA1,1	MA2,1
MA1,1	1.000	-0.070
MA2,1	-0.070	1.000

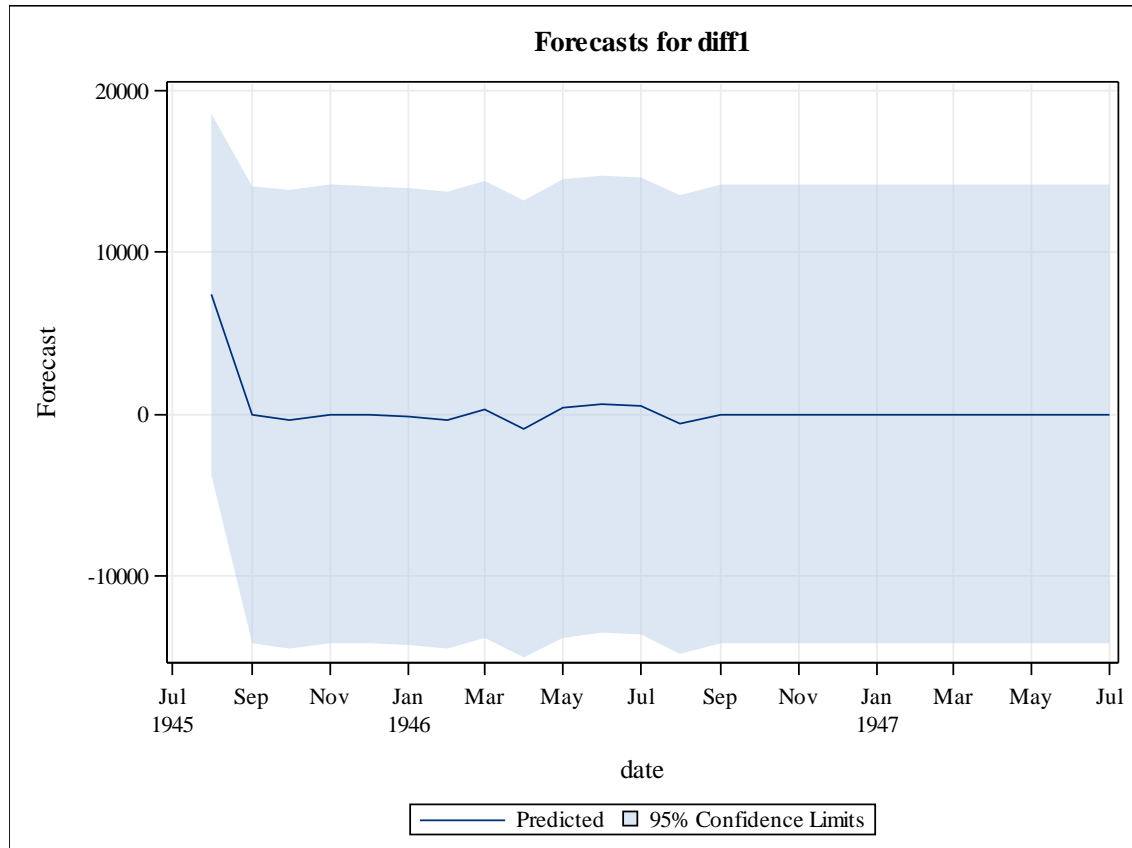
Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	0.62	4	0.9609	0.063	-0.047	-0.053	-0.008	-0.062	0.002
12	0.81	10	0.9999	0.016	0.006	-0.030	0.031	0.006	0.032
18	1.52	16	1.0000	-0.048	0.061	-0.000	-0.054	-0.035	-0.011
24	2.15	22	1.0000	-0.025	-0.030	0.024	0.018	-0.035	-0.054



Model for variable diff1

No mean term in this model.

Moving Average Factors	
Factor 1:	$1 - 0.77776 B^{**}(1)$
Factor 2:	$1 - 0.0824 B^{**}(12)$



The ACF plot of the residuals helps us to understand whether or not the model is a good fit. It appears to be a good fit because none of the lags are high. The QQ-Plot also indicates that the model is an appropriate fit.

Using PROC ARIMA, a graph of projected values, amongst other useful information, can be produced. This graph shows a 95% confidence interval for the estimated number of deaths that could have been incurred by the U.S. military if fighting had continued in the Pacific for 24 more months. This would mean that if the war had ended in July of 1947, the most extreme estimation of lives lost amongst the U.S. military would be approximately 343,650. This is about 68 percent of the estimated value that was

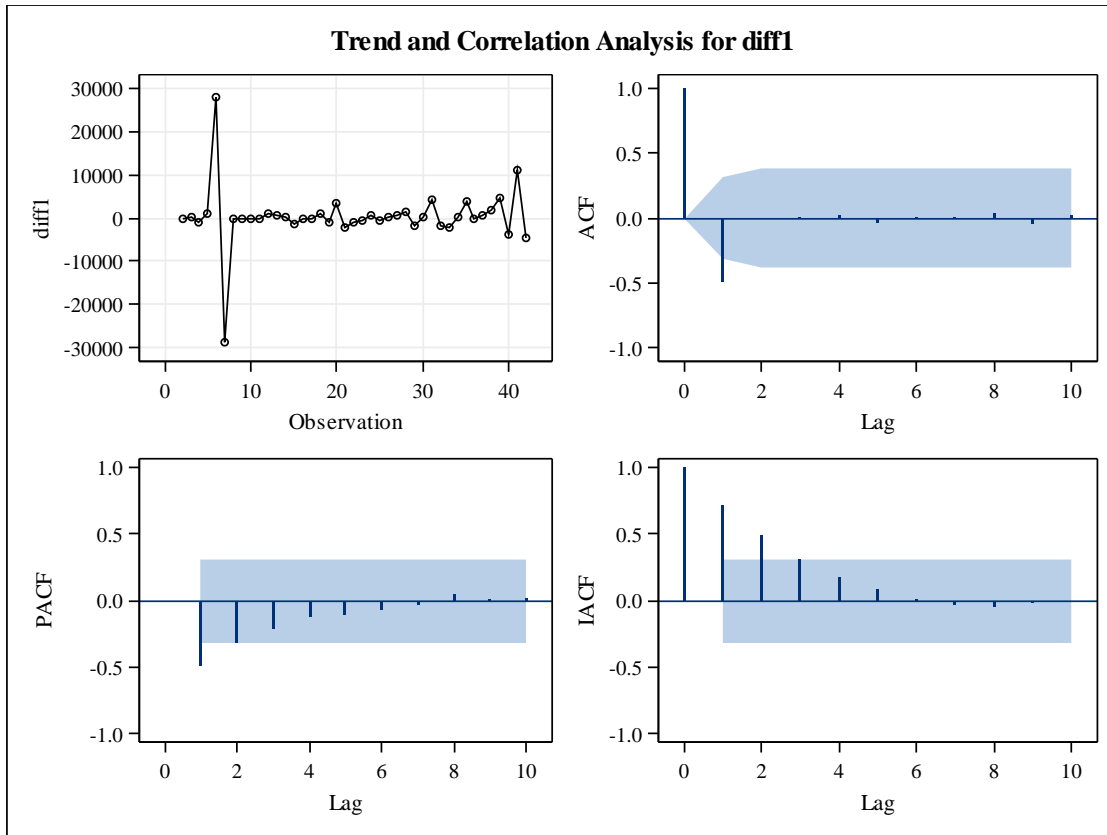
used to persuade the President to use the atomic weapons. It should be noted that the plan to invade the home islands of Japan would not have been implemented until March of 1946 which means that the war would likely have continued far past July. It is difficult to say whether or not the original estimate was too extreme without an indication as to how long the invasion of Japan would have taken.

I decided to try fitting another model after removing the data points that were obtained after May of 1945. At that point in time, there was not an approved plan to invade Japan. Harry Truman did not approve a plan to invade the home islands until after the U.S. military operations in the European theater was over. This meant that the U.S. military in the Pacific theater was unable to begin the invasion and thus resorted to continually bombing Japanese cities in hopes of a surrender without an invasion. This period of decreased troop involvement causes the number of deaths incurred to drastically decrease, but if they had not had this break in fighting, perhaps a more accurate estimate of the number of troops that would have died had the U.S. invaded the home islands of Japan could be calculated.

The code for the following output can be obtained in the appendix.

Name of Variable = diff1	
Mean of Working Series	360.9512
Standard Deviation	6716.392
Number of Observations	41

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	10.29	6	0.1128	-0.482	-0.008	0.011	0.016	-0.027	0.015

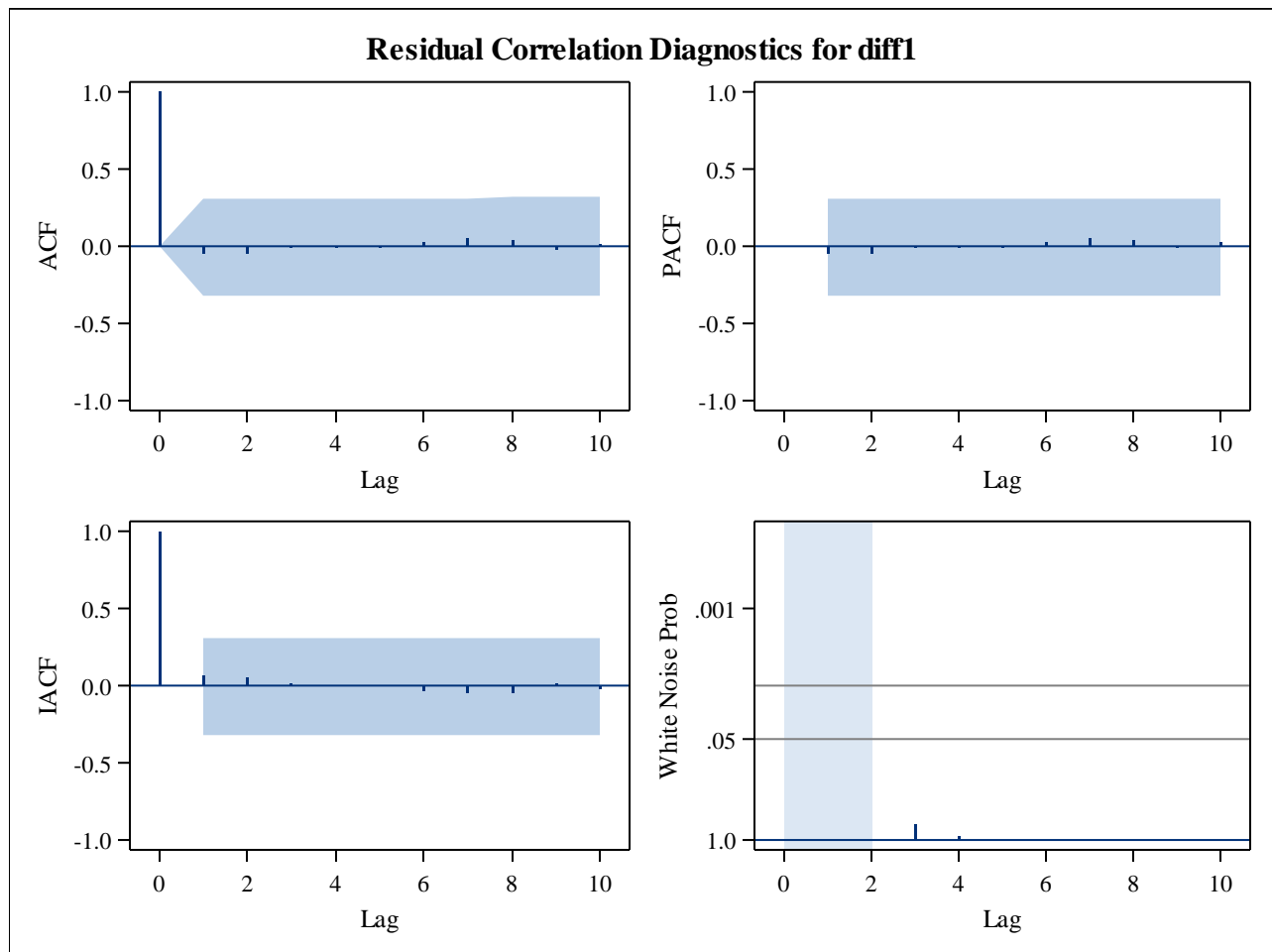


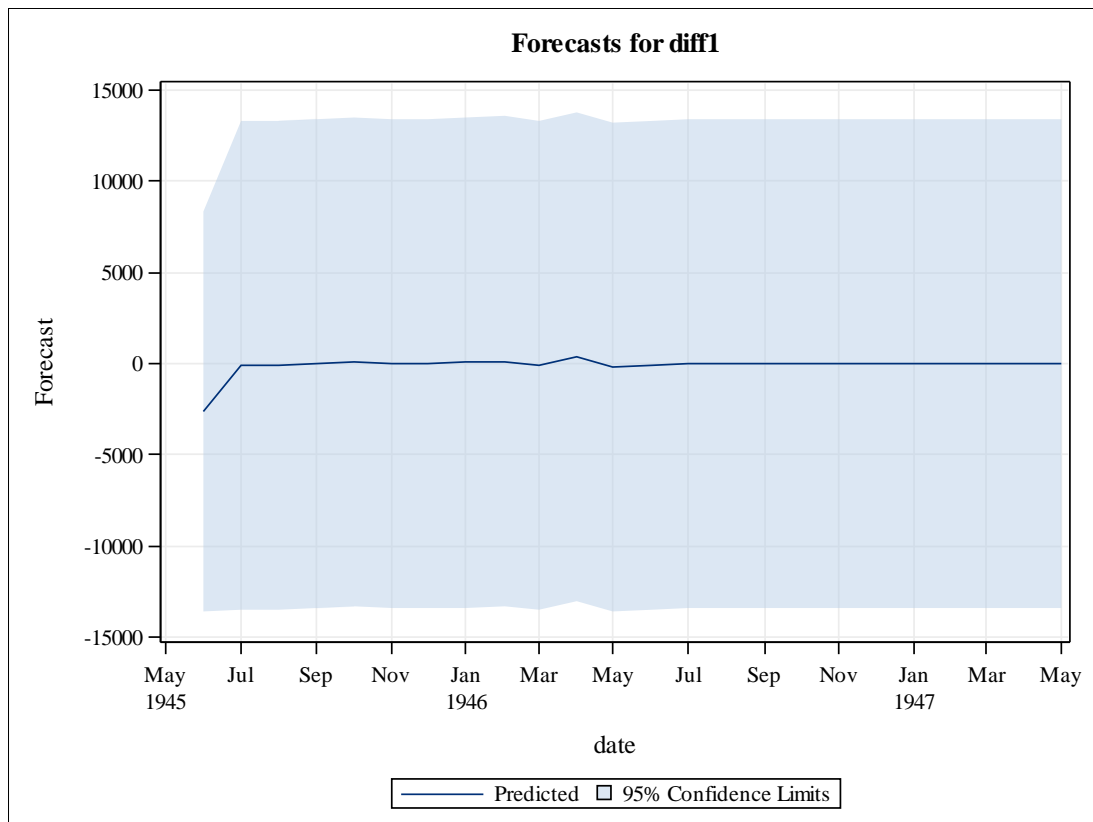
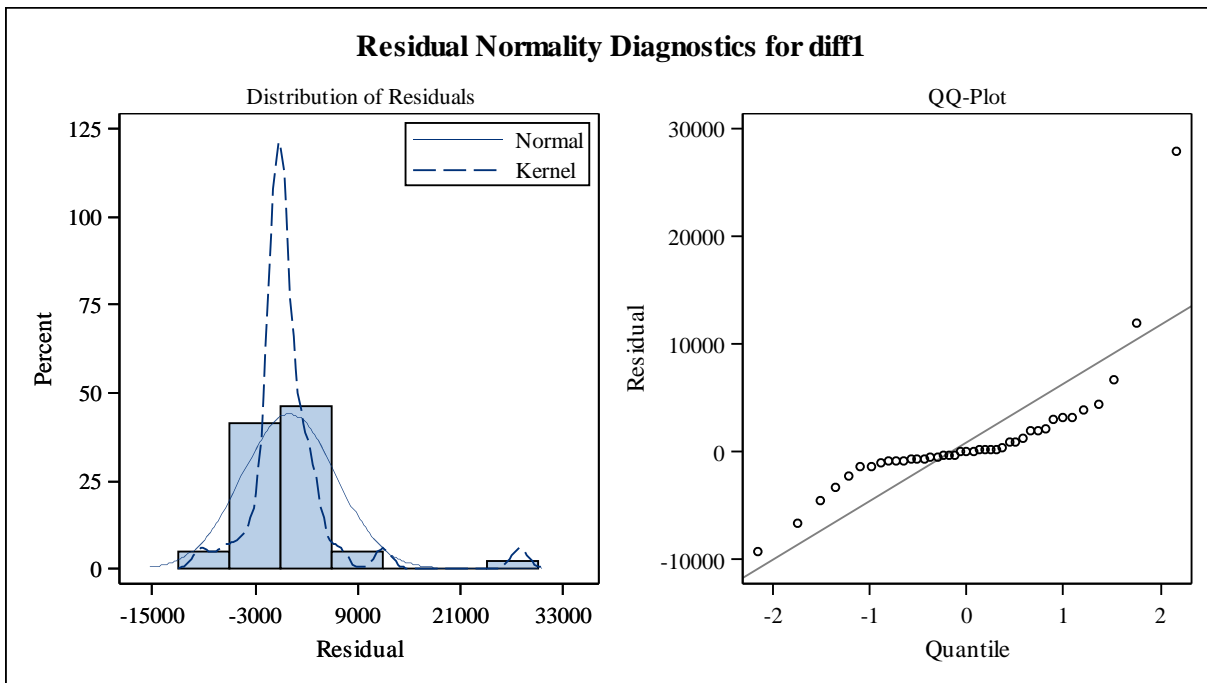
Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag
MA1,1	0.70019	0.13859	5.05	<.0001	1
MA2,1	-0.03215	0.17898	-0.18	0.8574	12

Variance Estimate	31361809
Std Error Estimate	5600.161
AIC	826.6931
SBC	830.1202
Number of Residuals	41

Correlations of Parameter Estimates		
Parameter	MA1,1	MA2,1
MA1,1	1.000	0.011
MA2,1	0.011	1.000

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	0.12	4	0.9983	-0.019	-0.029	0.012	0.009	-0.017	0.028
12	0.42	10	1.0000	0.048	0.041	-0.013	0.029	0.016	-0.018
18	1.01	16	1.0000	-0.017	0.084	0.021	-0.023	-0.023	-0.000
24	1.79	22	1.0000	-0.020	-0.031	0.018	0.033	-0.048	-0.055

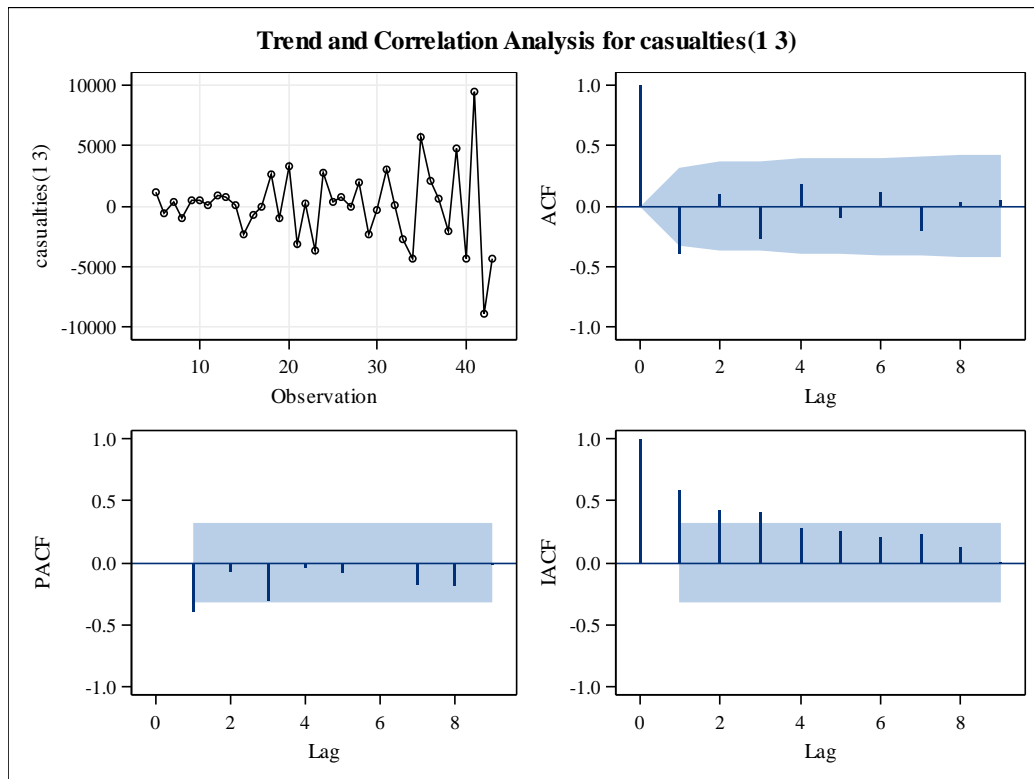


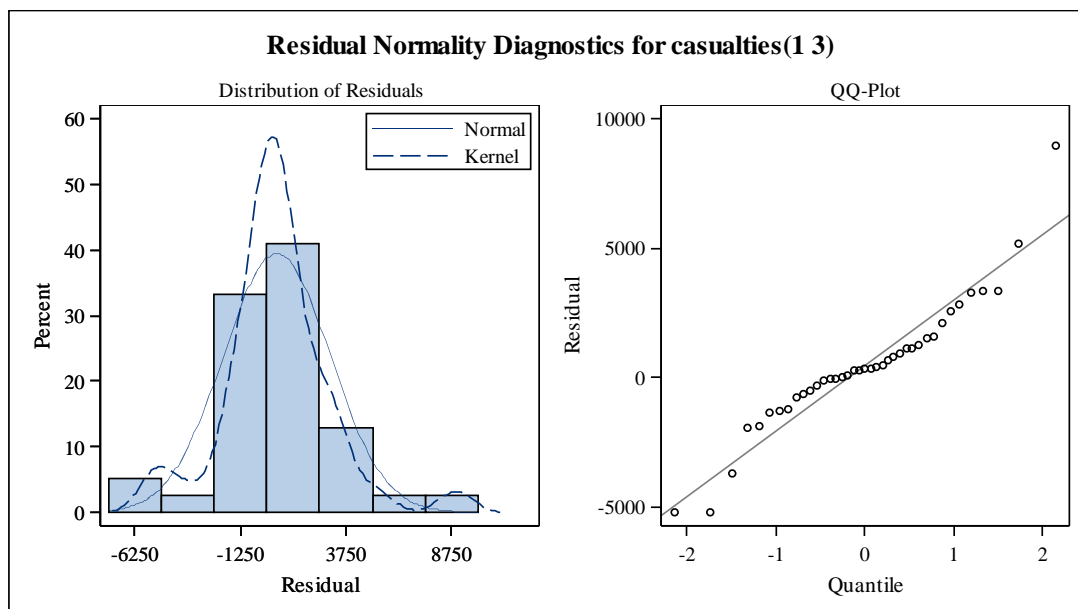
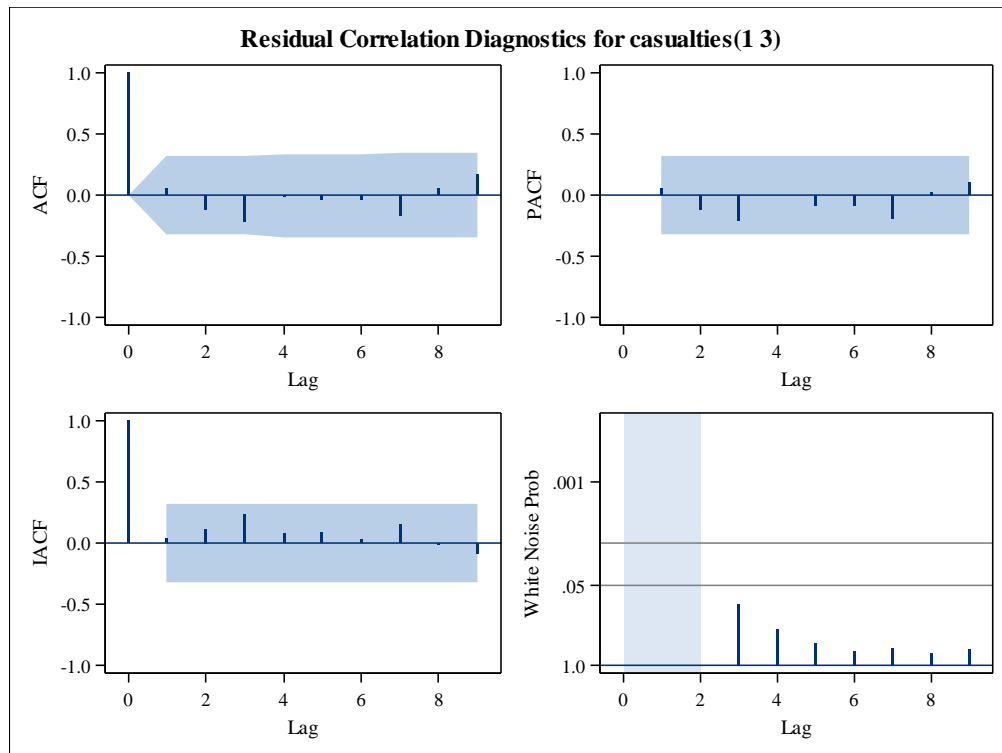


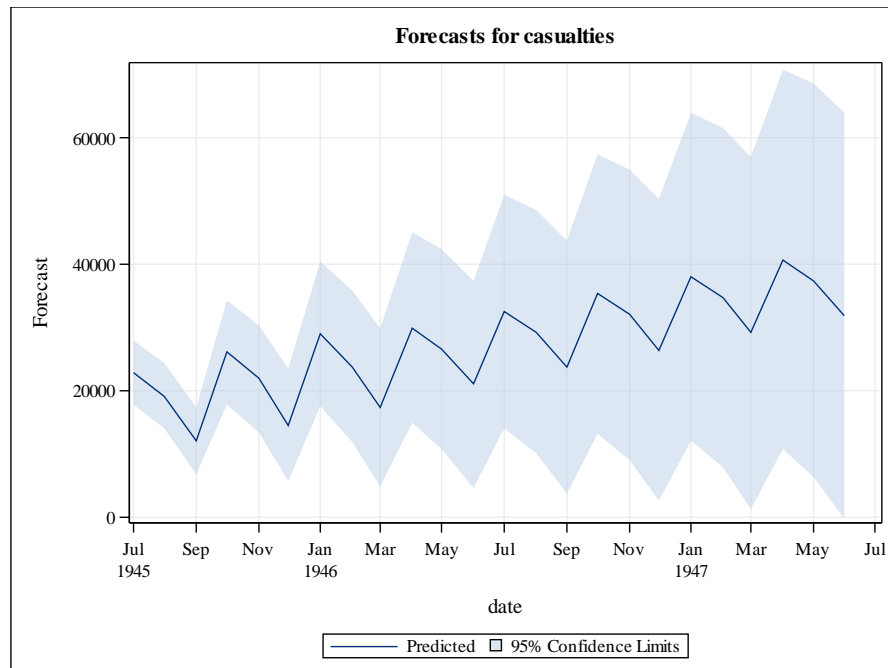
An ACF plot of the residuals will help us to understand whether or not the model is a good fit. It appears to be a good fit because none of the lags are high.

From this model fitting, the sum of the most extreme estimate for troop losses is 316,825.

However, the data point from May of 1942 could be disrupting the model. In an effort to produce a better estimate the data point was replaced by the average of April and June's casualties. Below is the result of removing the extreme data point and rerunning the new data with an appropriate model.







The most extreme estimate forecasted by the new model is 1,082,687.

Conclusion

Harry Truman's claim falls well within this bound. However, his claim that 500,000 lives would have been lost cannot be confirmed without an estimation as to how long it would have taken to achieve victory through invasion. While the argument cannot be made that one life is worth more than another, it is clear that the choice Truman made resulted in the least amount of deaths possible because the two bombs "only" killed 80,000 enemy civilians a number much smaller than any forecast of U.S. casualties. It was estimated that for every one U.S. soldier lost during the invasion of Okinawa, the Japanese lost nine soldiers.

The obtained estimates also show the limitations that need to be considered when using time series analysis. For some of the models the 95% confidence interval negative estimation as the lower bound. It does not make logical sense to have a negative number of people killed.

Further Research

It would be interesting to see what methodology was used to obtain the estimates of U.S. troop deaths in 1945. It would also be interesting to research how long the war was predicted to last in the Pacific theater if the atomic weapons had not been used.

Acknowledgment

I would like to thank Dr. Barbara Gannon, professor of History at the University of Central Florida, for suggesting this topic of research. I would also like to thank Dr. Liqiang Ni, professor of Statistics at UCF, for his guidance in exploring time series data. I would like to thank Kelcey Ellis, professor of Statistics at UCF, for her continued support and

encouragement. Lastly, I would like to thank Dr. Richards Plavnieks, professor of History at UCF, for his contributions to this paper and his efforts to review it.

References

Bowerman, Bruce L., and Richard T. Connell. *Forecasting, Time Series, and Regression: An Applied Approach*. 4th ed. Belmont, CA: Thomson Brooks/Cole, 2005.

Davison, John. *The Pacific War: Day by Day*. New York: Chartwell Books, 2011.

"HyperWar: Army Battle Casualties and Nonbattle Deaths in WW II [Intro/Summary]." HyperWar: Army Battle Casualties and Nonbattle Deaths in WW II [Intro/Summary]. Accessed March 15, 2016.
<http://www.ibiblio.org/hyperwar/USA/ref/Casualties/Casualties-Intro.html>.

[Intro/Summary]. Accessed April 26, 2015. Roosevelt Letter. Accessed May 1, 2015.
<http://www.fdrlibrary.marist.edu/archives/pdfs/docsworldwar.pdf>.

SAS/ETS(R) 9.2 User's Guide. (n.d.). Retrieved July 14, 2015, from
http://support.sas.com/documentation/cdl/en/etsug/60372/HTML/default/viewer.htm#etsug_arima_sect056.htm

Shaw, Antony. *World War II Day by Day*. New York: Chartwell Books, 2010.

Somerville, Donald. *World War II: An Authoritative Account of One of the Deadliest Conflicts in Human History with Analysis of Decisive Encounters and Landmark Engagements*. Hermes House, 2008.

"The Decision to Drop the Bomb." Ushistory.org. Accessed April 26, 2015.

"Understanding the Decision to Drop the Bomb on Hiroshima and Nagasaki." Understanding the Decision to Drop the Bomb on Hiroshima and Nagasaki. Accessed April 25, 2015.

Contact Information

Rachael Becker

Email: Leahcarbecker@knights.ucf.edu

Appendix

Original Dataset:

```
Data PacTheater;  
  input date MONYY7. casualties;  
datalines;  
Dec1941 1080  
Jan1942 818  
Feb1942 1005  
Mar1942 212  
Apr1942 1157  
May1942 28966  
Jun1942 231  
Jul1942 152  
Aug1942 156  
Sep1942 153  
Oct1942 168  
Nov1942 1046  
Dec1942 1758  
Jan1943 1821  
Feb1943 320  
Mar1943 282  
Apr1943 245  
May1943 1415  
Jun1943 383  
Jul1943 3679  
Aug1943 1692  
Sep1943 879  
Oct1943 476  
Nov1943 1269  
Dec1943 745  
Jan1944 1106  
Feb1944 1890  
Mar1944 3337  
Apr1944 1394  
May1944 1819  
Jun1944 6247  
Jul1944 4439  
Aug1944 2107  
Sep1944 2206  
Oct1944 6058  
Nov1944 5777  
Dec1944 6513  
Jan1945 8252  
Feb1945 12744  
Mar1945 9077  
Apr1945 20266  
May1945 15879  
Jun1945 7830  
Jul1945 1332  
Aug1945 683  
Sep1945 46  
Oct1945 10  
Nov1945 13  
Dec1945 4  
;  
run;
```

Code for estimation based from data gathered up until and including May of 1945:

```
data PacMay; set PacTheater; if date>-5358 then delete; Format date MONYY7.;
run;
Data FirstOrder2;
set PacMay;
diff1=DIF(casualties);
run;
ods rtf file = 'desktop\ARIMA_EXAMPLE2.rtf';
  Proc Arima data = FirstOrder2;
    identify var=diff1;
    estimate q=(1)(12) noint method=ml;
    forecast id=date interval=month printall out=ForcastedPac;
  run;
ods rtf close;
```

Revised code:

```
Data PacTheater2;
  input date MONYY7. casualties;
datalines;
Dec1941 1080
Jan1942 818
Feb1942 1005
Mar1942 212
Apr1942 1157
May1942 28966
Jun1942 231
Jul1942 152
Aug1942 156
Sep1942 153
Oct1942 168
Nov1942 1046
Dec1942 1758
Jan1943 1821
Feb1943 320
Mar1943 282
Apr1943 245
May1943 1415
Jun1943 383
Jul1943 3679
Aug1943 1692
Sep1943 879
Oct1943 476
Nov1943 1269
Dec1943 745
Jan1944 1106
Feb1944 1890
Mar1944 3337
Apr1944 1394
May1944 1819
Jun1944 6247
Jul1944 4439
Aug1944 2107
Sep1944 2206
Oct1944 6058
```

Nov1944	5777
Dec1944	6513
Jan1945	8252
Feb1945	12744
Mar1945	9077
Apr1945	20266
May1945	15879
Jun1945	7830

```

;
run;
Data Adjusted;
set PacTheater2;
    if Date = -6454 then casualties = (1157+231)/2;
run;

Proc Arima data = adjusted;
    identify var=casualties (1,3);
    estimate q=(1)(12) noint method=ML;
    forecast id=date interval=month printall out=ForecastedPac2;
run;

```

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies