Opioid Epidemic: Mortality Rate Forecasting in the State of Maryland

Alan Camarena (Perm: 8602062)

December 13, 2017

Introduction:

Introducing My Data

The following dataset that I worked on contains the amount of opioid drug induced deaths per month in the State of Maryland from the years 1999-2015

Over the past 7 years, the United States has been facing a drug epidemic problem. My data acknowledges the amount of deaths due to drug overdoses correlated directly to opioid use in the state of Maryland. Maryland is currently one of the states who is facing a unique difficulty with this epidemic. The inspiration to source this data came to my interest after watching a documentary on Netflix by the name of 'Heroin(e)'. This documentary follows three women; a fire chief, a judge and a street missionary - as they battle West Virginia's devastating opioid epidemic. The first responders are typically the fire department. They get there as soon as they receive notice that an individual is overdosing. Once the first responders are present, they administer a drug that goes by the name of Naloxene. Naloxene once administered, reverses opioid overdosing. It is essential for any firefighter, paramedic or even police officers to have many of these available. A unique problem that I found within this documentary was that many counties did not have enough personnel to respond as quickly as they could to overdoses.

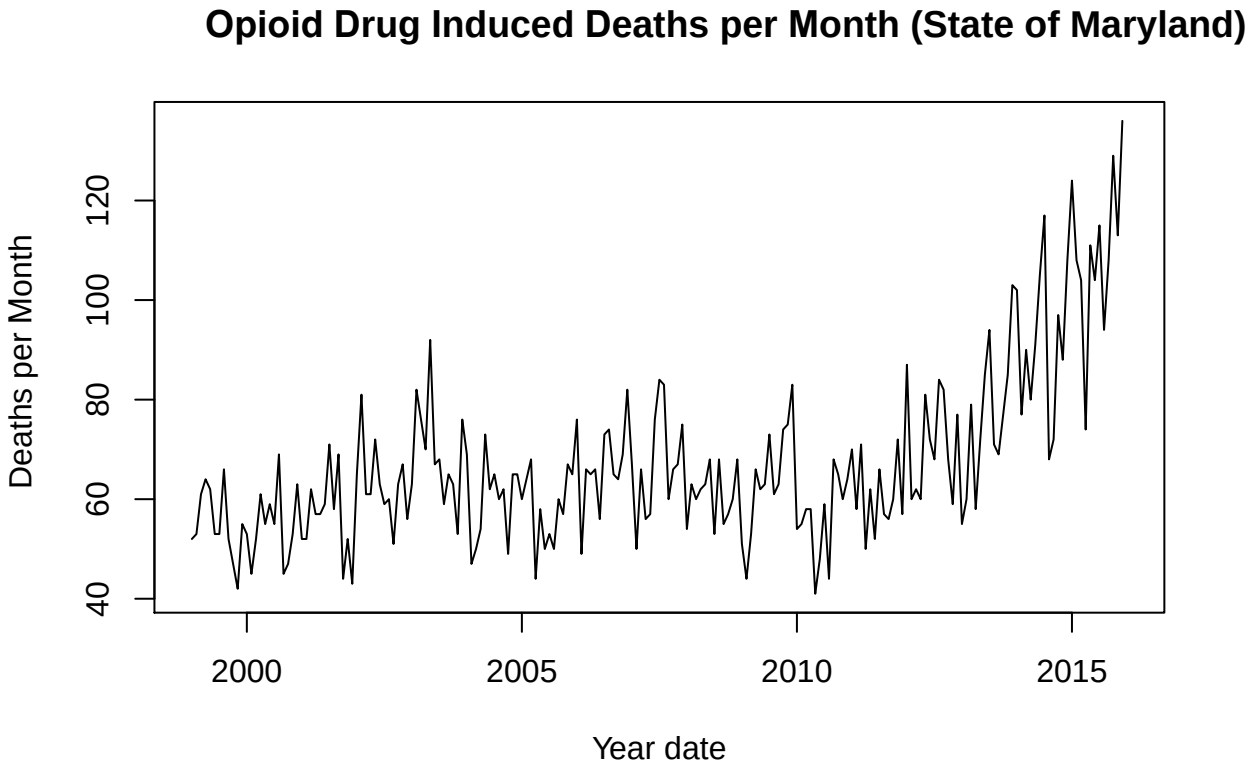
Abstract

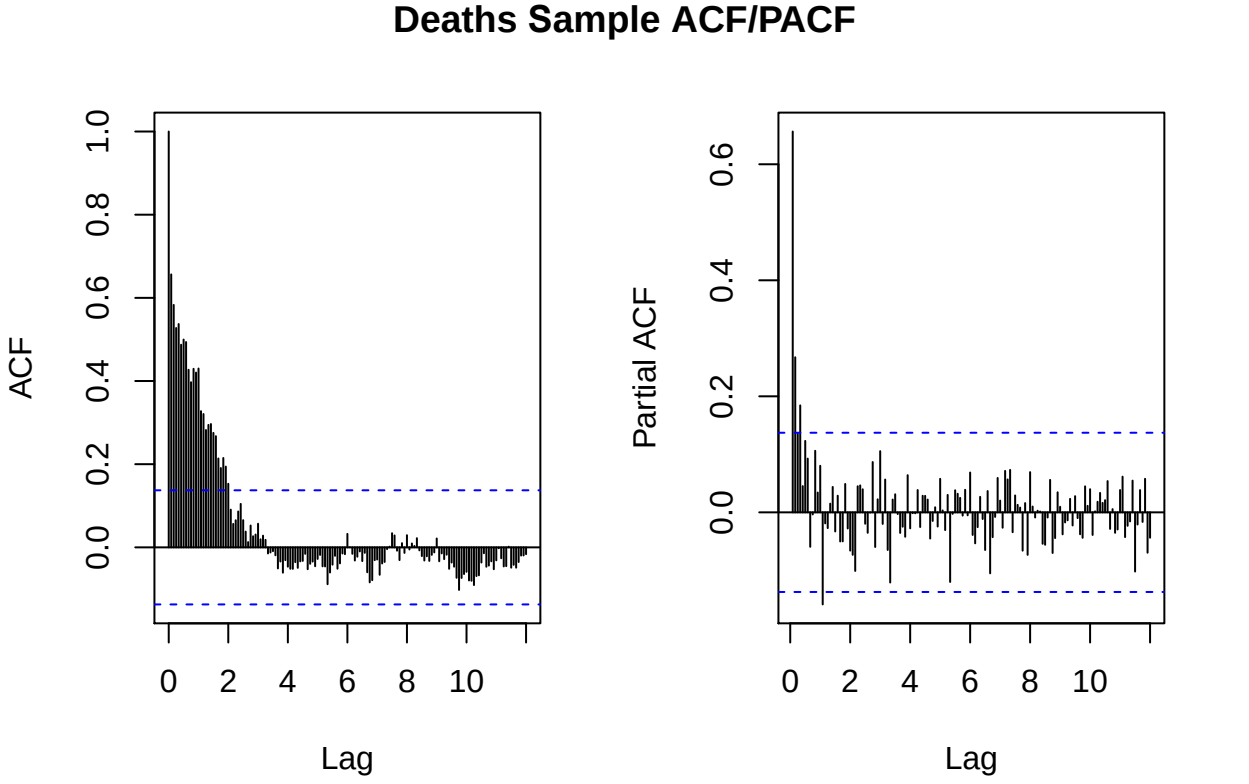
In my time series analysis, I want to address the amount of deaths that are probable to occur within the next year. By producing this information, this can alert officials in office to expand budgetary needs to first responders and to consider proper legislature to a problem that is affecting many communities in the State of Maryland.

The conclusion in my forecasting demonstrates that there is a high probability for more deaths to occur within the next year, the amount of deaths per months is still in range of 120 deaths per month. A 95% confidence interval was used in our prediction and it estimates a low limit of 70 death per month and a high limit of 250 deaths per month. I arrived through this conclusion from applying a set of time series techniques learned from my PSTAT 174 course. I manipulate the data, and apply the correct diagnostics to demonstrate that the model I built indeed works. There were small issues with normality, but regardless I was able to choose the best fitting model to predict the amount of opioid overdose deaths in the State of Maryland

I would like to acknowledge the CDC (Center for Control and Disease) for supplying that collected data, Professor Feldman, Teaching Assistant Zhipu Zhou and RStudio Software for making this project possible.

Orginial Data:





The variance is not constant throughout the time series, there are areas that have bigger ranges and others with smaller ranges.
It is also apparent that there is an increasing trend from the year 2010-2015, showing that there is an apparent change in behavior in comparison to the previous years. Neither do the time series or the acf/pacf plots insinuate seasonality.

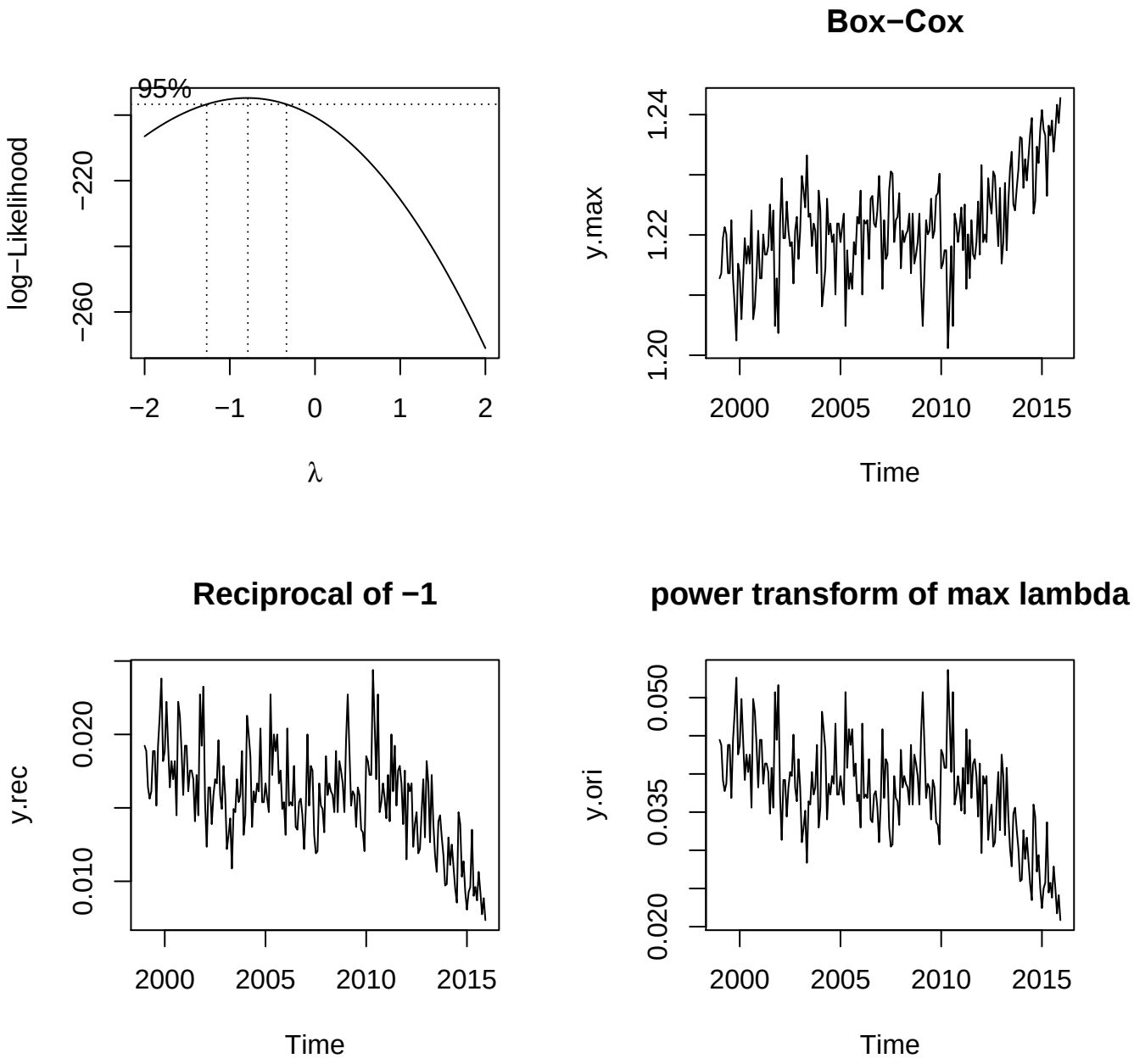
Steps to consider:

1) Transformation to lower variability
2) Difference at lag 1 to get rid of increasing trend

3) There is no seasonality present

Mean and Variance of original data. Mean: 67.1668 Variance:289.6864

Transformation:



In order to find the optimal power transformation for my time series, I employ the Box-Cox test.
I'm allowed to use anything within the confidence interval, thus I consider 3 different transformations.

1)Consider using the max value with Box-Cox formula.
2)Consider the power transformation of max lambda.
3)Consider the power transformation of -1.

Conclusion: The smallest obtained variance was acquired through the power transformation of -1, however, in the forecasting step, my data would be upside down, and in order to avoid this, I use the box-cox formula which also shows that the variance lowers and becomes stable(very similar to the power transformation of -1).

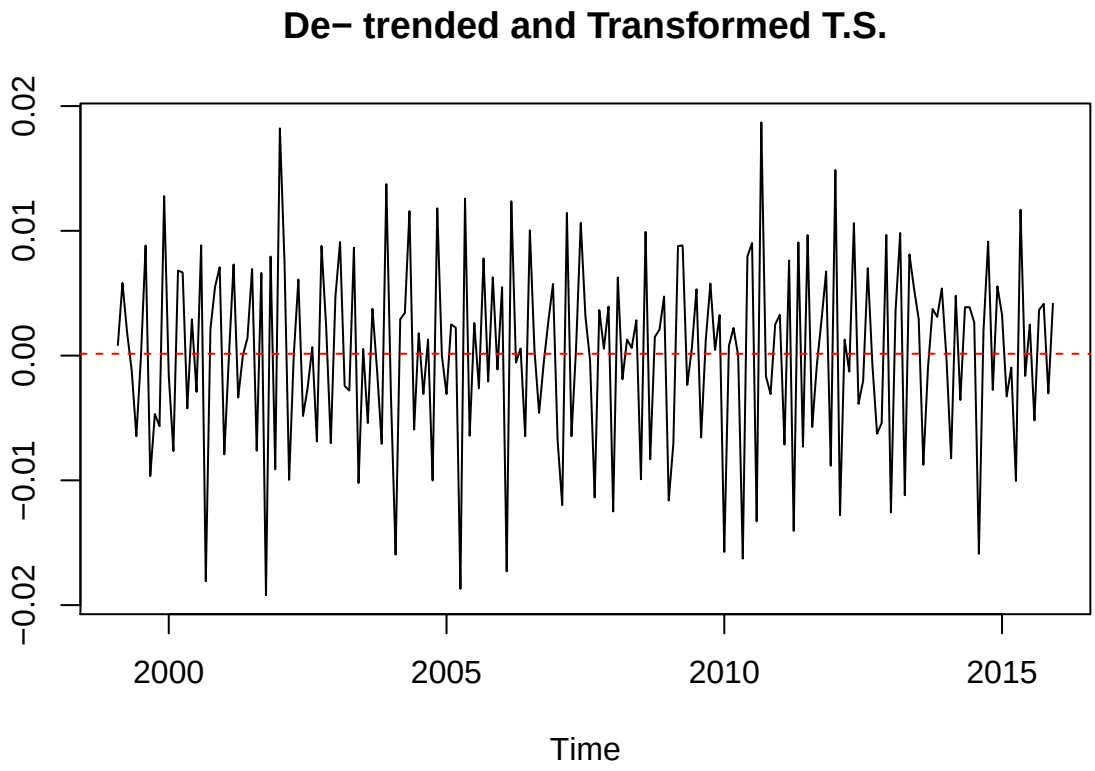
My variance is now stable yet not stationary, therefore my next step is to difference.

Box-Cox Formula:

$$y_i^\lambda = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} & x \neq 0 \\ log(y_i) & x = 0 \end{cases}$$

Variance of transformed data: .0000616034

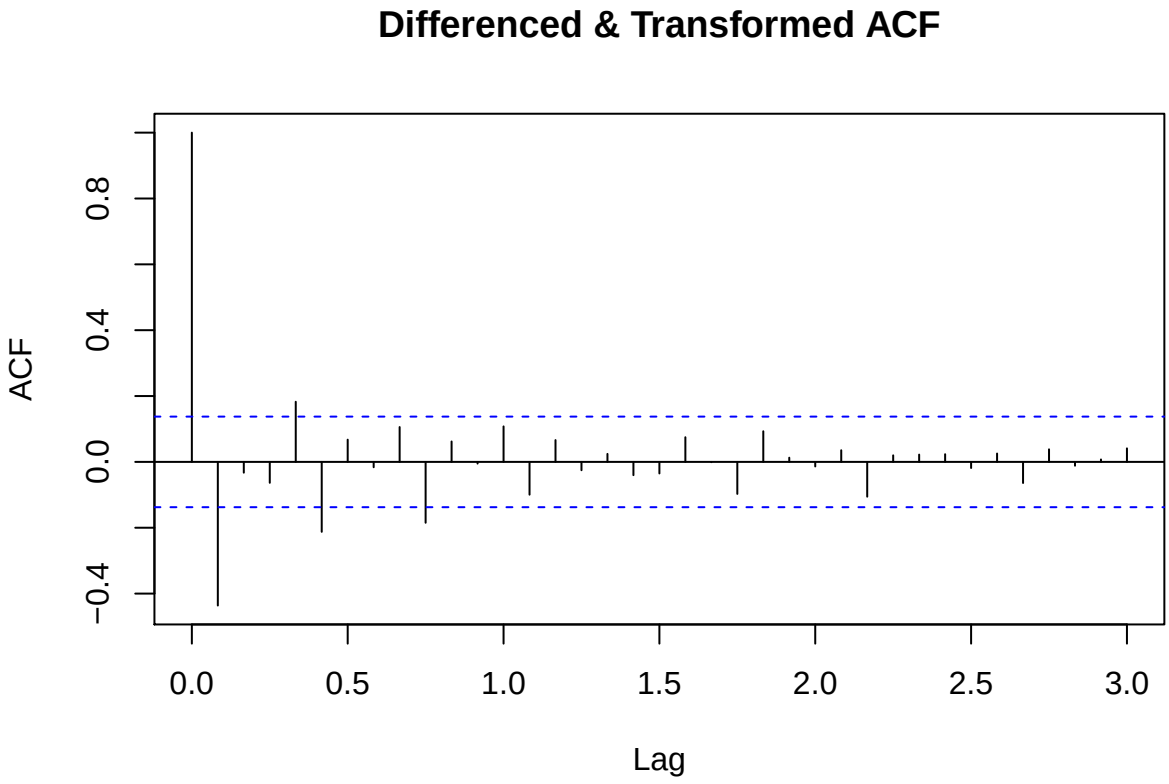
Differencing:

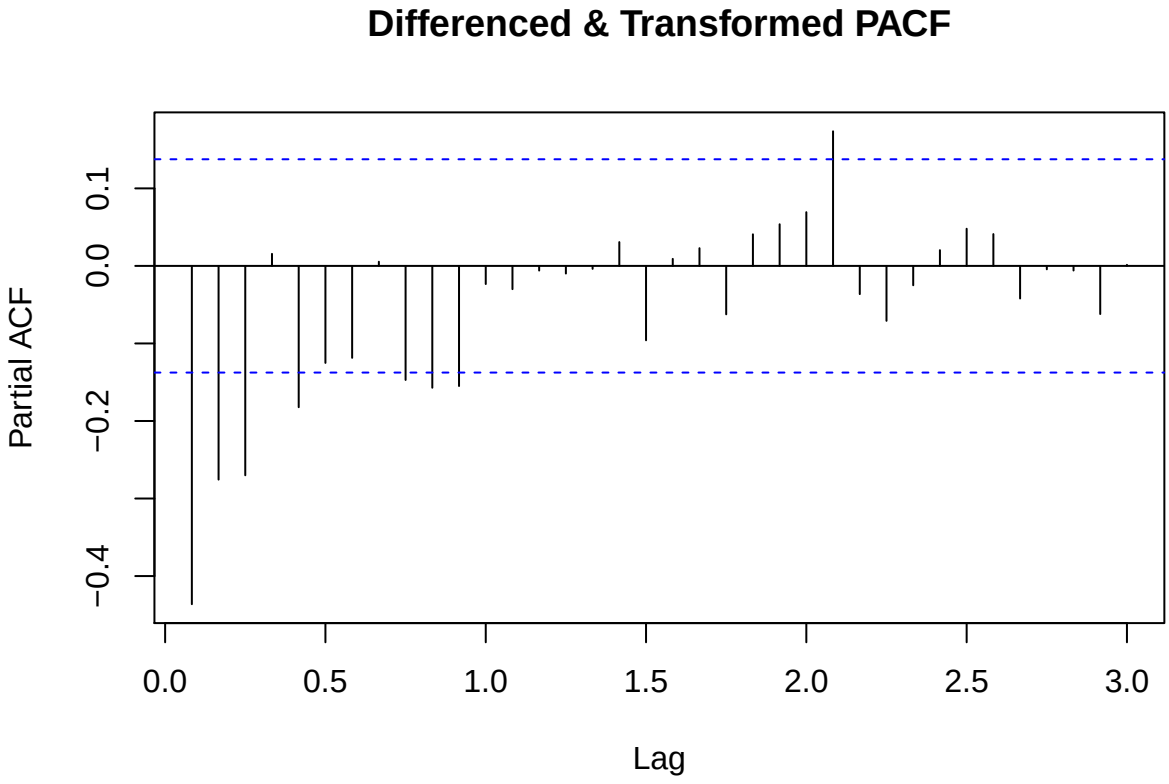


There seems to be no seasonality that is apparent, therefore I difference at lag 1 to remove an increasing trend component.
Plotting the time series, we can tell that it is now stationary. Note: Mean is almost 0 (Dashed red line)

I attempt to difference one more time to see if I can improve my time series but the variance only increases, therefore only one difference is required.

Model identification ~ Analysis of ACF/PACF





ACF/PACF Behavioral Analysis :

ACF: The lags decay and die out at lag 9

PACF: The lags decay and die at lag 11, however, there is a resurgence of another lag at (25) and cuts off

Conclusions: It seems that both the ACF/PACF tail off, thus I suspect that this is an arma model.
From consulting with Professor Feldman, I was suggested that this could also be an MA(9) model.

Consider:
1) MA(9)
2) AR(11) Given that the PACF cuts off after 11 lags

Model Estimation

```
##    q
## p           0         1         2         3         4         5         6
##   0 -1412.180 -1493.231 -1494.020 -1492.394 -1490.312 -1492.474 -1492.701
##   1 -1452.862 -1494.325 -1492.530 -1494.886 -1493.154 -1492.468 -1490.823
##   2 -1466.778 -1492.477 -1494.023 -1493.415 -1491.314 -1489.710 -1494.051
##   3 -1480.245 -1490.382 -1492.312 -1491.363 -1489.485 -1497.775 -1495.565
##   4 -1478.166 -1490.040 -1492.735 -1490.739 -1493.487 -1495.565 -1493.476
##   5 -1482.825 -1494.507 -1492.340 -1493.012 -1495.346 -1490.208 -1491.320
##   6 -1483.896 -1492.340 -1490.774 -1490.959 -1493.382 -1491.061 -1489.143
##   7 -1484.712 -1490.306 -1488.927 -1488.888 -1491.126 -1489.938 -1486.866
##   8 -1482.522 -1488.146 -1486.792 -1489.320 -1489.130 -1488.534 -1486.903
##   9 -1484.803 -1491.553 -1490.272 -1488.441 -1486.182 -1484.047 -1481.924
##    q
## p           7         8         9
##   0 -1490.796 -1488.622 -1486.493
##   1 -1488.608 -1488.362 -1486.599
##   2 -1487.822 -1489.785 -1487.526
##   3 -1493.329 -1490.887 -1488.915
##   4 -1491.299 -1489.070 -1487.022
##   5 -1488.992 -1486.731 -1484.487
##   6 -1485.897 -1488.059 -1486.975
##   7 -1487.305 -1481.764 -1491.115
##   8 -1485.989 -1480.134 -1485.979
##   9 -1479.557 -1482.015 -1486.114
```

At first, I execute two (AR) functions that automatically chose the amount of coefficients based off of the lowest AIC values. Both "MLE" & "Yule-Walker" method return an AR(11) model. Confirming my analysis of the PACF.

After fitting AR(11) to a model, it returns an AICc of [-1490.904].
After fitting MA(9) to a model, it returns an AICc of [-1486.493].

I then decide to run a for-loop, to have it estimate the numer of coefficients that would best suit my model. The for lop returns AICc values.

Based off the AICc values, the lowest AICc found an ARMA(3,5) model.

Consider the following 3 models chosen on basis of AICc:
1) ARMA(3,5) [-1497.775].
2) ARMA(1,3) [-1494.886].
3) ARMA(1,1) [-1494.325].

After comparing my previous models [AR(11) & MA(9)] to my new models obtained through the loop, I decided to leave AR(11) & MA(9) out of consideration on the basis of their AICc values and through rule of parsinomy.

Model Diagnostics:

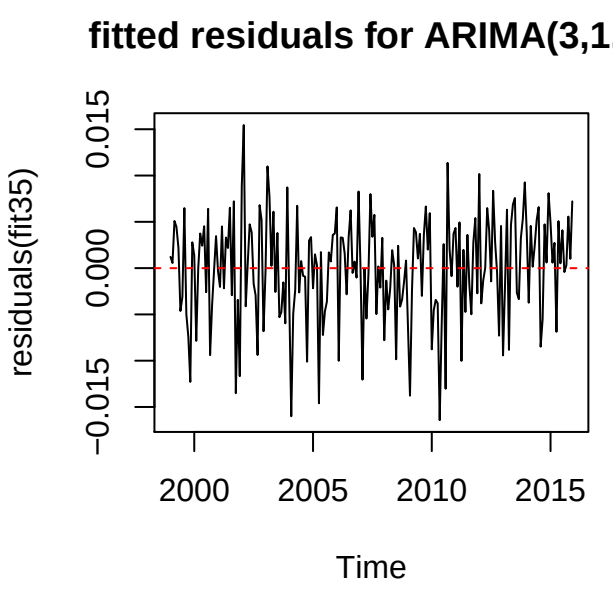
Models that will be considered for diagnostics:

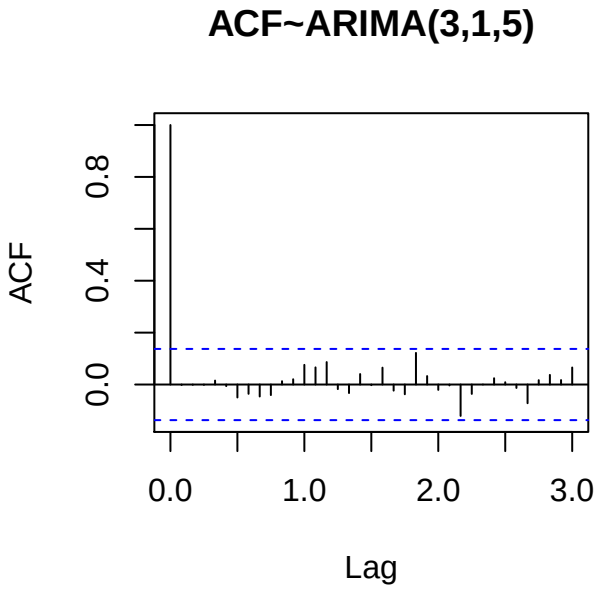
1)ARIMA(3,1,5)
$X_t + 0.8136X_1 + 0.8489X_2 + 0.8208X_3 = 0.1066Z_1 + 0.2348Z_2 + 0.1382Z_3 - 0.6642Z_4 - 0.2763Z_5 + Z_t$

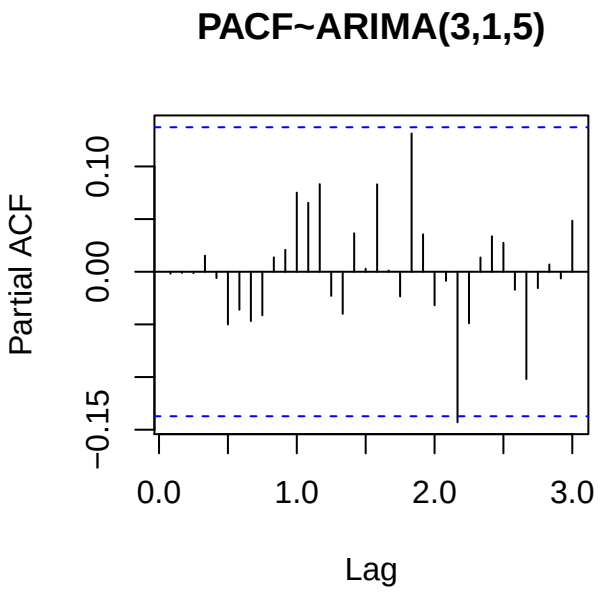
2)ARIMA(1,1,3)
$X_t + 0.9132X_1 + = 0.2194Z_1 - 0.7379Z_2 - 0.1953Z_3 + Z_t$

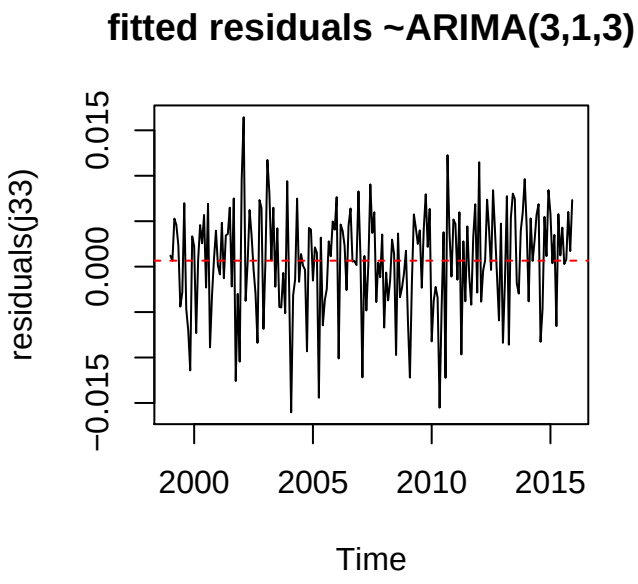
3)ARIMA(1,1,1)
$X_t - 0.1359X_1 = -0.8474Z_1 + Z_t$

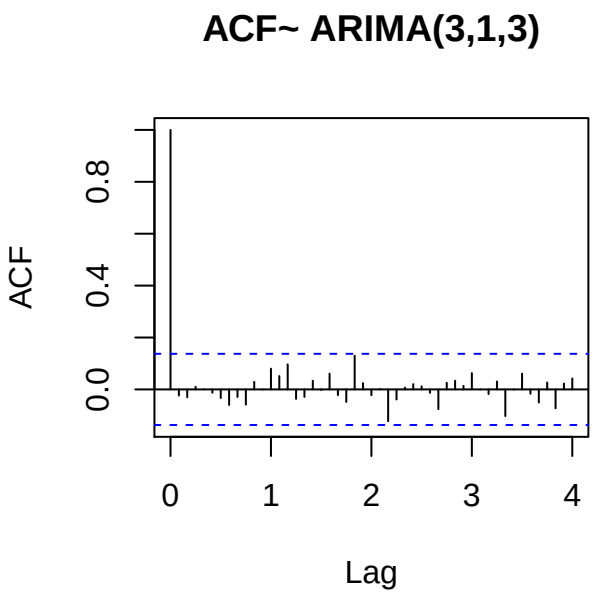
First I run diagnostics on my ARIMA(3,1,5) model

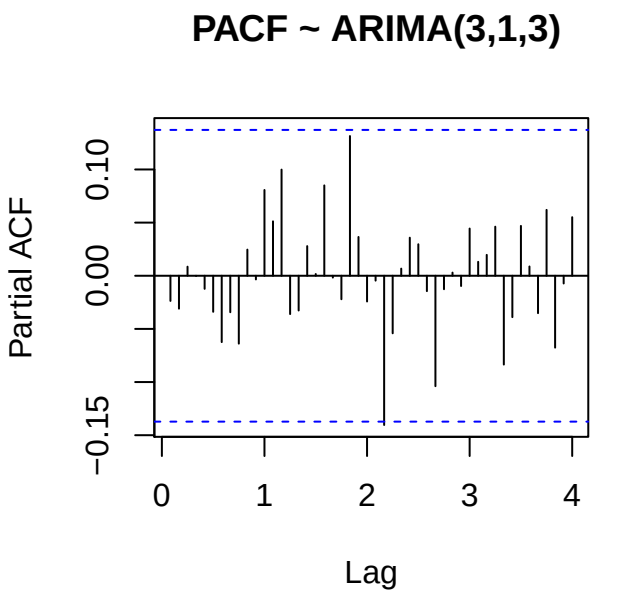


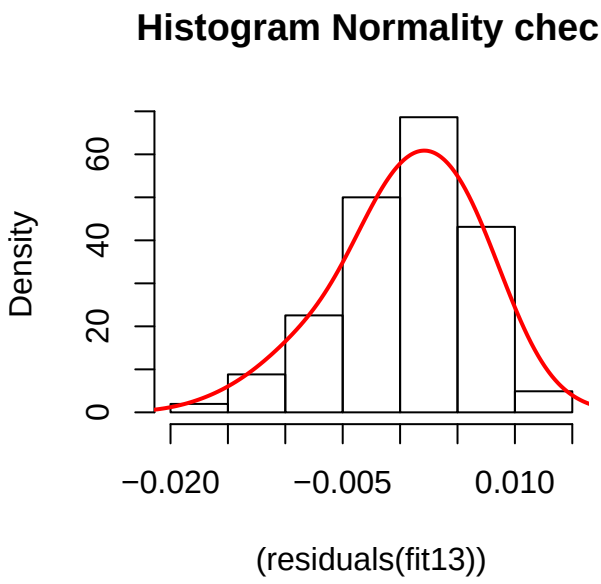


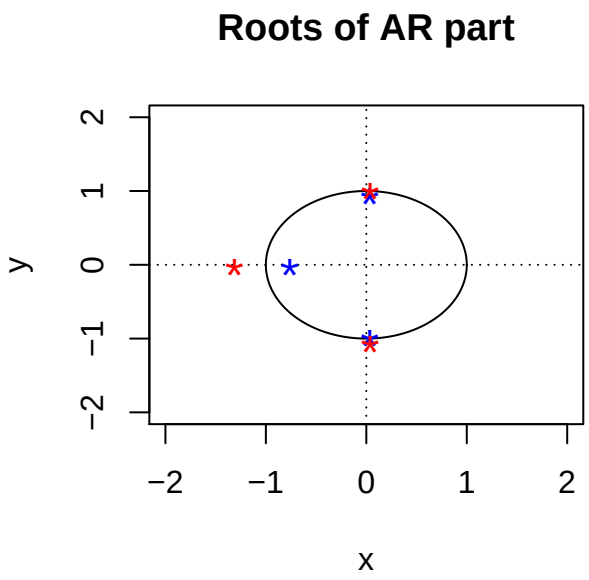

PACF~ARIMA(3,1,5)
Partial ACF
0.10
0.00
−0.15
0.0
1.0
2.0
3.0
Lag

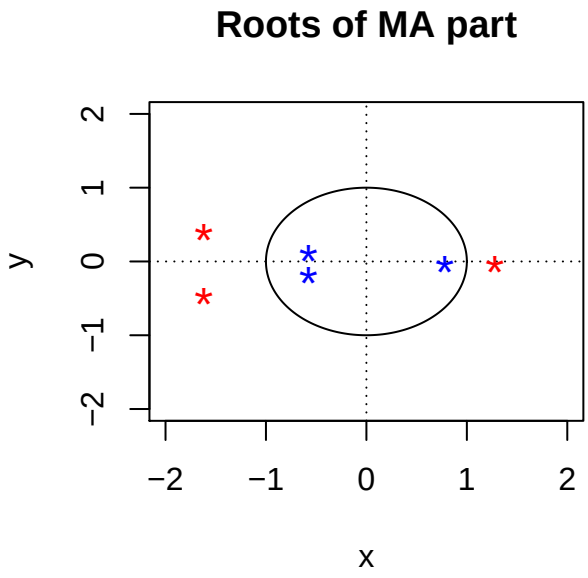

fitted residuals ~ARIMA(3,1,3)
residuals(j33)
0.015
0.000
−0.015
2000
2005
2010
2015
Time


ACF~ ARIMA(3,1,3)
ACF
0.8
0.4
0.0
0
1
2
3
4
Lag


PACF ~ ARIMA(3,1,3)
Partial ACF
0.10
0.00
−0.15
0
1
2
3
4
Lag








```
##
##  Box-Pierce test
##
## data:  residuals(j33)
## X-squared = 6.094, df = 8, p-value = 0.6367
```

```
##
##  Box-Ljung test
##
## data:  residuals(j33)
## X-squared = 6.4995, df = 8, p-value = 0.5915
```

```
##
##  Box-Ljung test
##
## data:  residuals(j33)^2
## X-squared = 8.3591, df = 14, p-value = 0.8698
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(j33)
## W = 0.98001, p-value = 0.005298
```

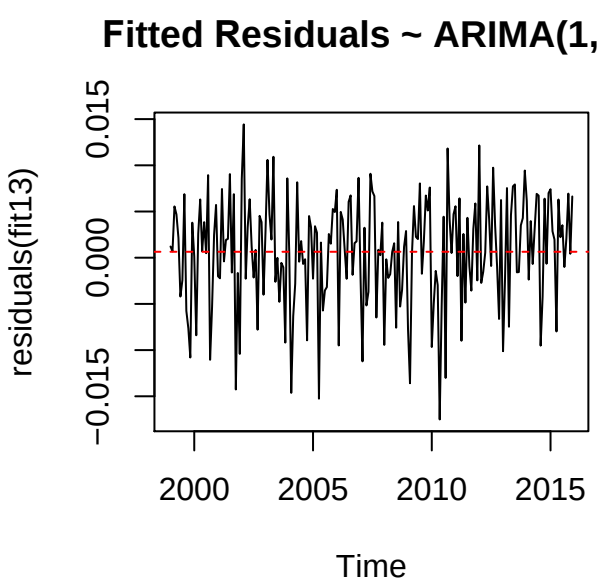
Findings:

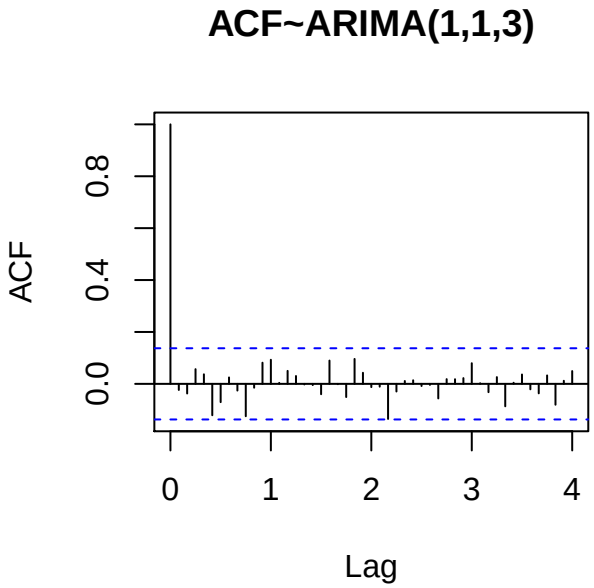
First I check to see if my residuals are stationary and uncorrelated.
Everything looks fine in my acf, but my pacf has a residual that is just over the confidence interval, therefore I have one significant value that can be included.

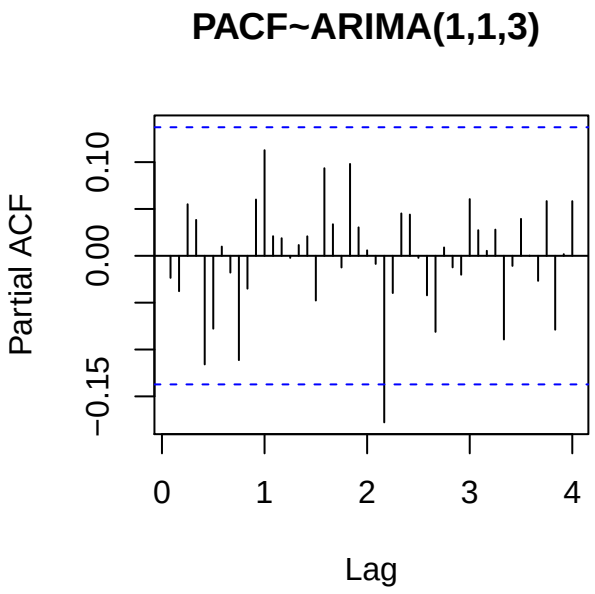
I zero out some coefficients that have a 0 associated with their confidence interval, and my model improves to ARIMA(3,1,3) with a total of 6 coefficients now, my AICc does imrpove and lowers to [-1500.68].

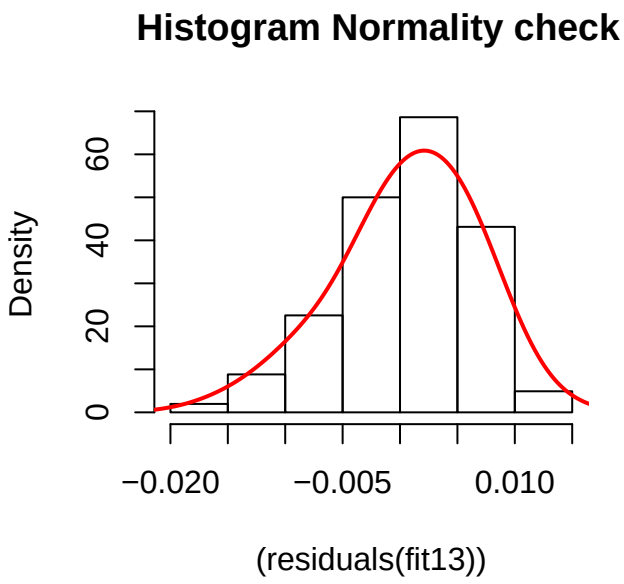
New ARIMA(3,1,3) model: New model is Casual and Invertible.
I again run my residuals plot, ACF/PACF and find that the value of the PACF at lag (26) has some improvement and is barely touching the confidence interval line.

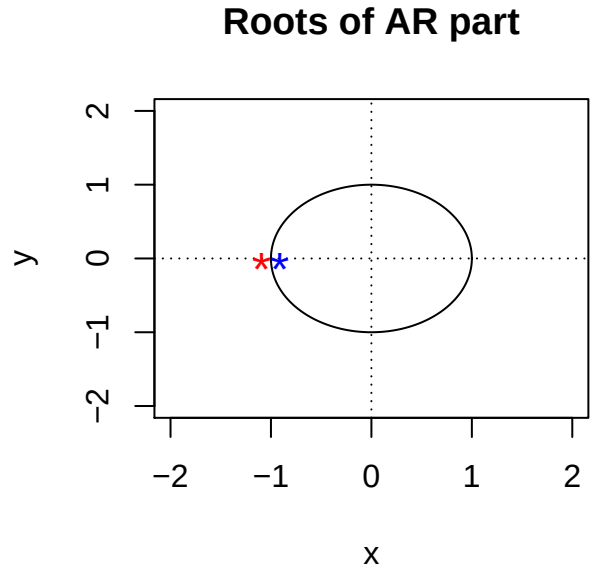
I decide to run another set of diagnostics, and I find that everything passes with a high p-value except for my shapiro-wilk test.
This insinuates that my data is not normal. I move on to my next model in consideration, ARIMA(1,1,3)

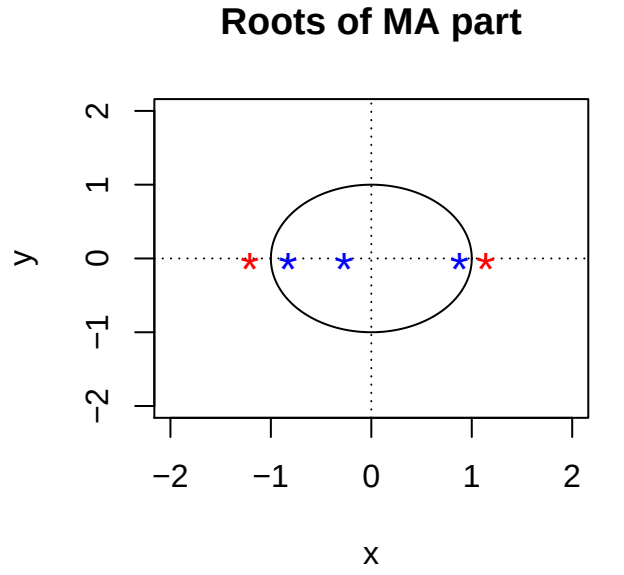













```
##
##  Box-Pierce test
##
## data:  residuals(fit13)
## X-squared = 12.443, df = 10, p-value = 0.2565
```

```
##
##  Box-Ljung test
##
## data:  residuals(fit13)
## X-squared = 13.079, df = 10, p-value = 0.2193
```

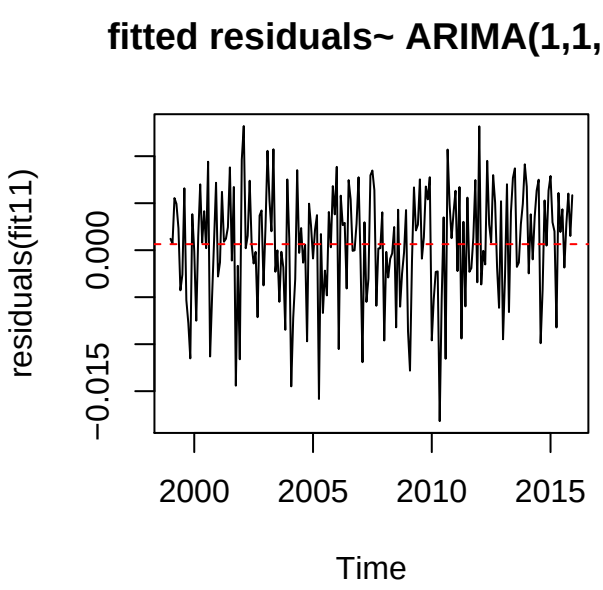
```
##
##  Box-Ljung test
##
## data:  residuals(fit13)^2
## X-squared = 9.4626, df = 14, p-value = 0.8003
```

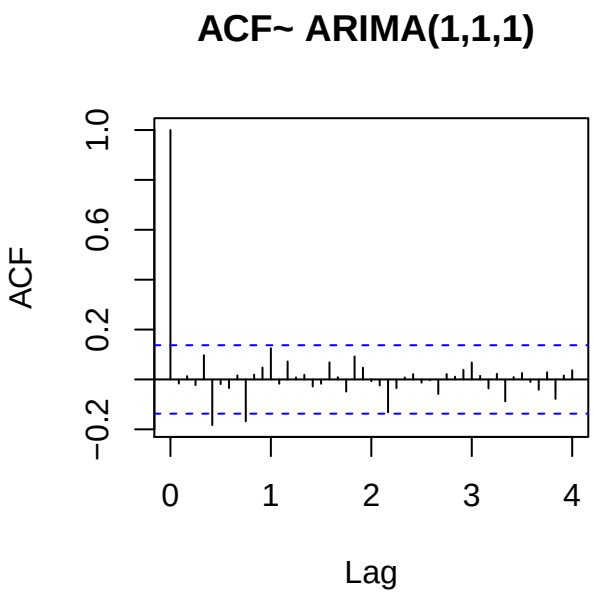
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit13)
## W = 0.97662, p-value = 0.001764
```

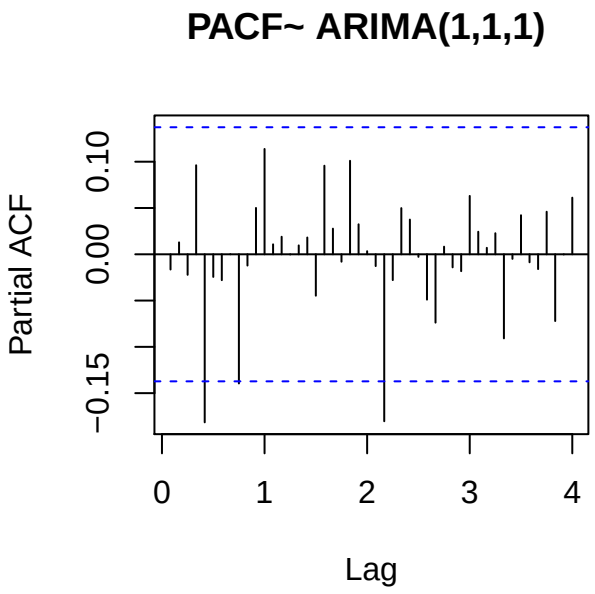
Findings:

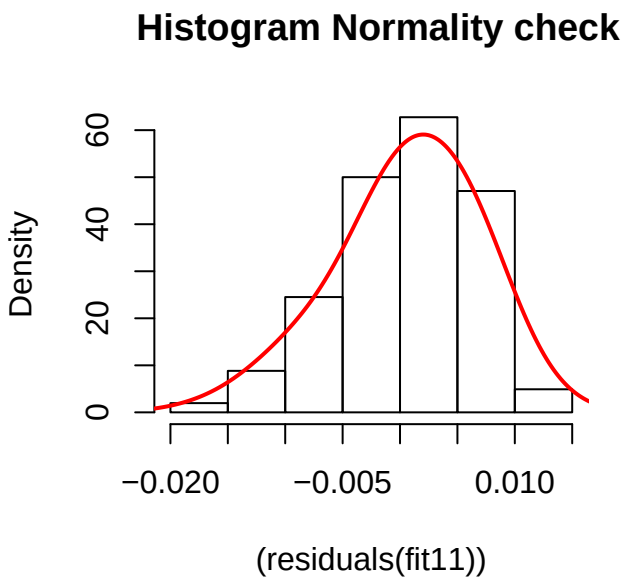
First I check to see if my residuals are stationary and uncorrelated.
Both my ACF/PACF how a residual that is far out of the confidence interval, coincidentally the same value as in my previous model.
I attempt to zero out coefficients, but they're all significant.

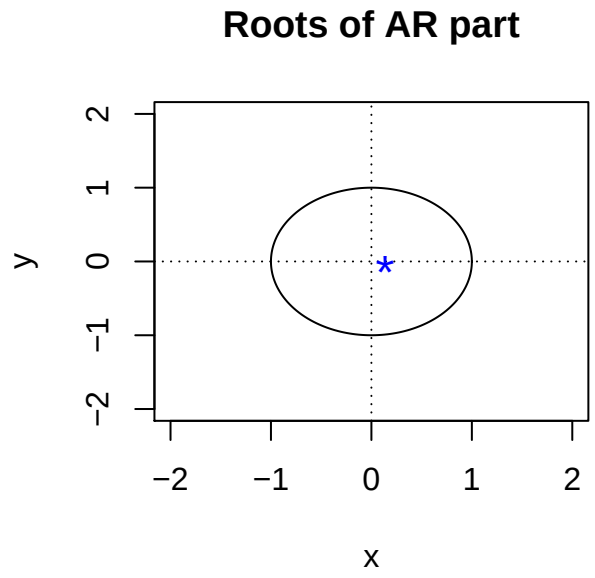
Invertibility and casualness is ok, all roots are outside of unit circle. I decide to run my set of diagnostics and find that everything passes, however my shapiro wilk test does not and therefore my data is not normal. I move on to my next model in consideration: ARIMA(1,1,1).

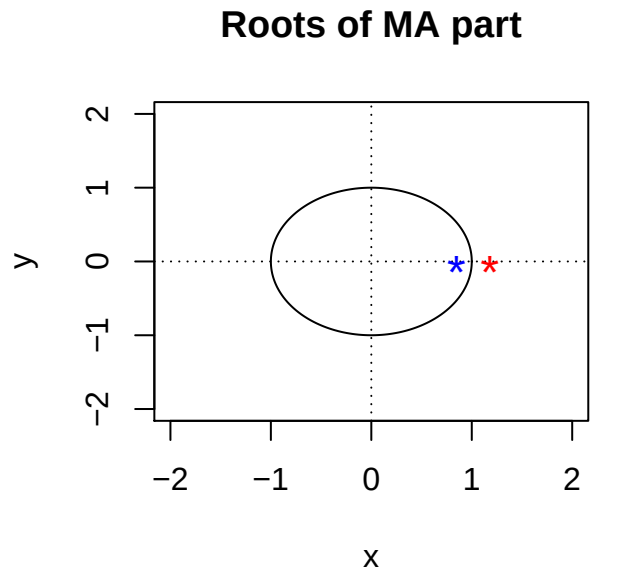













```
##
##  Box-Pierce test
##
## data:  residuals(fit11)
## X-squared = 19.911, df = 12, p-value = 0.06879
```

```
##
##  Box-Ljung test
##
## data:  residuals(fit11)
## X-squared = 20.916, df = 12, p-value = 0.05163
```

```
##
##  Box-Ljung test
##
## data:  residuals(fit11)^2
## X-squared = 8.0341, df = 14, p-value = 0.8875
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit11)
## W = 0.97436, p-value = 0.0008748
```

Findings:

First I check to see if my residuals are stationary and uncorrelated.
Both my ACF/PACF have 3 residual that are far out of the confidence interval, I attempt to zero out a coefficient which does nothing to my AICc, and in turn fails all of my diagnostic test, therefore, I decide to keep the model as ARIMA(1,1,1)

Invertibility and casualness is ok, all roots are outside of unit circle. I decide to run my set of diagnostics with ARIMA(1,1,1) and find the same result as the previous models, normality does not pass.

Conclusion:

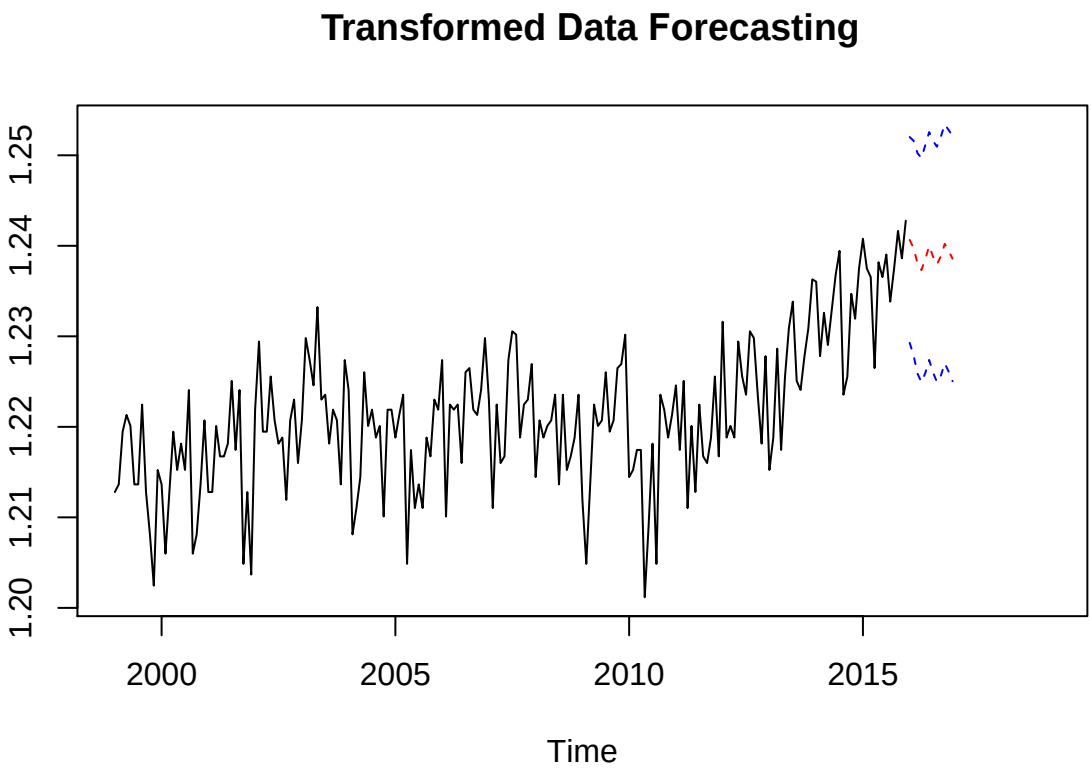
After discarding ARIMA(1,1,1) model, it leaves me to decide between two models on my list.
1) ARIMA(3,1,3)
2) ARIMA(1,1,3)
I exclude ARIMA(1,1,3) because its residual was too significant, also, it's AICc was -1494.886 in comparison to ARIMA(3,1,3) [-1500.68]

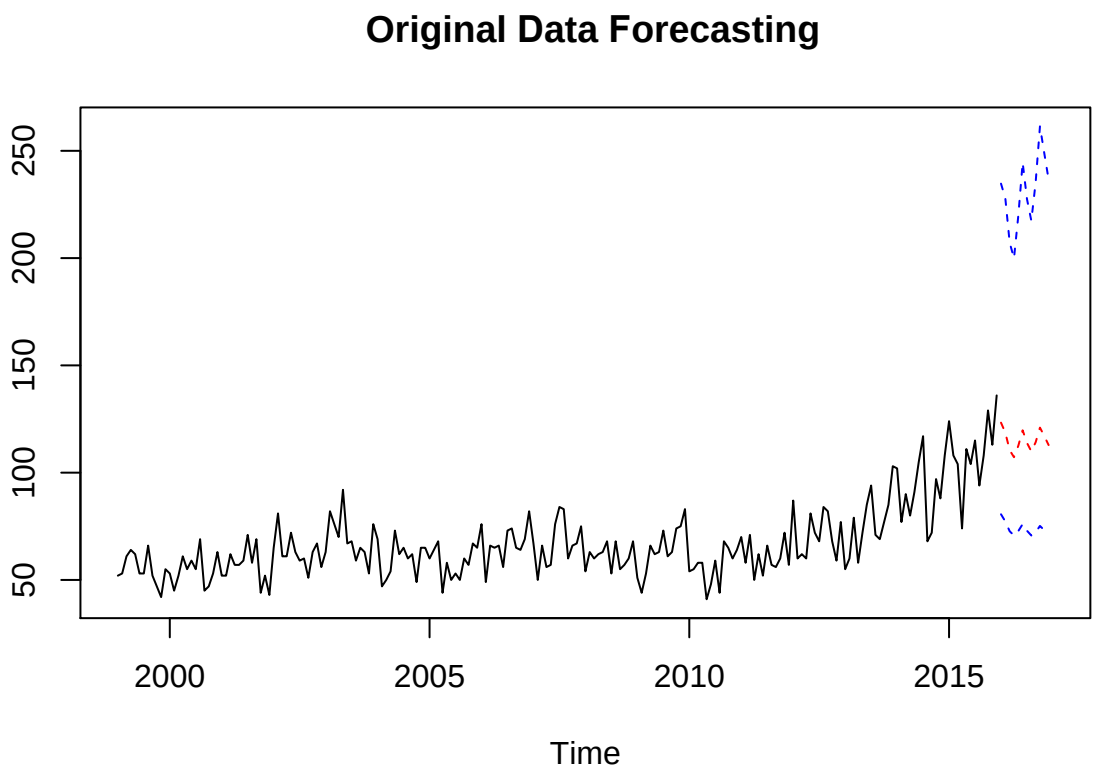
After careful consideration, based off of MLE & AICc I decide that my best model is ARIMA(3,1,3).

Model: $X_t + 0.6879_1 + 0.8715X_2 + 0.7006X_3 = 0.38Z_1 - 0.5383Z_2 - 0.2786Z_3 + Z_t$

Forecasting Stage:

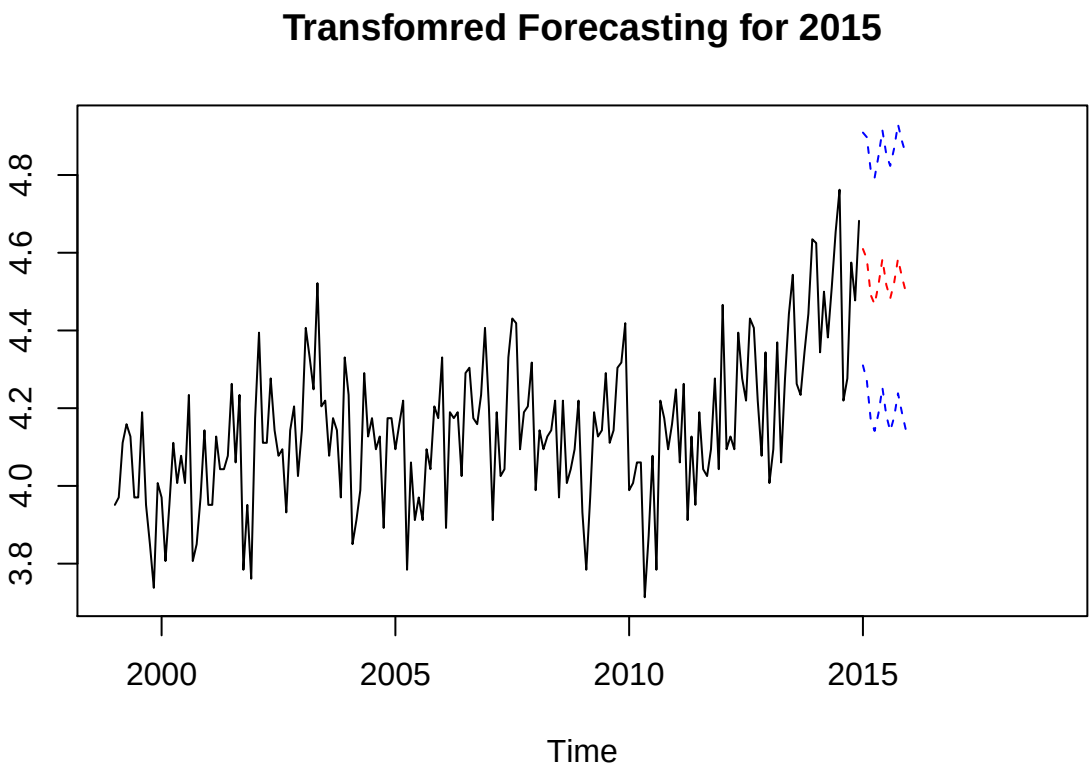
The following graphs will show forecasting for the year 2016. Note: Original data is from (1999-2015).
I will first forcast my transformed data and I will then reverse the number and produce the same results with my non transformed time series.

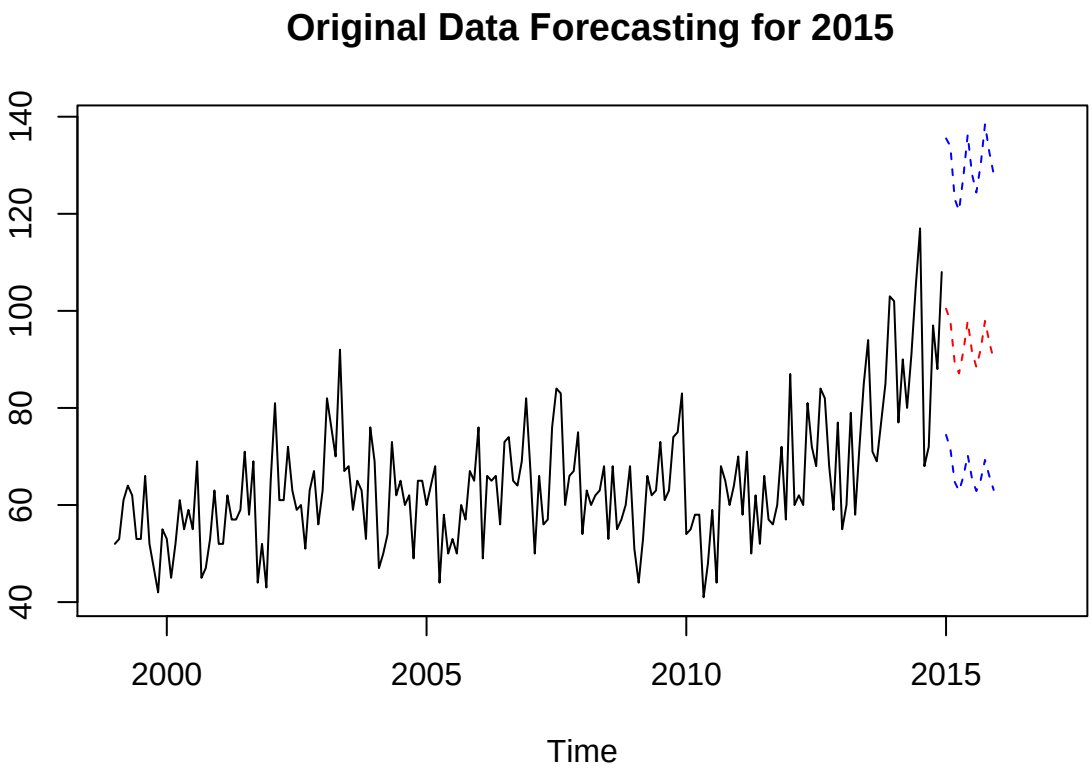

Transformed Data Forecasting
1.20
1.21
1.22
1.23
1.24
1.25
2000
2005
2010
2015
Time

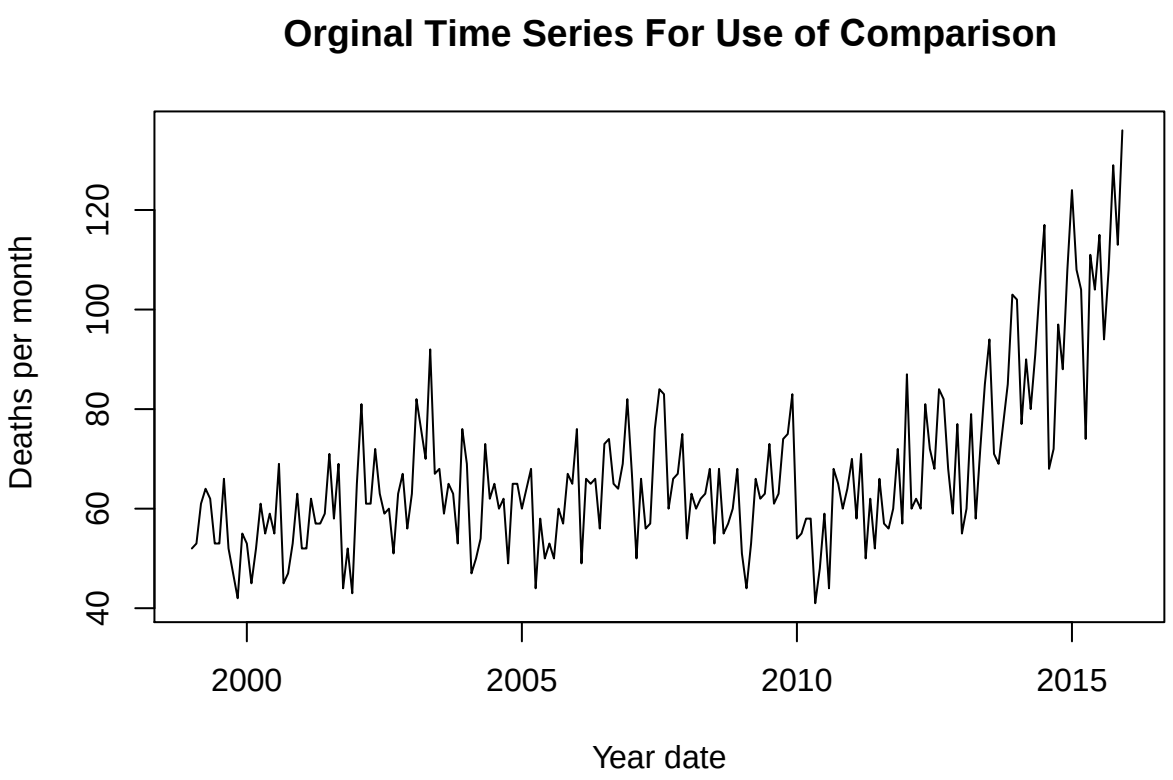


My forecasting seems to predict that the deaths in 2016 will reduce very sligthly.

In order to validate my forecasting for 2016, I will remove 12 obervations (2015 data) and forecast 2015. The following graphs will show a red line which is my prediciton and blue lines that are my confidence intervals. My confidence intervals represent the range in which my prediction can also be valid.


Transfomred Forecasting for 2015
3.8
4.0
4.2
4.4
4.6
4.8
2000
2005
2010
2015
Time


Original Data Forecasting for 2015
140
120
100
80
60
40
2000
2005
2010
2015
Time



Although our predicting line does not neccessarily follow our originals data prediction, our upper confidence interval does follow through and enable every value to be a possibility, therefore we can conclude that this model is adequate enough to make a prediction.

Conclusion:

After applying box-jenkins methodology in my project, I was able to find a model that was adequate for my data. My model predicts the amount of deaths that will occur within 2016. I removed values from the original time series and predicted 2015(the information that I already had). My forecasting for 2015 was similar to the original data, and therefore I concluded that ARIMA(3,1,5) was adequate enough for forecasting 2016.

I did encounter one problem in my diagnostics: none of my models that I chose passed the Shapiro Wilk test for normality, therefore I had to choose the best model I could based off of other diagnostic and criteria (Ljung, McLeod-Li, AICC, MLE)

Final Model: ARIMA(3,1,5) $X_t + 0.8136X_1 + 0.8489X_2 + 0.8208X_3 = 0.1066Z_1 + 0.2348Z_2 + 0.1382Z_3 - 0.6642Z_4 - 0.2763Z_5 + Z_t$

References

List of references:

Professor Feldman [Lecture notes, Lecture slides, Lecture, Office Hours] Thank you, everything was very helpful.
T.A. Zhipu Zhou [Section, Office Hours]

CDC Website https://www.cdc.gov/
Jessica Beshir [Director and Writer of Heroin(e) Documentary]

Appendix

This includes all R code used for this project.