Joe Kinderman

Armed Robberies Time Series Analysis

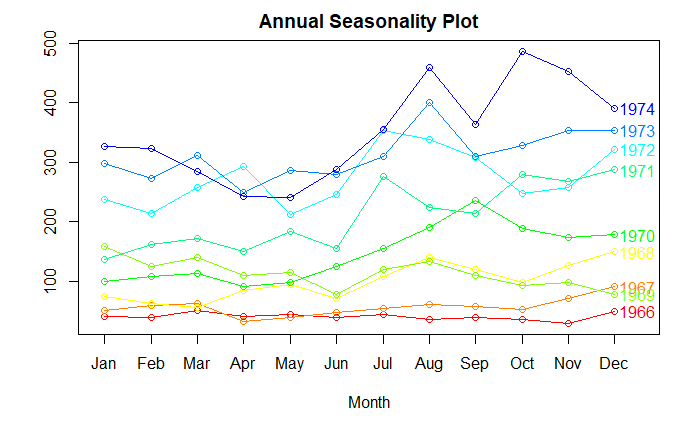
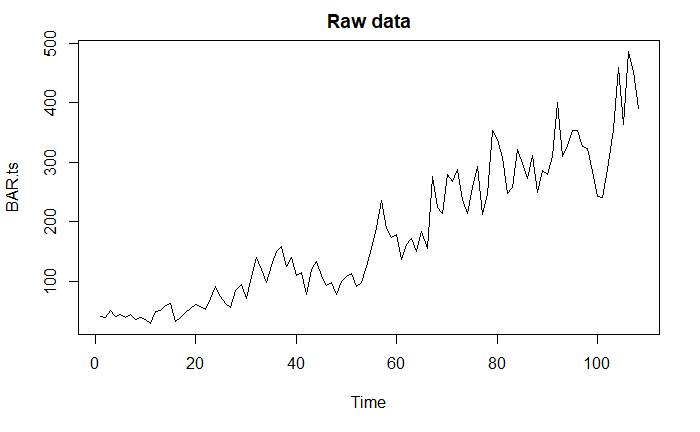
Abstract

I analyzed a dataset containing the amount of armed robberies committed every month from January 1966 to October 1975. I first removed the last 10 datapoints for testing purposes and then analyzed the descriptors of the base dataset. I plotted each year on a seasonality plot but did not find seasonality to be significant. I transformed my data using Box-Cox transformation, finding a lambda of approximately .3. The Box-Cox confidence interval did not include 0 which would suggest a log transformation, or 1 which would suggest that no transformation is necessary. The variance decreased significantly after the transformation. Next, I proceeded to difference the data. Although I concluded the data was not seasonal based off the seasonality plot I tested seasonal differencing to see if it would decrease variance. I found that the differencing that resulted in the lowest variance was differencing at lag 1 once for trend with no differencing for seasonality. I tested for stationarity using the Dickey-Fuller test and found that my differenced data was stationary. After model selection and diagnostics I concluded my final model as ARIMA(0,1,2) . Finally, using this model I predicted the next 10 data points and plotted them against the transformed and original data.

Introduction

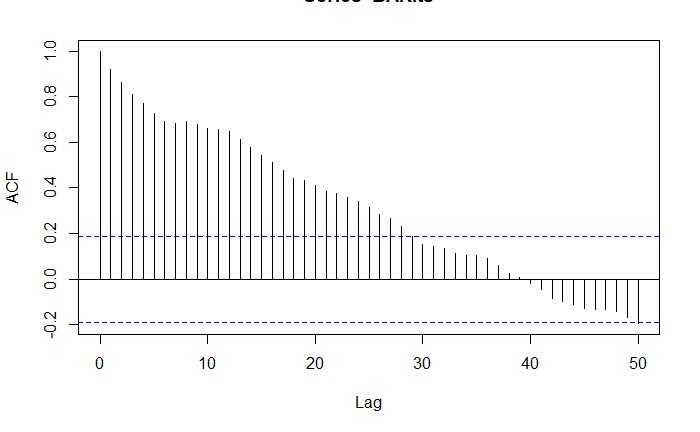
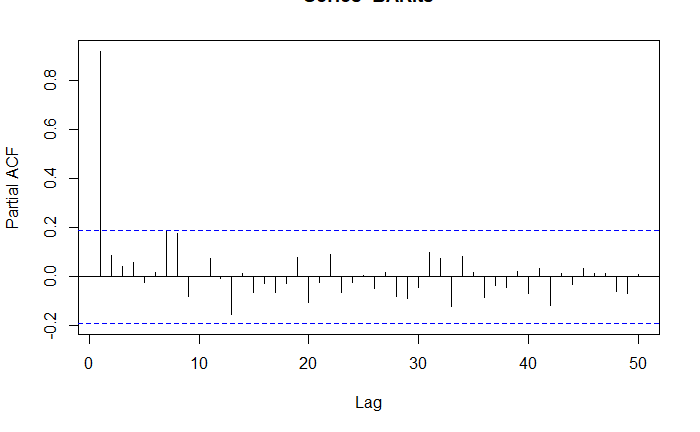
*The Minority Report* by Phillip K. Dick is a short story about a group of detectives that can predict crime. This story was so popular that it was remade into both a movie and television series. However, what if this science fiction story could be brought to real life. I used time series techniques in R to analyze a data set of monthly armed robberies in Boston from 1966 to 1975. This data set is 484 in the tsdl R package. This analysis is important as predicting the number of crimes in a given month will allow citizens and police to prepare in advance. Knowing when a robbery is most likely to occur is helpful in preventing that robbery from succeeding. I plan to forecast the number of armed robberies in Boston committed in a month. Ideally, this method could be expanded upon and used for any crime in any city, fully allowing police and people to prepare for danger. After transforming my data, I found that an ARIMA(0,1,2) model was the best model for predicting Boston armed robberies. The lack of an AR component means that this model quickly reverts to the mean and is only able to predict two months into the future before repeating the last prediction.

Data Analysis



The time series plot of the raw data shows a very strong positive trend and an increase in variance over time. This indicates the data should be transformed and differenced to remove the increasing variance and trend. The seasonality plot does not show a strong trend depending on the month, indicating there is no seasonal component to the time series.

Original data ACF Original data PACF

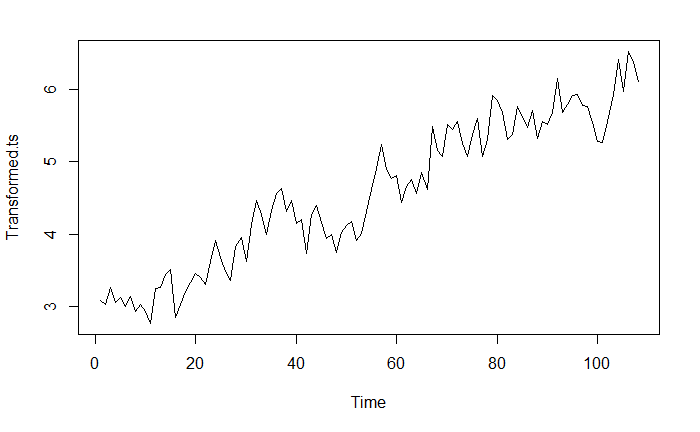
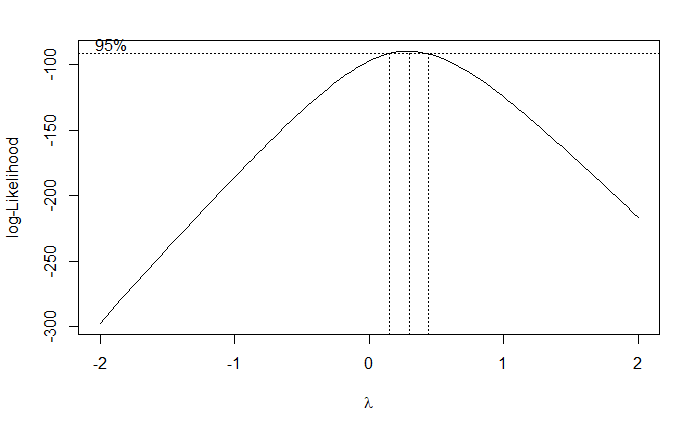


Initially, the original data ACF and PACF indicate the possibility of an AR(1) model. However, this result is likely to change after transformation.

Data Transformation

The initial data does not pass the Dickey-Fuller test for stationarity, so I performed a Box-Cox transformation. The Box-Cox transformation resulted in a lambda value of approximately .3. As seen in below, the Box-Cox confidence interval did not include lambda=1. This suggests that a transformation is necessary. The interval did not contain lambda=0 for a log transformation or lambda = .5 for a square root transformation, so I raised my time series to the power of lambda. The variance decreased significantly after the transformation. The transformed time series still contains a trend, but has lessened the increasing variance.

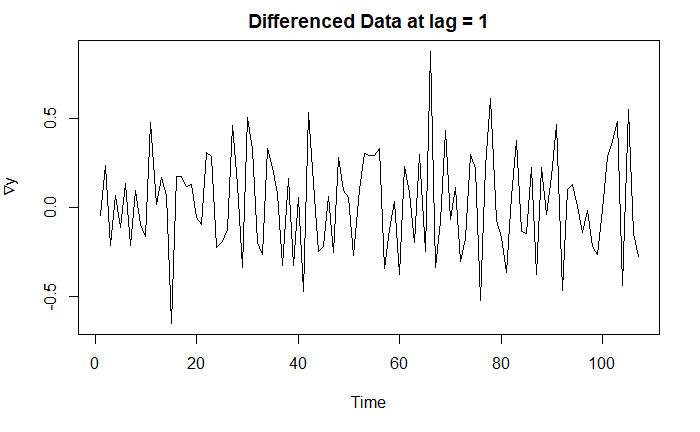
Box-Cox Transformation Confidence Interval Transformed Time Series



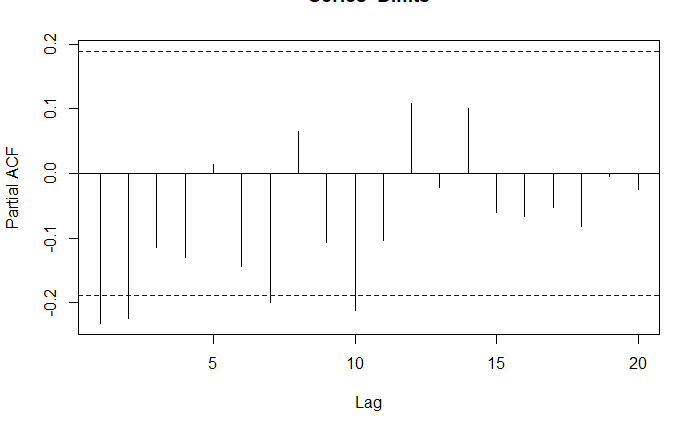
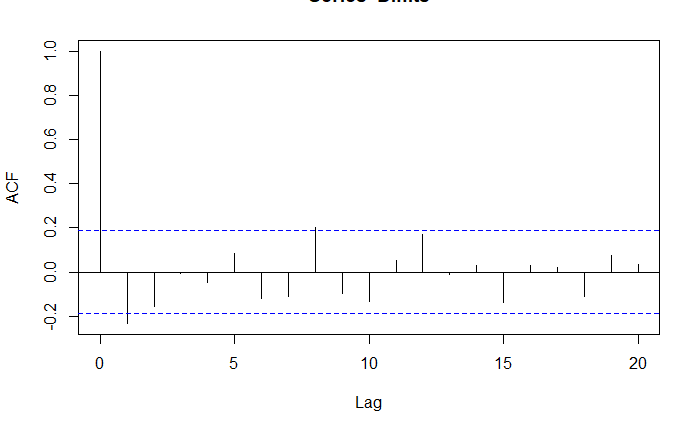
Next, I proceeded to difference the data. Although I concluded the data was not seasonal based off the seasonality plot I tested seasonal differencing to see if it would decrease variance, but it did not. I differenced at lag = 1 to adjust for the significant trend that was apparent in the time series plot. I attempted to difference for trend further but found a second differencing increased variance. I found that the differencing that resulted in the lowest variance was differencing at lag 1 once for trend with no differencing for seasonality. I tested for stationarity using the Dickey-Fuller test and found that my differenced data was stationary. The figure below details the decrease in variance after transformation and differencing.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Original | Transformed | Differenced for trend |
| Variance | 13509.71 | 1.025131 | 0.08085865 |

The transformed and differenced data appears to be stationary. There is no apparent trend and variance does not seem to increase.



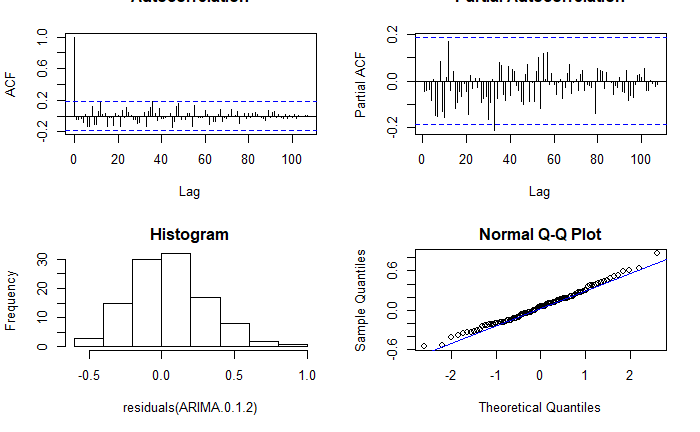
ACF of Differenced Data PACF of Differenced Data



We can now analyze the ACF and PACF of the differenced data. The ACF plot cuts off at lag 8, indicating a possible MA term of 8. The PACF plot cuts off at lag 10 indicating a possible AR term of 10. I entered both MA(8) and AR(10) into consideration and ran a matrix or ARMA models from p=0 to 5 and q=0 to 5 to determine the best model based on AICc. The matrix shows that ARMA(1,1) has the lowest AICc. However, all these models are based off the differenced data. So, we must formulate models using these terms based of the transformed data without differencing. In doing this we also find coefficients for ARIMA(0,1,8), ARIMA(10,1,0), and ARIMA(1,1,1) that maximize likelihood. In addition to these models I run auto.arima to choose a model based off the original data, selecting ARIMA(0,1,2). I then find the coefficients that maximize likelihood for this model and proceed to diagnostics.

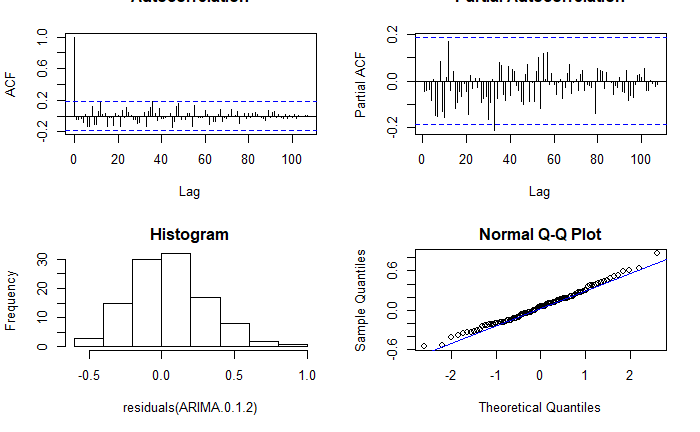
Diagnostics

In order to check for normality, I plotted the histogram and Normal Q-Q plot for the residuals of each model. Every model’s residuals had a histogram that approximately followed the normal curve and a Q-Q plot that approximately followed the Q-Q line, indicating that the data is normal. Finally, I ran each model through the Shapiro-Wilk normality test and tested against the alternative hypothesis that the residuals are not normal. Every model had a P-Value greater than .05, indicating that we fail to reject normality. Below is an example of the ARIMA(0,1,2) model appearing normal in the histogram and Q-Q plot.

ARIMA(0,1,2) Residuals Histogram ARIMA(0,1,2) Residuals Q-Q plot 

We preform the Box-Pierce and Box-Ljung, tests to see if the residuals resemble white noise. ARIMA(0,1,2), ARIMA(1,1,1), and ARIMA(0,1,8) pass the tests indicating that they are not correlated. From here we analyze for the squares of the residuals. We run the McLeod-Li test and find that all remaining models squared residuals are uncorrelated. Additionally, we can analyze the ACF and PACF plots of each models’ residuals. If all the values are within the confidence interval, the model passes for white noise.

ARIMA(0,1,2) Residuals ACF ARIMA(0,1,2) Residuals PACF



The results for the remaining 3 models are summed up below. Each model has a P Value greater than .05, so they pass the diagnostic test.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Box-Ljung test | Box-Pierce test | McLeod-Li test | Shapiro-Wilk normality test |
| ARIMA(0,1,2) | P = 0.3099 | P = 0.3774 | P = 0.9115 | P = 0.7429 |
| ARIMA(0,1,8) | P = 0.1061 | P = 0.1298 | P = 0.7712 | P = 0.2135 |
| ARIMA(1,1,1) | P = 0.1865 | P = 0.2415 | P = 0.9165 | P = 0.7328 |

ARIMA(0,1,2), ARIMA(1,1,1), and ARIMA(0,1,8) passed all of the diagnostic tests, so we will evaluate which model is best using AIC, AICc, and BIC.

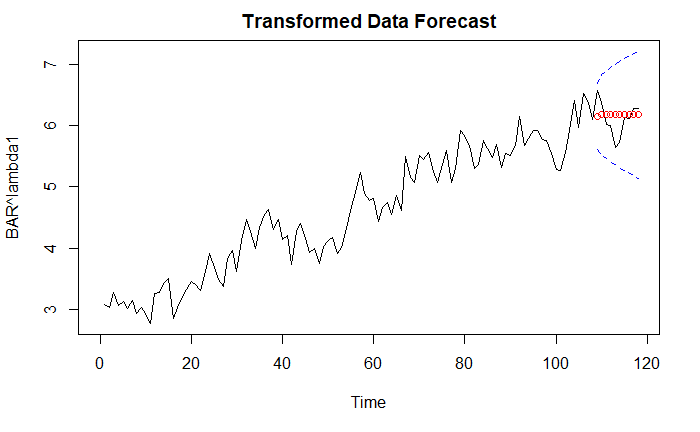
|  |  |  |  |
| --- | --- | --- | --- |
|  | AIC | AICc | BIC |
| ARIMA(0,1,2) | 29.17 | 29.4 | 37.19 |
| ARIMA(0,1,8) | 35.32 | 37.1 | 59.38 |
| ARIMA(1,1,1) | 29.47 | 29.71 | 37.49 |

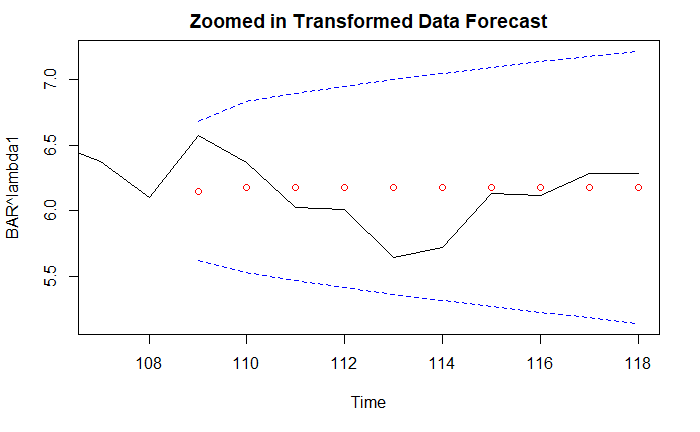
ARIMA(0,1,2) has the lowest AIC, AICc, and BIC so it will serve as the final model.

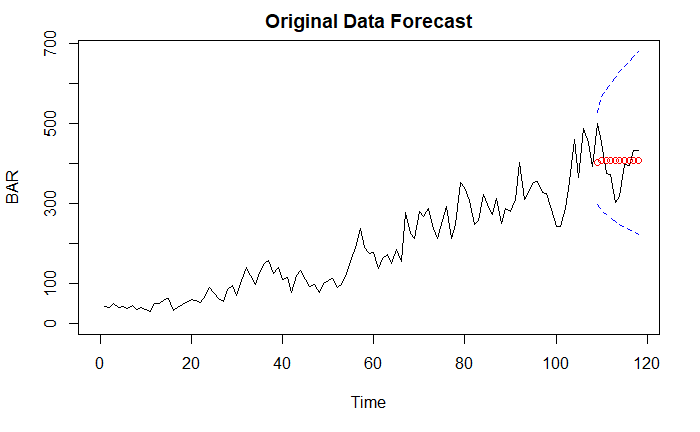
Final Model

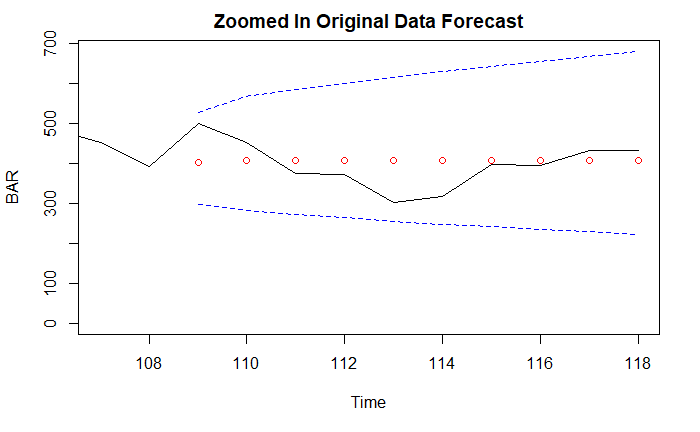
This model is not the same as the ones suggested by ACF or PACF, but is passes all the diagnostic tests and is the most accurate according to a multitude of methods so I deem it to be satisfactory.

Forecasting

The black line is the number of armed robberies in Boston each month from 1966 to 1975. The red dots are my final model’s predictions for the number of armed robberies in Boston for the ten months after the training set. The blue dashed line is my model’s 95% confidence interval for number of armed robberies in Boston.







Conclusion

I set out to predict the number of armed robberies committed in Boston in a given month. I was successful in creating a model that can predict this, but it has limitations. Due to the nature of the data, the model lacks an AR term. This means that we only have a moving average component and the prediction quickly reverts to the mean. However, the prediction values appear to be accurate compared to the actual values. Additionally, all actual values fall within the 95% confidence interval for my model. It is difficult to predict crime as there are many hidden factors that are not readily apparent, so I am happy with my model’s performance.