**Module 5 Project  
  
INTERMEDIATE ANALYTICS  
(TIME SERIES ANALYSIS WITH R)  
  
 NORTHEASTERN UNIVERSITY  
College of Professional Studies**

**SUBMITTED TO: PROFESSOR ROY WADA**

**SUBMITTED BY: REETIKA CHATURVEDI**

**[NEU ID-001081374]**

**INTRODUCTION**

This assignment gave me an opportunity to conduct time series analysis on a dataset. I selected an inbuilt dataset in R which is Australian population data. This dataset deals with time series that contains data of different time periods in terms of years. I tried to understand the features that were there in the dataset, also the changes that were predicted and can be improved.

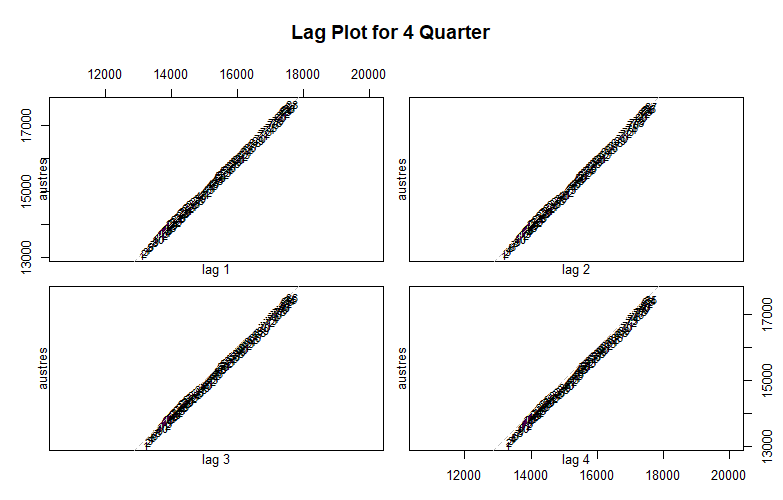
Techniques of time series were used to forecast the population. This data has yearly population values that are divided into four quarters. The data was analyzed and the stationarity and cycles were decomposed to use ARIMA (auto regressive integrated moving average) function. This function enabled to obtain meaningful characteristics and predict the future population in Australia.

**ANALYSIS**

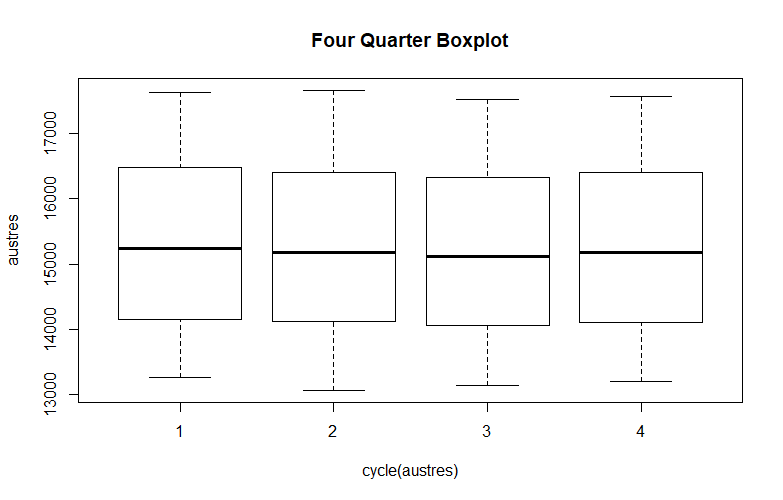
In order to do a time series analysis on my data, I installed the following packages- factoextra, ggplot2, ggcorrplot, dplyr, tseries, astsa and forecast. I used the dataset which is there in the R library- ‘austres’. The structure of the dataset consists of the population rate of Australia from 1971 to 1993 given on a quarterly basis. The summary statistics of the data is given below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Quartile | Median | Mean | 3rd Quartile | Max |
| 13067 | 14110 | 15184 | 15273 | 16399 | 17662 |

After this, I tried to find the regression values on different interval lags. The following is the graph that represents the same.



We can analyze that there is a linear proportionality of the data to the regression line.

After analyzing the linearity of the data, I tried to visualize the structure of data using box plots. I wanted to see if there were outliers in the data.

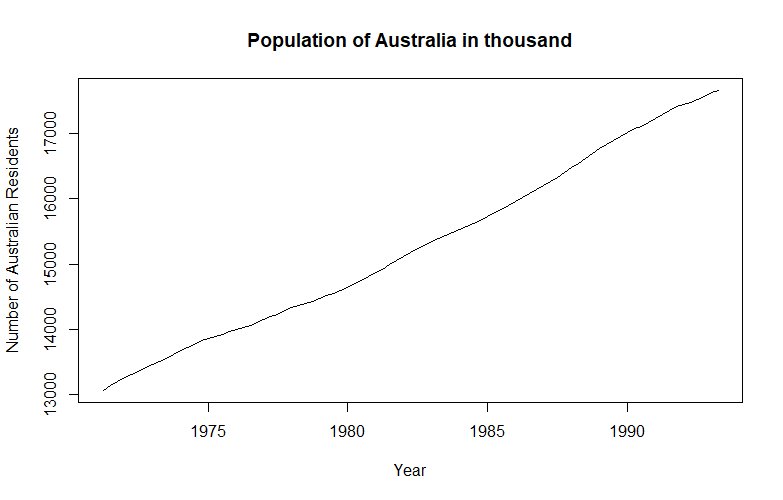
The upper quartiles and lower quartiles could be analyzed. However, there are no major outliers in the data.

After evaluating the box plots, I tried to run the model using prediction function. Sampling was performed to obtain a new dataset which was created in ctree. The prediction has two numbers 1 and 0. 1 represents the conditions that are satisfied whereas 0 represents the conditions that are false. The best combination will help in finding the best predicted values. It also helps us to identify the efficiency that our model possesses.

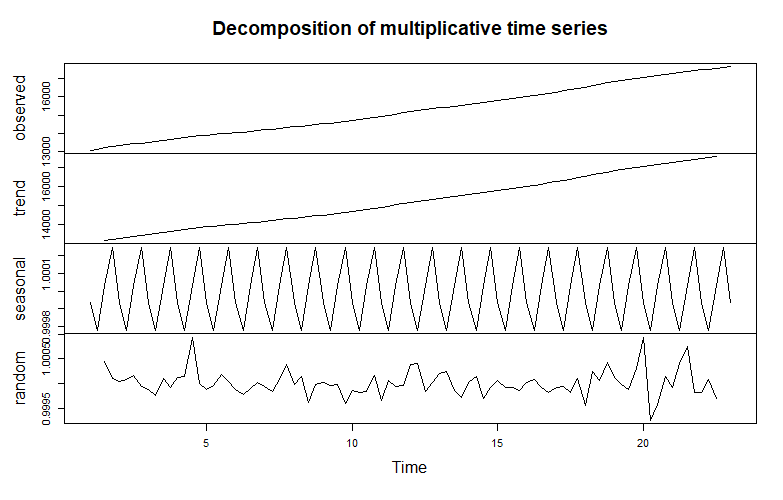
US\_pred <- predict(US.ctree, newdata = US\_test)  
table(US\_pred, US\_test$Rape)

##   
## US\_pred 8.3 9.5 15.6 16.4 18 20 21 22.5 40.6 46  
## 3.17 0 1 0 0 0 0 0 0 0 0  
## 5.6 0 0 0 1 1 1 1 0 0 0  
## 11.7052631578947 1 0 1 0 0 0 0 1 1 1

I tried to visualize the stationarity of the data. The graph that depicts the following is as follows:



It can be seen that the Australian population has increased over the years. The trend is rising and the residents are increasing over the years.

After this, I tried to decompose the time series. This is done in order to combine the cycle and the trend representing a trend-cycle element. This not only helps to understand the time series but also helps to forecast accurately so can be said as one of the important steps.

Once I saw the trend-cycle component, I constructed the Arima model which works on the moving averages and since the data has different time involved, this time series forecasting is required for the Australian population. The best model can be found in this way. The values of Arima (with AIC) will tell us the accuracy of our model. If the AIC values will be low, they are considered to be the best in the model.

The ARIMA model, coefficients and sigma estimation with AIC values is given below:

#Building Arima Model  
model\_austres <- auto.arima(austres)  
model\_austres

## Series: austres   
## ARIMA(0,2,1)(1,0,0)[4]   
##   
## Coefficients:  
## ma1 sar1  
## -0.6051 0.1921  
## s.e. 0.0974 0.1075  
##   
## sigma^2 estimated as 103.7: log likelihood=-322.93  
## AIC=651.86 AICc=652.15 BIC=659.26

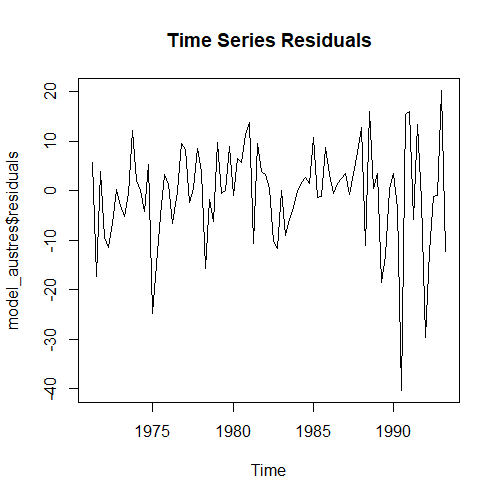
After this auto Arima function was used that entails all the combinations of the model which are possible with the least AIC values.

auto.arima(austres, ic= "aic", trace = TRUE)

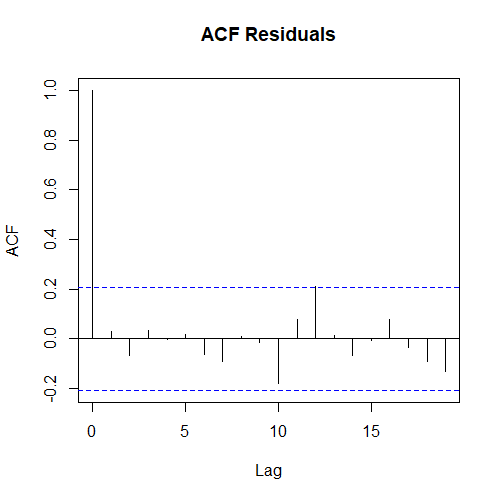
##   
## ARIMA(2,2,2)(1,0,1)[4] : Inf  
## ARIMA(0,2,0) : 672.4575  
## ARIMA(1,2,0)(1,0,0)[4] : 661.8007  
## ARIMA(0,2,1)(0,0,1)[4] : 652.086  
## ARIMA(0,2,1) : 652.9911  
## ARIMA(0,2,1)(1,0,1)[4] : Inf  
## ARIMA(0,2,1)(0,0,2)[4] : 654.0706  
## ARIMA(0,2,1)(1,0,0)[4] : 651.8645  
## ARIMA(0,2,1)(2,0,0)[4] : 653.493  
## ARIMA(0,2,1)(2,0,1)[4] : Inf  
## ARIMA(0,2,0)(1,0,0)[4] : 671.992  
## ARIMA(1,2,1)(1,0,0)[4] : 652.956  
## ARIMA(0,2,2)(1,0,0)[4] : 652.7974  
## ARIMA(1,2,2)(1,0,0)[4] : 654.0543  
##   
## Best model: ARIMA(0,2,1)(1,0,0)[4]

## Series: austres   
## ARIMA(0,2,1)(1,0,0)[4]   
##   
## Coefficients:  
## ma1 sar1  
## -0.6051 0.1921  
## s.e. 0.0974 0.1075  
##   
## sigma^2 estimated as 103.7: log likelihood=-322.93  
## AIC=651.86 AICc=652.15 BIC=659.26

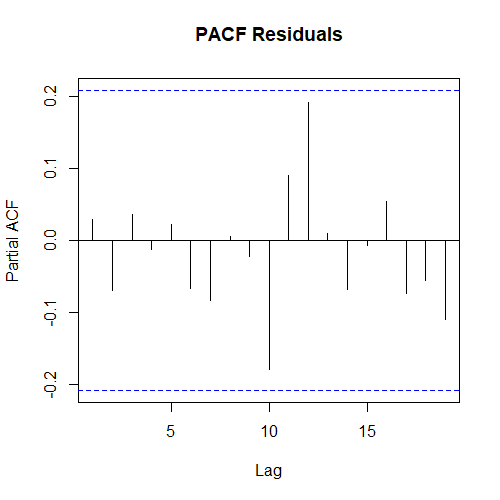
The residuals from all the years were plotted and the graph below represents the same:



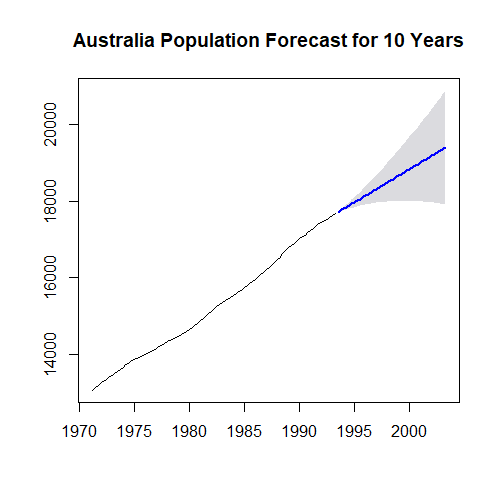
After visualizing the residuals, auto correlation amongst the residual values was computed. The graph below represents that there is a positive and strong correlation in the lower lags whereas there is a weak correlation in the upper lags.



Partial auto correlation function was used to find out the partial correlation of the time series as well as the lagged values. It regresses the shorter lags. It can be seen from the graph below.



After this, there was a forecasting of the Australian population. This was done using the best values predicted by our Arima model. The blue line represents the predicted values (the R file contains all the values).



Below are few predicted values.

## Forecasts:  
## Point Forecast Lo 95 Hi 95  
## 1993 Q3 17704.82 17684.86 17724.78  
## 1993 Q4 17748.00 17713.74 17782.26  
## 1994 Q1 17794.20 17744.70 17843.70  
## 1994 Q2 17835.79 17769.82 17901.76  
## 1994 Q3 17879.09 17792.99 17965.19  
## 1994 Q4 17922.36 17814.65 18030.08  
## 1995 Q1 17966.22 17835.46 18096.97

Ljung- Box testing function was used in the last to find if there are correlations that are different from zero. It is used to test the overall randomness in place of checking different lags. It can be found that our values are above the significance level which can point out towards the correctness of our model. This is a perfect model.

#validate model  
Box.test(model\_austres$residuals, lag = 3, type = "Ljung-Box")

##   
## Box-Ljung test  
##   
## data: model\_austres$residuals  
## X-squared = 0.61819, df = 3, p-value = 0.8923

Box.test(model\_austres$residuals, lag = 8, type = "Ljung-Box")

##   
## Box-Ljung test  
##   
## data: model\_austres$residuals  
## X-squared = 1.8464, df = 8, p-value = 0.9854

Box.test(model\_austres$residuals, lag = 12, type = "Ljung-Box")

**CONCLUSION**

After performing time series analysis, I conclude that time series help to find the hidden structures in the given observations. After carefully observing the time series model, I was able to predict and monitor the data. I was able to understand the structure and the main deterministic features in the observed data thereby was able to choose a suitable model that helped me to analyze. So, it can be said that time series help in decision making.

**REFERENCES**

Hayes, A. (2020, February 5). Trend Analysis. Retrieved from <https://www.investopedia.com/terms/t/trendanalysis.asp>

Fuqua School of Business. (n.d.). Retrieved from https://people.duke.edu/~rnau/411arim.htm