**TIME SERIES ANALYSIS -2**

**REPORT**

Shamini Puthooppallil Baby

[Shaminibaby000@gmail.com](mailto:Shaminibaby000@gmail.com)

Masters In Data Science

Table of Contents

[Introduction 4](#_Toc8552310)

[Data Preparation 5](#_Toc8552311)

[Time Series Analysis 6](#_Toc8552312)

[Model Building -Regression Approach 7](#_Toc8552313)

[Models- Stationary Time Series 8](#_Toc8552314)

[Linear Model 8](#_Toc8552315)

[Quadratic Trend 9](#_Toc8552316)

[Cyclical or Seasonal trend. 10](#_Toc8552317)

[Models - Non-stationary Time Series 11](#_Toc8552318)

[ACF and PACF plots 11](#_Toc8552319)

[ADF Test for Non-Stationarity. 11](#_Toc8552320)

[Transformation 13](#_Toc8552321)

[Differencing 14](#_Toc8552322)

[Order of AR/ MA ( p and q) 16](#_Toc8552323)

[Parameter Estimation 18](#_Toc8552324)

[Model Diagnostics 23](#_Toc8552325)

[Analysis of residuals 23](#_Toc8552326)

[Overfitting and Parameter Redundancy 26](#_Toc8552327)

[Prediction 28](#_Toc8552328)

[Conclusion 29](#_Toc8552329)

[References 30](#_Toc8552330)

[Appendix 31](#_Toc8552331)

[1. Dataset 31](#_Toc8552332)

[2. Required Libraries 31](#_Toc8552333)

[3. Pre validations 31](#_Toc8552334)

[4. Time series plot of R time series object 32](#_Toc8552335)

[5. Plot - relationship between pairs of consecutive egg deposition: 32](#_Toc8552336)

[6. Linear Trend 32](#_Toc8552337)

[7. Quadratic Trend 33](#_Toc8552338)

[8. ACF and PACF plots 33](#_Toc8552339)

[9. ADF Test to analyses non stationarity 33](#_Toc8552340)

[10. Ar() and adfTest to check normality. 34](#_Toc8552341)

[11. Box cox transformation, CI level and ADF for the raw data and transformed data. 34](#_Toc8552342)

[12. Differencing 35](#_Toc8552343)

[13. ACF and PACF plot of egg deposition data with d=4 38](#_Toc8552344)

[14. Extended auto correlation Function 38](#_Toc8552345)

[15. Bayesian Information Criterion (BIC) 38](#_Toc8552346)

[16. Parameter Estimation (Maximum Likelihood) 38](#_Toc8552347)

[17. Parameter Estimation (Least squares) 38](#_Toc8552348)

[18. Sort AIC and BIC for all candidate models 39](#_Toc8552349)

[19. Sort.score-user defined function 39](#_Toc8552350)

[20. Function- ‘residual.analysis’ 39](#_Toc8552351)

[21. Overfitting models 40](#_Toc8552352)

# Introduction

This report describes time series analysis of Egg deposition of age-3 Lake Huron Bloaters ([Coregonus hoyi](https://en.wikipedia.org/wiki/Coregonus_hoyi)) by year,1981-1996 in millions).This is a data frame of 16 observations on the following 2 variables:

* year -Year of data (1981-1996)
* eggs Millions of eggs deposited

The dataset is attached to [Appendix 1](#_Dataset)

The main aim is to analyze the data by using the Time series analysis methods. The analysis trying to find the best fitting trend model  among a set of possible models for this dataset to and give predictions of egg depositions for the next 5 years.

The analysis starts with a data cleansing part followed by model specification by descriptive analysis. Various models will be testing using statistical methods and out of all the fitted models, a suitable model will be selected for future prediction values.

## 

# Data Preparation

Before starting the data analysis and modelling, the main duty of the data scientist is to make sure that the data provided is in correct format. If the dataset is not in proper format, the entire work needs to be repeated.

The required libraries for the time series analysis is in [Appendix 2](#_Required_Libraries_1).

Dataset is loaded to the R software using read.csv function and performed required pre-validations for the loaded dataset. The time series analysis is performed using the R Markdown and the common statistical tools in the upcoming sessions.

Structure of the dataset, column values, null values and impossible values are checked using some basic R data preprocessing packages. Sample data is viewed and make sure that the data loaded correctly to the R software. This code snippet is available in [Appendix 3](#_Pre_validations)

# Time Series Analysis

In this stage, the dataset is analyzed using more statistical and descriptive tools to get more insights of the time series characteristics.

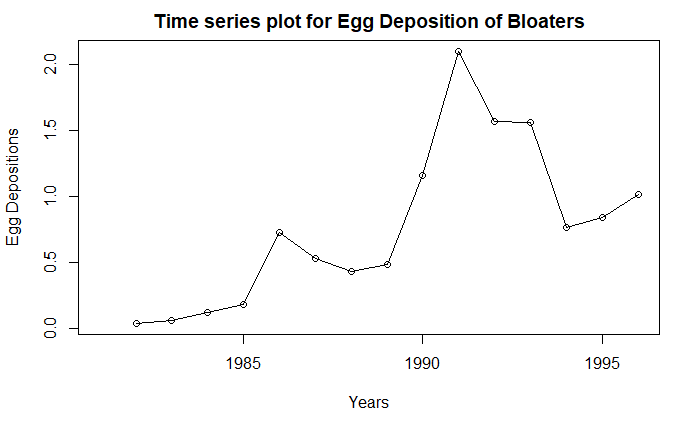
The data analysis stage considered the time series characteristics of the dataset and figured out the elements of suitable and successful data analysis of time series data. The dataset needs to be a R time series object to perform the time series analysis. So, the dataset converted to timeseries object using the TSA package function. Before converting to the time series object, egg deposition column is extracted from the dataset by eliminating the year column.

The ts() function will convert the data frame into an R time series object. Next step is to plot the time series dataset and analyses the common characteristics of time series plot.

You can find the code snippet in [Appendix 4](#_Time_series_plot)

A time series plot was drawn initially to analyze the common characteristics of the dataset.

Figure 1: Time series plot for Egg Deposition of Bloaters



We observe considerable variation in egg deposition of Bloaters over the years from this time series plot. The year 1990 reported the highest egg deposition, while 1981 was the lowest. We can consider 1990 as an intervention point. From 1990 onwards the trend of the series decreasing up to 1993 and from 1994 onwards there is a slight increase in the egg deposition values. From this 16-year period, we cannot spot any obvious changing variance according to the time series plot. Also, the data set is not showing any seasonality during this year period .The plot shows an almost an upward trend up to the year 1990 and from 1991 onwards, the plot is falling downwards by depicting a downward trend. Last ,3 years shows a slightly increasing behavior after this big fall.

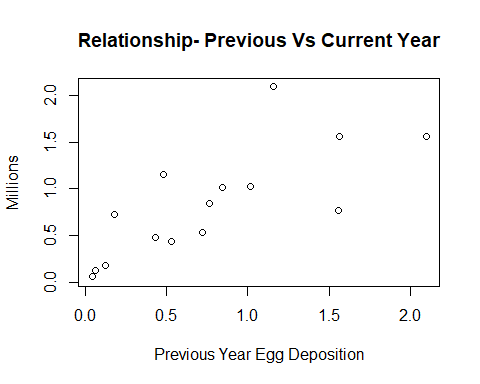
The above are the assumptions we made from the first glance of the time series plot. We will analyze the above points further in the subsequent sessions.

To investigate more on the relationship between pairs of consecutive egg deposition, a scatter plot is drawn with current year and previous year data in the egg deposition of Bloater’s dataset.

The code chunk for this relationship is available in [Appendix 5](#_Plot_-_relationship).

A scatter plot of the adjacent years plotted using zlag function.

Figure 2: Scatterplot represents the relationship between consecutive years.



This plot shows a slight upward trend. We observe correlation between egg deposition of succeeding years However, it is impossible to observe seasonality for this scatter plot. Here, neighboring values are very closely related up to half of the plot. In the second half of the plot, large changes in egg deposition occur from one year to the next. Correlation between the previous year observation and the current year observation has the value 0.7445657 which is a strong one.

# Model Building -Regression Approach

In this phase, we are trying to fit our dataset to the best suitable stationary model. The main objective of this process is to find a suitable model which is appropriate and satisfy most of the statistically significant tests and assumptions.

## Models- Stationary Time Series

Time series plot of the egg deposition of bloaters shows different trends during the years. This is an indication of non-stationarity. So, initial analysis concluded that linear and quadratic models are not suitable for this time series dataset. The below sections show the summary statistics to provide relevant information to reject these models.

### Linear Model

Initially the data is fitted to a linear trend model and analyzed the characteristics and compared these characteristics with the rest of the models.

## Call:  
## lm(formula = egg\_Bloaters ~ time(egg\_Bloaters))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.4048 -0.2768 -0.1933 0.2536 1.1857   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -165.98275 49.58836 -3.347 0.00479 \*\*  
## time(egg\_Bloaters) 0.08387 0.02494 3.363 0.00464 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4598 on 14 degrees of freedom  
## Multiple R-squared: 0.4469, Adjusted R-squared: 0.4074   
## F-statistic: 11.31 on 1 and 14 DF, p-value: 0.004642

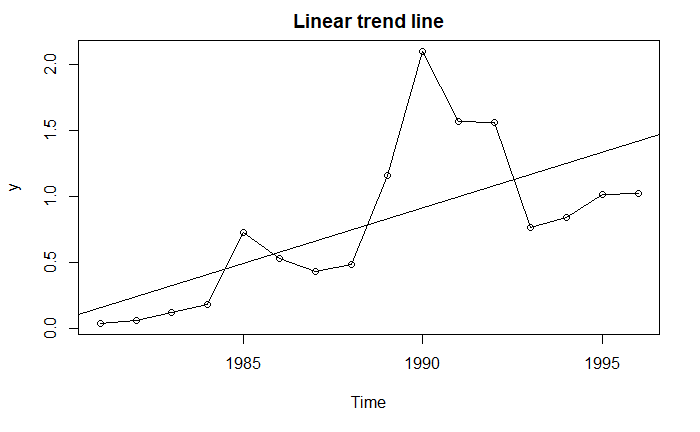
According to multiple R^2, only 40% of the variation in the linear model series is explained by the linear time trend. Linear model is not suitable for egg deposition dataset.

After fitting the model, a trend line is drawn along with the original dataset to analyze the behavior of the linear model.

From Figure 3, it is observed that the trend line touching the mean of the values. And most of the data points couldn’t touch by the linear trend model and can try some better models instead of linear.

The R codes used for linear model is available in [Appendix 6](#_Linear_Trend_1)

Figure 3: Fitted Linear Trend Line



### Quadratic Trend

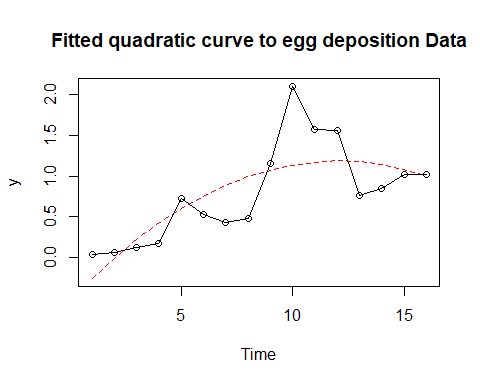
Next, the data is fitted to a quadratic trend model and analyzed the characteristics and compared these characteristics with the rest of the models.

## Call:  
## lm(formula = egg\_Bloaters ~ t + t2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.50896 -0.25523 -0.02701 0.16615 0.96322   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.647e+04 2.141e+04 -2.170 0.0491 \*  
## t 4.665e+01 2.153e+01 2.166 0.0494 \*  
## t2 -1.171e-02 5.415e-03 -2.163 0.0498 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4092 on 13 degrees of freedom  
## Multiple R-squared: 0.5932, Adjusted R-squared: 0.5306   
## F-statistic: 9.479 on 2 and 13 DF, p-value: 0.00289

According to multiple R^2, only 50% of the variation in the quadratic model series is explained by the quadratic time trend. We could conclude that quadratic model is not suitable for egg deposition dataset.

After fitting the model, a trend curve is drawn along with the original dataset to analyze the behavior of the quadratic model.

Figure 4: Fitted Quadratic Trend curve



We can observe that the trend line touching the mean of the values. And most of the data points couldn’t touch by the quadratic trend model and can try some better models instead of quadratic.

In this quadratic model, it’s clear that the fitted curve cannot captures the mean level throughout the time period. So, this model may not fit in this dataset.

The R codes used for quadratic model is available in [Appendix 7](#_Quadratic_Trend)

### Cyclical or Seasonal trend.

The time series plot does not show any sign of seasonality or cyclic behavior. I’m not proceeding with the cyclic trend analysis for this dataset

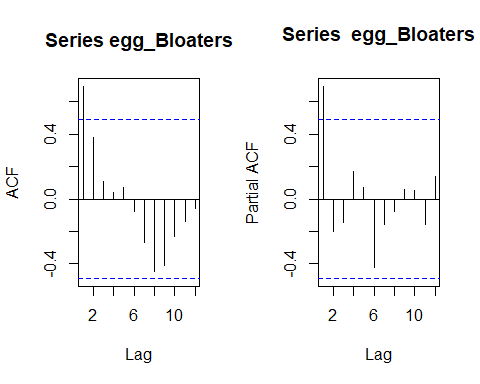
## Models - Non-stationary Time Series

From the stationary time series analysis and model building approach we got an indication that this series depicts non stationarity. The following session gives more thorough analysis of non-stationarity followed by time series model building approaches.

### ACF and PACF plots

In time series analysis, trend is apparent from ACF AND PACF plots.

Figure 5: ACF and PACF plots for Bloaters egg data.



Slowly decaying pattern in the auto correlation function plot is a high indication of trend in the series. In the ACF plot, we can observe an almost wave like pattern which is repeating upwards and downwards, So, from the initial assumption we can conclude that this series is non stationary and shows trend.

The Partial auto correlation function plot shows a very high first correlation. This is a strong evidence of non-stationarity in the egg deposition dataset.

The R code chunks used for drawing these plots are available in [Appendix 8](#_ACF_and_PACF)

### ADF Test for Non-Stationarity.

The augmented Dickey-Fuller (ADF) test statistic is the t-statistic of the estimated coefficient of a from the method of least squares regression. However, the ADF test statistic is not approximately t-distributed under the null hypothesis; instead, it has a certain nonstandard large-sample distribution under the null hypothesis of a unit root.[3]

ADF test is applied with default settings in the egg deposition data.

## Augmented Dickey-Fuller Test

## data: egg\_Bloaters

## Dickey-Fuller = -2.0669, Lag order = 2, p-value = 0.5469

## alternative hypothesis: stationary

With a p-value of 0.5469, we cannot reject the null hypothesis stating that the series is non-stationary.

The R code chunks used for ADF Test is available in [Appendix 9](#_ADF_Test_to)

From the initial estimates, the raw data contains a trend in the series. In order to get rid of the trend and non-stationarity, from the model initially a box cox transformation is applied to the dataset and compared it with the actual dataset if there is any improvement.

ADF unit root test for checking the non-stationarity is applied to the dataset .the order of the series return from the test was ‘0’ and Augmented Dickey-Fuller Test is applied to the dataset with order 0.

The R codes used for f the ar() function and adfTest is available at [Appendix 10](#_Ar()_and_adfTest)

**Ar()**

## Call:  
## ar(x = diff(egg\_Bloaters))  
##   
##   
## Order selected 0 sigma^2 estimated as 0.1841

**adfTest**

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 0  
## STATISTIC:  
## Dickey-Fuller: -0.4911  
## P VALUE:  
## 0.452   
##   
## Description:  
## Sat May 11 11:32:16 2019 by user: SIA

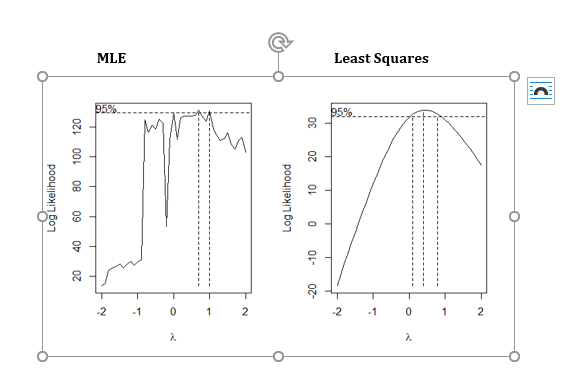
The P value obtained from the adfTest is 0.452 which confirms that this dataset is not stationarity and we need to perform further transformations or differencing for the further analysis

### Transformation

In order to apply transformations in the dataset ,initially the lambda value for the box cox transformation is obtained and from the selected lambda value , egg deposition dataset is transformed.

The default method for fitting the lambda value is MLE . Since, we are not getting a smooth curve from that model, we fitted the dataset with least squares and selected the lambda value as 0.45 , midpoint in between 0.1 and 0.8

Figure 6: Range of Lambda values from MLE AND Least Square methods.



As a second step, adf test is performed on the initial dataset and the transformed dataset to check how the stationarity affected.

Code chunk for box cox transformation and adf test are available at [Appendix 11](#_Box_cox_transformation,)

egg1.transform**$**ci

## [1] 0.7 1.0

egg1.transform**$**ci

## [1] 0.1 0.8

##   
## Augmented Dickey-Fuller Test  
##   
## data: egg\_Bloaters  
## Dickey-Fuller = -2.0669, Lag order = 2, p-value = 0.5469  
## alternative hypothesis: stationary

**adf.test**(BC.egg)

##   
## Augmented Dickey-Fuller Test  
##   
## data: BC.egg  
## Dickey-Fuller = -1.6769, Lag order = 2, p-value = 0.6955  
## alternative hypothesis: stationary

From the adf test we realized that the box cox transformation cannot improve the stationarity of the series, also from the p value we can conclude that the probability of non-stationarity increased ( p value from 0.5 to 0.6) this means that box cox transformation didn’t help the dataset much.

So here we are the CI level from MLE is selected (0.7 1.0) Since, 1 is in the range. The dataset is preserved as the original dataset without any transformation.

### Differencing

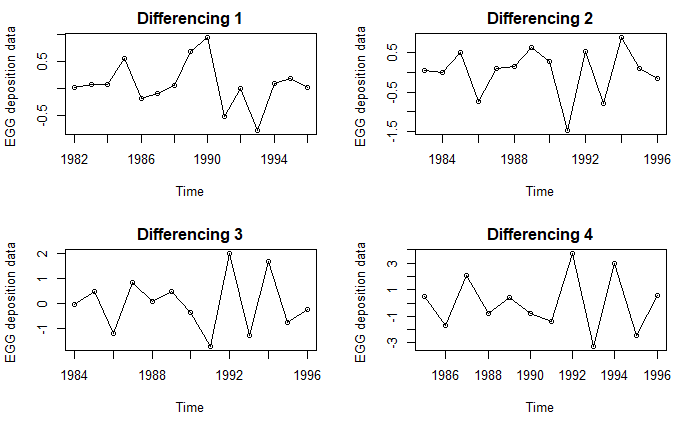
The statistical tests performed above shows that the egg deposition dataset shows trend and non - stationarity. To make the series non stationary and out of trend, different differencing is applied in order and after each differencing, applied the ADF unit-root test to test the existence of non-stationarity. Probably value from every differencing is compared and selected the p value with 5% of significance and rejected all other differenced series.

[Figure 7](#_Differencing) shows the time series plot for differenced datasets. From the plots, we can observe that first difference and second difference show a slight trend but third and fourth could eliminate the trend and the series looks almost non stationary.

The code chink for differencing and the corresponding ADF test is available at [appendix 12.](#_Differencing_1)

After the time series plot analysis ,adfTest performed on each differenced dataset to check the significant level of non-stationarity.

Figure 7: Time series plot of (d=1,2,3,4).



In fourth difference p value is less than 0.05 (0.03368) this means that this series is stationary, and we can proceed with the modelling from this point.

Value obtained from the fourth difference is shown below.

## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 2  
## STATISTIC:  
## Dickey-Fuller: -2.1524  
## P VALUE:  
## 0.03368   
##   
## Description:  
## Sat May 11 12:55:36 2019 by user: SIA

From the Augmented Dickey-Fuller Test its better to proceed with this dataset with a d value 4.Before finalizing the differenced dataset a shapiro test is performed and checked the normality of the series.

Result of the normality test shows that a higher p value(~.8) this means that the dataset is normally distributed, and we can proceed with our assumption.

##   
## Shapiro-Wilk normality test  
##   
## data: diff.egg  
## W = 0.96712, p-value = 0.8785

The following session contains the methods find out the order of AR and MA models in the fourth difference of the egg deposition data.

### Order of AR/ MA ( p and q)

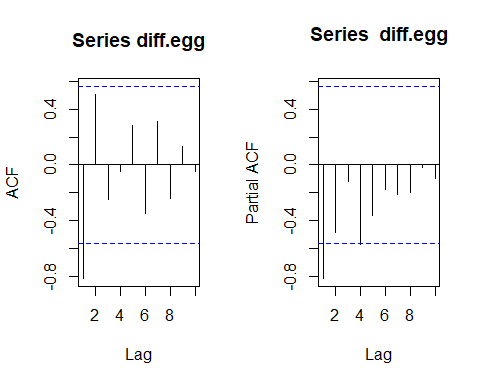
Once the time series data is fixed with a difference 4, next step is to find the order of the Auto regressive and Moving average models. To find these orders the following methods are adopted from the statistical visualization tools and pre-defined functions.

1. **ACF / PACF plots**

ACF and PACF plots will give a clear indication of trend and seasonality in the series. From the probability value of .03 we cannot conclude that the series is non stationary and has no trend,

Figure 8 shows ACF and PACF plots for the egg deposition data with a difference of 4.

Figure 8: ACF and PACF plot for egg deposition data with difference(d=4)



ACF plot shows one significant line (negative) and all the other lines are inside the dashed line. Also, this plot is not showing any repeating pattern or any indication of seasonality, trend and non-stationarity. From this observation we can select the order of MA to be 1(q)

PACF plot also shows a negative significant line in the first phase of the plot and one line is slightly touching the dashed line. It’s our choice to include this or not. From my analysis I could lake the order of the 1 or 2 (p).

From the above plots, the candidate models are ARIMA(1,4,1) and ARIMA(2,4,1)

R code chunk used for creating these plots are available at [Appendix 13.](#_ACF_and_PACF_1)

1. **Extended Auto Correlation Function (EACF)**

The eacf function allows maximum of 3 for AR model and a maximum of 2 for MA model.

Here is the EACF table obtained from the function eacf()

AR/MA

0 1 2

0 x o o

1 o o o

2 o o o

3 o o o

From the eacf table its able to figure out more candidate models and we consider all the possible models from this stage and eliminate them in the model diagnosis and parameter estimation stage.

Possible candidate models eacf table are { ARIMA(0,4,1), ARIMA(0,4,2), ARIMA(1,4,1), ARIMA(1,4,2)}

R code chunk used for EACF IS available at [Appendix 14.](#_Extended_auto_correlation)

1. **Bayesian Information Criterion (BIC)**

It would be helpful to examine a few best subset ARMA models (in terms of, BIC) in order to arrive at some helpful tentative models for further study.

The pattern of which lags of the observed time series and which of the error process enter the various best subset models can be summarized in a display like that shown below.

Figure 9: Various candidate models from BIC

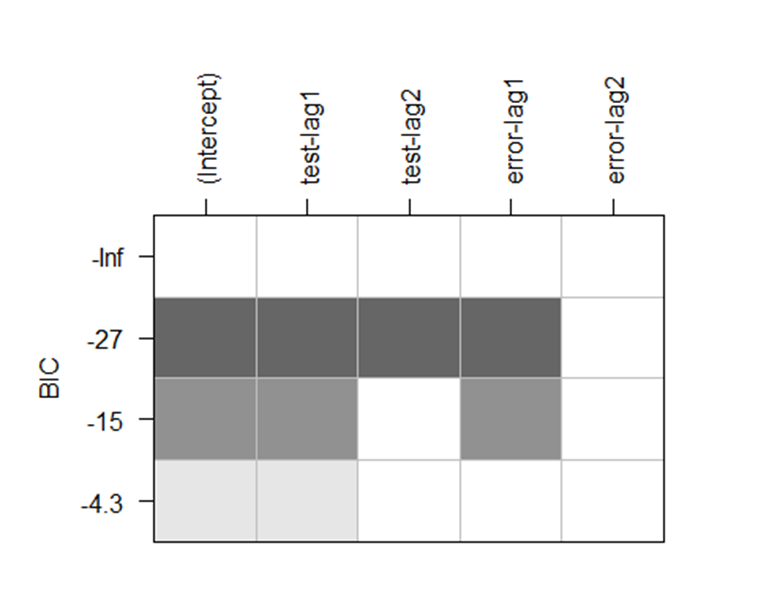


Figure 9 suggests the models AR (1) and MA (1).If we generalize our view, we can add AR(2) model as well which is obtained from initial acf and pacf plots.

The candidate models from the BIC statistics are ARIMA(1,4,1) and ARIMA(2,4,1) and this subset is same as the models we got initially from the ACF and PACF plot.

R code used for the above is available at [Appendix 15](#_Bayesian_Information_Criterion)

From the above 3 order estimation methods, 5 models are finalized as candidate models abd They are :

1. ARIMA(0,4,1)
2. ARIMA(0,4,2)
3. ARIMA(1,4,1)
4. ARIMA(1,4,2)
5. ARIMA(2,4,1)

This report explains more about the parameter selection, model diagnosis and model elimination criteria in the following sessions.

## Parameter Estimation

After ensuring the series is stationary and specification of orders of autoregressive and moving average elements of ARIMA models, the next step is to estimate parameters of the specified tentative models.

The following session gives the estimation of parameters for all the selected models. After the parameter specification, the model with highly significant values will select. Also, based on the AIC and BIC, the model with lower value of these will select for forecasting.

1. **Maximum Likelihood Estimate**

The below results show coefficient test for maximum likelihood estimates.

coeftest(model\_041\_ml)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ma1 -0.97878 0.20781 -4.71 2.477e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(model\_141\_ml)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ar1 -0.63807 0.19490 -3.2739 0.001061 \*\*   
## ma1 -0.97188 0.23748 -4.0924 4.269e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(model\_042\_ml)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ma1 -1.86751 0.32411 -5.7619 8.317e-09 \*\*\*  
## ma2 0.94443 0.31996 2.9517 0.003161 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(model\_142\_ml)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ar1 -0.32699 0.27811 -1.1758 0.23969   
## ma1 -1.82562 0.45095 -4.0484 5.157e-05 \*\*\*  
## ma2 0.87601 0.46015 1.9038 0.05694 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(model\_241\_ml )

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ar1 -0.80414 0.28948 -2.7779 0.0054713 \*\*   
## ar2 -0.21424 0.28760 -0.7449 0.4563070   
## ma1 -0.96672 0.25708 -3.7604 0.0001696 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

From the results obtained its obvious that all the models except ARIMA(1,4,2) and ARIMA(2,4,1) are not significant. In ARIMA(1,4,2)only MA(1) is significant and all the others are not .Also, in ARIMA(2,4,1) AR(2) is not significant.

The R code chunks used for parameter estimation of maximum likelihood method is given in [Appendix 16](#_Parameter_Estimation_(Maximum).

1. **Least squares estimate**

The below results show coefficient test for least square estimates

#significance tests for each parameter coeftest()  
coeftest(model\_041\_css)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ma1 -1.25372 0.10027 -12.504 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(model\_141\_css)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ar1 -0.71015 0.21880 -3.2456 0.001172 \*\*   
## ma1 -0.90595 0.12263 -7.3877 1.494e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(model\_042\_css)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ma1 -2.29095 0.14287 -16.0351 < 2.2e-16 \*\*\*  
## ma2 1.41267 0.16555 8.5331 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(model\_142\_css)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ar1 -0.53170 0.36720 -1.4480 0.147627   
## ma1 -1.30688 0.49188 -2.6569 0.007886 \*\*  
## ma2 0.43860 0.55530 0.7898 0.429623   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(model\_241\_css)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ar1 -1.02602 0.30042 -3.4153 0.0006371 \*\*\*  
## ar2 -0.38402 0.31022 -1.2379 0.2157511   
## ma1 -0.79038 0.17250 -4.5818 4.609e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Least square estimates also give the same result as maximum likelihood. As per the ML the last two models were not significant. Likewise, In ARIMA(1,4,2)only MA(1) is significant and all the others are not .Also, in ARIMA(2,4,1) AR(2) is not significant.

The R code chunks used for parameter estimation of maximum likelihood method is given in [Appendix 17.](#_Parameter_Estimation_(Least)

From the above two methods, the models with all coefficients significant are

1. ARIMA(1,4,1)
2. ARIMA(0,4,1)
3. ARIMA(0,4,2)
4. **AIC & BIC Values**

For now, we will consider AIC and BIC values of the models to decide the best one within the subset of possible models. Here we include all the models we selected for BIC and AIC value calculation. The results obtained from the coeftest() cannot eliminate without finalizing the best model obtaining from these values.

Listing the AIC and BIC values obtained from all the candidate models below.

**AIC**

## df AIC  
## model\_042\_ml 3 42.84286  
## model\_142\_ml 4 43.76585  
## model\_141\_ml 3 44.37815  
## model\_241\_ml 4 45.87398  
## model\_041\_ml 2 49.07553

BIC

## df BIC  
## model\_042\_ml 3 44.29758  
## model\_142\_ml 4 45.70547  
## model\_141\_ml 3 45.83287  
## model\_241\_ml 4 47.81361  
## model\_041\_ml 2 50.04535

# Both AIC and BIC select ARIMA(0,4,2) model for this series

From the sorted AIC values its obvious that ARIMA(0,4,2) has the minimum value (42.84286).Also, in the BIC values same model shows the least estimates(44.29758) .

This code snippet is available in [Appendix 18](#_Sort_AIC_and)

‘Sort.score’ function used in for sorting the model values are attached in [Appendix 19](#_Sort.score-user_defined_function)

Since, both estimates point out the same model, we can fix this model as our final model. But, before proceeding to the forecast, we must analyze some model accuracy by some popular model diagnosis techniques.

# Model Diagnostics

After specification of orders of autoregressive and moving average parts, we estimated parameters of specified ARMA models. This session includes the diagnostic checks to validate the accuracy of the finalized model by some statistical parameters.

Here, we will concern with testing the goodness of fit of the model. If the goodness of fit is poor, we will suggest appropriate modifications

## Analysis of residuals

If the model is correctly specified and the parameter estimates are reasonably close to the true values, then the residuals should have nearly the properties of white noise. They should behave roughly like independent, identically distributed normal variables with zero means and common standard deviations.

For the residual analysis residual plots are considering listing out the plots used for the residual analysis below.

* Time series plot of standardized residuals
* Shapiro Test for normality
* Histogram of standardized residuals
* QQ plot of standardized residuals
* ACF of standardized residuals
* Ljung-Box test for Residuals

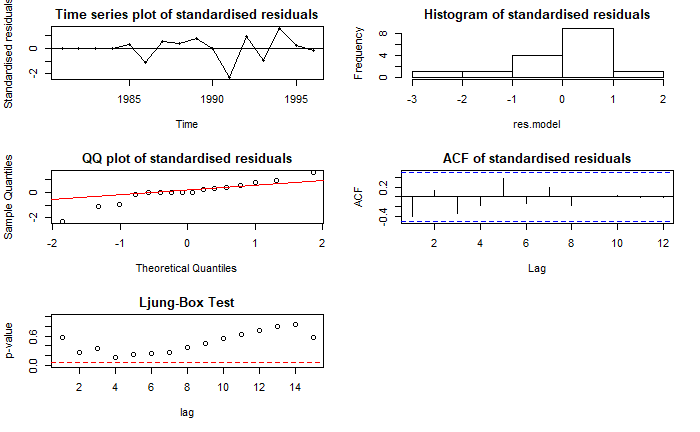
All the above listed methods are included and created a function named ‘residual.analysis’.The function applied to the selected model.

Residual analysis unction declaration attached to [Appendix 20](#_Function-__‘residual.analysis’)

Attaching the result obtained from the residual analysis below.

Deviations from these properties can help us discover a more appropriate model

Figure 10: Residual Analysis Plots



.

* **Time series plot of standardized residuals**

If the model is adequate, we expect the plot to suggest a rectangular scatter around a zero-horizontal level with no trends whatsoever in the time series plot of standardized residual. In this case we can observe that Residuals are around the 0 mean and all the points are inside the range -3 to +3.

* **Shapiro Test for normality**

And the Shapiro-Wilk test for the residual of the model fitted to the egg deposition series is:

##   
## Shapiro-Wilk normality test  
##   
## data: res.model  
## W = 0.91548, p-value = 0.1426  
##

The test statistic for this test is W = 0.91548 corresponding to a p-value of 0.1426, and we would not reject normality based on this test.

* **Histogram of standardized residuals**

Histogram of the residuals shows normal distribution. If we observe deeper, we can see a slight left skewness in the histogram, but the shapiro test shows 5% significance and gives the evidence of normality.

* **QQ plot of standardized residuals**

There is a fluctuation from the normal line from -2 to -1. From -1 to 2 the residuals follow the normal line. There is only one point(initial) which is far away from the normal line and we can conclude that residuals follow normality.

* **ACF of standardized residuals**

To check on the independence of the noise terms in the model, we consider the sample autocorrelation function of the residuals.

There is no significant value in the in the acf plot of residuals. Also, we can observe all the values are near to 0 and only one or two values shows lengthy line. But, all of them are below the significant level. This suggest that there is no autocorrelation between the residual values, and this leads to a good fit.

We conclude that the graph does not show statistically significant evidence of nonzero autocorrelation in the residuals

* **Ljung-Box test for Residuals**

The Ljung-Box test provides an overall test for looking at residual correlations as a whole

## Box-Ljung test  
##   
## data: res.model  
## X-squared = 10.465, df = 6, p-value = 0.1064

Here the p value is significant at 5% level. So, we have no evidence to reject the null hypothesis that the error terms are uncorrelated.

To prove this point, we drawn a sequence plot of the standardized residuals, the sample ACF of the residuals, and p-values for the Ljung-Box test statistic for a whole range of values of from 5 to 15.

From the resulted visualization ,the estimated ARIMA(0,4,2) model seems to be capturing the dependence structure egg deposition time series quite well.

## Overfitting and Parameter Redundancy

We can use overfitting as another tool to detect anomalies in terms of goodness of fit.

After specifying and fitting what we believe to be an adequate model, we fit a slightly more general model; that is, a model “close by” that contains the original model as a special case.

Now, to ensure overfitting, we fit ARIMA(0,4,3) and ARIMA(1,4,2) to this series and compare the results with that of ARIMA(0,4,2).

coeftest(model\_043\_ml)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ma1 -1.945451 0.409099 -4.7555 1.98e-06 \*\*\*  
## ma2 1.040986 0.591817 1.7590 0.07858 .   
## ma3 -0.021509 0.296991 -0.0724 0.94226   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

MA(2) and MA(3) are insignificant according to MLE. When we increase the MA value to 3 both MA(2) and MA(3) became insignificant.

coeftest(model\_142\_ml)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ar1 -0.32699 0.27811 -1.1758 0.23969   
## ma1 -1.82562 0.45095 -4.0484 5.157e-05 \*\*\*  
## ma2 0.87601 0.46015 1.9038 0.05694 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

When we try AR(1) in the fitted model we got the above result. Here it clearly indicates that AR(1) model is insignificant with p value 0.23969.

The R code used for over fitting is available at [Appendix 21](#_Overfitting_models)

From the model diagnostic methods explained above clearly suggests that the model ARIMA(0,4,2) is the best model out of the selected candidate models.

The parameter estimates of the best fitted model is mentioning below:

coeftest(model\_042\_css)

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ma1 -2.29095 0.14287 -16.0351 < 2.2e-16 \*\*\*  
## ma2 1.41267 0.16555 8.5331 < 2.2e-16 \*\*\*

Also, above estimates are statistically significant at 5% of significant level.

# Prediction

After ensuring that the fitted model is suitable for prediction purposes, we use the model to find forecasts. For time series regression models, this task is simply based on the straightforward use of the fitted regression model.

In the forecasts differencing will already been taken back. We need to specify the lambda of Box-Cox transformation if any.

The plot() function can plot the time series data and its predictions with 95% prediction bounds.

We can plot these forecasts next to the time series of interest as shown below.

Figure 11: Forecast model -Egg deposition data

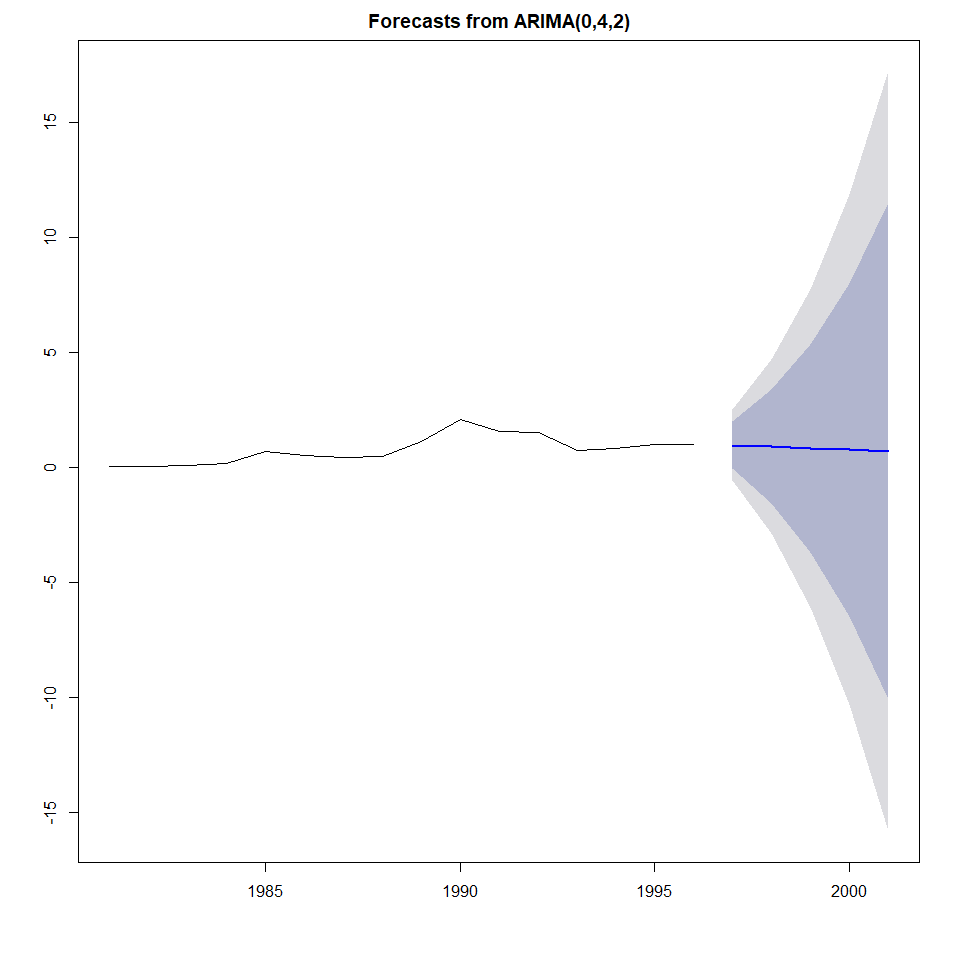


Figure 11 shows forecasts from the ARIMA(0,4,2) model successfully follow the pattern in the original series. Here the overall trend of the series is maintained by the forecast model. The forecast years shows 2 different prediction intervals as well which is in accordance with the original model.

From the overall analysis, it is concluded that the model ARIMA(0,4,2) is the best fit for Egg deposition of age-3 Lake Huron Bloaters dataset .

# Conclusion

The goal of the time series analysis is to find the best fitting trend model to this egg deposition of age-3 Lake Huron Bloaters and give predictions of yearly changes for the next 5 years. The dataset is provided with egg deposition with 16 observations. During the initial analysis, the time series plot shows different trends with no seasonal objects and no interventions.

Model specification and model building performed on different trends such as linear, quadratic. Since the statistical test show the existence of non-stationarity. Model building is done in accordance with the non-stationary time series. All the candidate models are selected during the initial analysis and eliminated one by one by verifying the significance of the model ,model coefficients and parameter estimations.

Residual analysis is performed for the fitted model. The statistical tests and significance tests show that the ARIMA(0,4,2) MODEL is more suitable for egg deposition dataset. So, ARIMA(0,4,2) model fitted for the dataset.

From the fitted model, the egg deposition for the next 5 years is predicted, and the results are plotted in the dataset. The predicted values almost follow the selected model and concluded that model ARIMA(0,4,2) is the best fit for Egg deposition of age-3 Lake Huron Bloaters dataset .

# References

1. Time Series Analysis With Applications in R(Jonathan D. Cryer, Kung-Sik Chan)
2. <https://cran.r-project.org/web/packages/FSAdata/FSAdata.pdf>
3. <https://handbook.unimelb.edu.au/subjects/ecom30004>

# Appendix

## Dataset



## Required Libraries

#Required libraries  
library(readr)  
library(TSA)  
library(dplyr)  
library(ks)  
library(tseries)  
library(fUnitRoots)  
library(lmtest)  
source('C:/Users/SIA/Desktop/SEM 3/Time series Analysis/sort.score.R')

## Pre validations

data1 <- read.csv("C:/Users/SIA/Desktop/SEM 3/Time series Analysis/eggs.csv",header=TRUE)  
#View the sample  
head(data1)

## year eggs  
## 1 1981 0.0402  
## 2 1982 0.0602  
## 3 1983 0.1205  
## 4 1984 0.1807  
## 5 1985 0.7229  
## 6 1986 0.5321

#check the structure  
str(data1)

## 'data.frame': 16 obs. of 2 variables:  
## $ year: int 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 ...  
## $ eggs: num 0.0402 0.0602 0.1205 0.1807 0.7229 ...

#check for impossible and null values  
data1[is.na(data1$Thickness),]

## Warning in is.na(data1$Thickness): is.na() applied to non-(list or vector)  
## of type 'NULL'

## [1] year eggs  
## <0 rows> (or 0-length row.names)

data1[is.nan(data1$Thickness),]

## [1] year eggs  
## <0 rows> (or 0-length row.names)

## Time series plot of R time series object

The time series plot for this data is generated with the following code chunk

#convert into time series objectby eliminating year specific field  
egg\_Bloaters <- ts(as.vector(data1[,-1]),start=1981, end=1996)  
#check the class of the dataset  
class(egg\_Bloaters)

## [1] "ts"

The time series plot for this data is generated with the following code chunk

plot(egg\_Bloaters,type='o',ylab='Egg Depositions',xlab='Years',main="Time series plot for Egg Deposition of Bloaters")

## Plot - relationship between pairs of consecutive egg deposition:

The following code chunk generates a scatter plot to investigate the relationship between pairs of consecutive ozone thickness:

#Plot the relationship diagram

#scatter plot for previous and current year egg deposition   
plot(y=egg\_Bloaters,x=zlag(egg\_Bloaters),ylab='Millions', xlab='Previous Year Egg Deposition',main='Relationship- Previous Vs Current Year')

#Correlation between previous year and current year  
x=egg\_Bloaters  
y=zlag(egg\_Bloaters)  
index=2:length(x)  
cor(y[index],x[index])

## [1] 0.7445657

## Linear Trend

The below R chunk will fit the data into a linear model.

#Linear Model  
model.egg\_Bloaters.ln = lm(egg\_Bloaters~time(egg\_Bloaters))   
summary(model.egg\_Bloaters.ln)

The trend line is plotted Using the below R code:

#plot the linear trend line

plot(egg\_Bloaters,type='o',ylab='y')  
plot.new

abline(model.egg\_Bloaters.ln)

## Quadratic Trend

The below R chunk will fit the data into a quadratic model.

# create the time variables for quadratic trend

t = time(egg\_Bloaters)  
t2 = t^2  
model.egg\_Bloaters.qa = lm(egg\_Bloaters~ t + t2)

# label the quadratic trend  
summary(model.egg\_Bloaters.qa)

The trend curve is plotted Using the below R code:

# Fitted quadratic trend

plot(ts(fitted(model.egg\_Bloaters.qa)), ylim = c(min(c(fitted(model.egg\_Bloaters.qa),  
 as.vector(egg\_Bloaters))), max(c(fitted(model.egg\_Bloaters.qa),as.vector(egg\_Bloaters)))),  
 ylab='y' , main = "Fitted quadratic curve to egg deposition Data", type="l",lty=2,col="red")  
plot.new

lines(as.vector(egg\_Bloaters),type="o")

## ACF and PACF plots

#ACF AND PACF plots to analyses non stationarity   
par(mfrow=c(1,2))  
#Trend is apparent from ACF and PACf plots  
acf(egg\_Bloaters)   
pacf(egg\_Bloaters)

# Slowly decaying pattern in ACF and very high first correlation in PACF  
# implies the existence of trend and nonstationary.

## ADF Test to analyses non stationarity

# Apply ADF test with default settings  
adf.test(egg\_Bloaters)

# With a p-value of 0.5469, we cannot reject the null hypothesis stating that  
# the series is non-stationary.

## Ar() and adfTest to check normality.

The following R chunk shows standardized residuals from linear model of the ozone thickness data fitted by means of the data:

ar(diff(egg\_Bloaters))

##   
## Call:  
## ar(x = diff(egg\_Bloaters))  
##   
##   
## Order selected 0 sigma^2 estimated as 0.1841

adfTest(egg\_Bloaters, lags = 0, title = NULL,description = NULL)

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 0  
## STATISTIC:  
## Dickey-Fuller: -0.4911  
## P VALUE:  
## 0.452   
##   
## Description:  
## Sat May 11 11:32:16 2019 by user: SIA

## Box cox transformation, CI level and ADF for the raw data and transformed data.

# The default method for fitting is MLE here  
# If do not get a smooth curve with MLE, you can change the method to least squares or Method of Moments  
par(mfrow=c(1,2))  
egg.transform = BoxCox.ar(egg\_Bloaters,main ,'MLE')   
egg1.transform = BoxCox.ar(egg\_Bloaters, method = "yule-walker",main='Least Squares')

#check the range of lambda  
egg.transform$ci

## [1] 0.7 1.0

egg1.transform$ci

## [1] 0.1 0.8

#apply boxcox transformation  
lambda = 0.45  
BC.egg = (egg\_Bloaters^lambda-1)/lambda  
adf.test(egg\_Bloaters)

## Differencing

#d=1

diff.egg = diff(egg\_Bloaters,differences = 1)

#plot series d=1

plot(diff.egg,type='o',ylab='Quarterly earnings ')

#calculate the order

order = ar(diff(diff.egg))$order

#adf test

adfTest(diff.egg, lags = order, title = NULL,description = NULL)

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 4  
## STATISTIC:  
## Dickey-Fuller: -0.7808  
## P VALUE:  
## 0.3601   
##   
## Description:  
## Sat May 11 12:55:34 2019 by user: SIA

#d=2

diff.egg = diff(egg\_Bloaters,differences = 2)

#plot series d=2

plot(diff.egg,type='o',ylab='Quarterly earnings ')

#calculate the order

order = ar(diff(diff.egg))$order

#adf test  
adfTest(diff.egg, lags = order, title = NULL,description = NULL)

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 4  
## STATISTIC:  
## Dickey-Fuller: -1.3974  
## P VALUE:  
## 0.1643   
##   
## Description:  
## Sat May 11 12:55:35 2019 by user: SIA

d=3

diff.egg = diff(egg\_Bloaters,differences = 3)

#plot series d=3

plot(diff.egg,type='o',ylab='Quarterly earnings ')

#calculate the order

order = ar(diff(diff.egg))$order

#adf test  
adfTest(diff.egg, lags = order, title = NULL,description = NULL)

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 4  
## STATISTIC:  
## Dickey-Fuller: -0.7284  
## P VALUE:  
## 0.3767   
##   
## Description:  
## Sat May 11 12:55:35 2019 by user: SIA

#d=4

diff.egg = diff(egg\_Bloaters,differences = 4)

#plot series d=4

plot(diff.egg,type='o',ylab='Quarterly earnings ')

#calculate the order

order = ar(diff(diff.egg))$order

#adf test

adfTest(diff.egg, lags = order, title = NULL,description = NULL)

##   
## Title:  
## Augmented Dickey-Fuller Test  
##   
## Test Results:  
## PARAMETER:  
## Lag Order: 2  
## STATISTIC:  
## Dickey-Fuller: -2.1524  
## P VALUE:  
## 0.03368   
##   
## Description:  
## Sat May 11 12:55:36 2019 by user: SIA

#Check the normality of the differenced series with d=4

shapiro.test(diff.egg)

## ACF and PACF plot of egg deposition data with d=4

#ACF AND PACF plots to identify the order   
par(mfrow=c(1,2))  
acf(diff.egg)   
pacf(diff.egg)

## Extended auto correlation Function

#eacf table for d=4  
eacf(diff.egg,ar.max = 3, ma.max = 2)

## Bayesian Information Criterion (BIC)

#BIC  
res = armasubsets(y=diff.egg,nar=2,nma=2,y.name='test',ar.method='ols')

plot(res)

## Parameter Estimation (Maximum Likelihood)

#Maximum likelihood of the AR and MA coefficient with significance tests   
# ARIMA(0,4,1)  
model\_041\_ml = arima(egg\_Bloaters,order=c(0,4,1),method='ML')  
# ARIMA(1,4,1)  
model\_141\_ml = arima(egg\_Bloaters,order=c(1,4,1),method='ML')  
# ARIMA(0,4,2)  
model\_042\_ml = arima(egg\_Bloaters,order=c(0,4,2),method='ML')  
# ARIMA(1,4,2)  
model\_142\_ml = arima(egg\_Bloaters,order=c(1,4,2),method='ML')  
# ARIMA(2,4,1)  
model\_241\_ml = arima(egg\_Bloaters,order=c(2,4,1),method='ML')

#significance tests for each parameter coeftest()  
coeftest(model\_041\_ml)

coeftest(model\_141\_ml)

coeftest(model\_042\_ml)

coeftest(model\_142\_ml)

coeftest(model\_241\_ml )

## Parameter Estimation (Least squares)

#Least squares estimate of the AR and MA coefficient with significance tests   
# ARIMA(0,4,1)  
model\_041\_css = arima(egg\_Bloaters,order=c(0,4,1),method='CSS')  
# ARIMA(1,4,1)  
model\_141\_css = arima(egg\_Bloaters,order=c(1,4,1),method='CSS')  
# ARIMA(0,4,2)  
model\_042\_css = arima(egg\_Bloaters,order=c(0,4,2),method='CSS')  
# ARIMA(1,4,2)  
model\_142\_css = arima(egg\_Bloaters,order=c(1,4,2),method='CSS')  
# ARIMA(2,4,1)  
model\_241\_css = arima(egg\_Bloaters,order=c(2,4,1),method='CSS')

#significance tests for each parameter coeftest()  
coeftest(model\_041\_css)

coeftest(model\_141\_css)

coeftest(model\_042\_css)

coeftest(model\_142\_css)

coeftest(model\_241\_css)

## Sort AIC and BIC for all candidate models

sort.score(AIC(model\_041\_ml,model\_141\_ml,model\_042\_ml,model\_142\_ml,model\_241\_ml), score = "aic")

sort.score(BIC(model\_041\_ml,model\_141\_ml,model\_042\_ml,model\_142\_ml,model\_241\_ml), score = "bic" )

# Both AIC and BIC select ARIMA(0,4,2) model for this series

## Sort.score-user defined function



## Function- ‘residual.analysis’

residual.analysis <- function(model, std = TRUE){  
 library(TSA)  
 library(FitAR)  
 if (std == TRUE){  
 res.model = rstandard(model)  
 }else{  
 res.model = residuals(model)  
 }  
 par(mfrow=c(3,2))  
 plot(res.model,type='o',ylab='Standardised residuals', main="Time series plot of standardised residuals")  
 abline(h=0)  
 hist(res.model,main="Histogram of standardised residuals")  
 qqnorm(res.model,main="QQ plot of standardised residuals")  
 qqline(res.model, col = 2)  
 acf(res.model,main="ACF of standardised residuals")  
 print(shapiro.test(res.model))  
 print(Box.test(res.model, lag = 6, type = "Ljung-Box", fitdf = 0))  
 k=0  
 LBQPlot(res.model, lag.max = length(model$residuals)-1 , StartLag = k + 1, k = 0, SquaredQ = FALSE)  
 par(mfrow=c(1,1))  
}

## Overfitting models

# ARIMA (0,4,3)

model\_043\_ml = arima(egg\_Bloaters,order=c(0,4,3),method='ML')

coeftest(model\_043\_ml)

# ARIMA (1,4,2)

model\_142\_ml = arima(egg\_Bloaters,order=c(1,4,2),method='ML')  
coeftest(model\_142\_ml)