MATH1318 - Time Series Analysis

Assignment 2 Project Report

Forecasting
on
Egg Depositions(in millions) of age 3 Lake Huron Bloaters data

By

Ayaz Aziz Mujawar (s3751555)

Abstract

The report aims to analyze and find the best fitting model to forecast the egg depositions of lake Huron bloaters by using a set of possible ARIMA(p,d,q) models.

Introduction

Bloater is the species of freshwater whitefish found in water bodies like lake Huron and they inhabit underwater slopes. The number of eggs produced by this type of female fish varies between 3000 and 12000. Time Series Analysis techniques are the best way to measure and choose the best model among the set of possible models and forecast the egg depositions for the next 5 years.

Methodology

The dataset provided represents the egg depositions(in millions) of age 3 bloaters between the years 1981 and 1996. The dataset is taken from the FSA data package After reading the data into the data frame it is then converted into a time series object and then the time series plot is applied to it. Observation of the plot is done based on trend, variance, seasonality, autocorrelation, and intervention. Based on these observations we have done our further analysis.

General Analysis

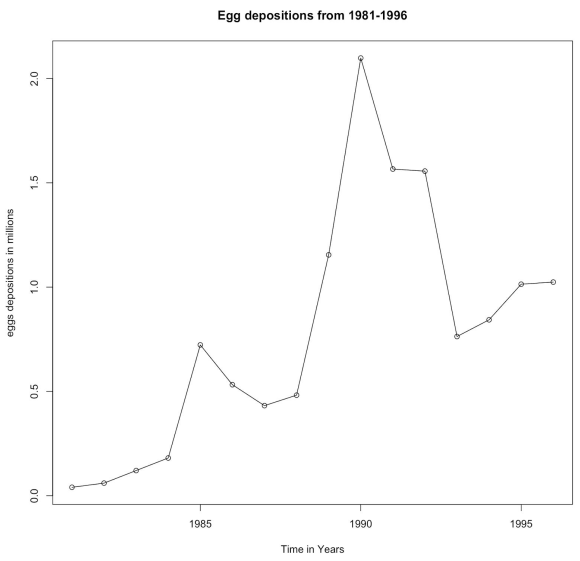


Fig: 1.1

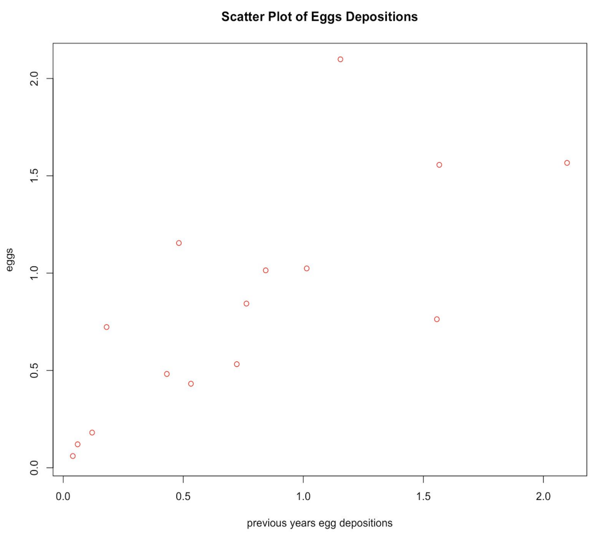


Fig: 1.2

From the figure(Fig 1.1), We can observe that there seems to be an upward trend as well as a slight change in the value of mean over time. There is no repeating pattern present in the time series so there seems to be no obvious seasonality present in the plot. Also, a change in variance is not observed in the time series. There is also no clear intervention point observed. There might be an intervention in the year 1990 but research is needed to verify whether it is an intervention. From the time series plot, we can also observe that the series is in stochastic form and successive points are also observed through the time in the series which is an indicator of autoregressive behavior.

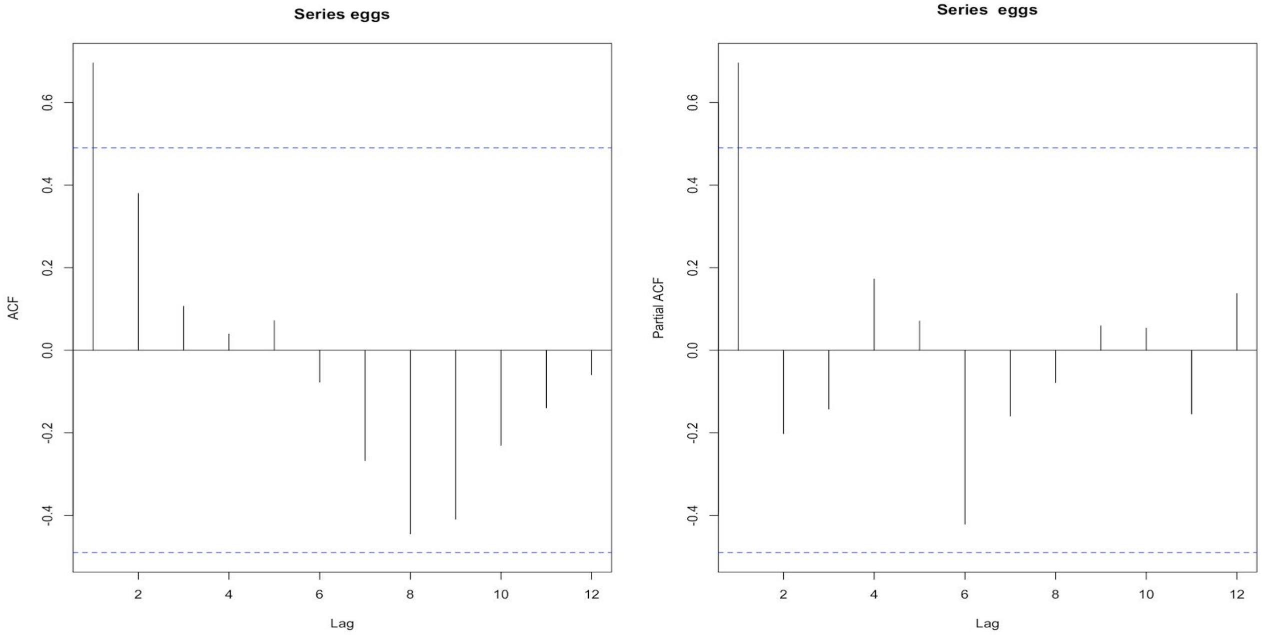


Fig: 1.3

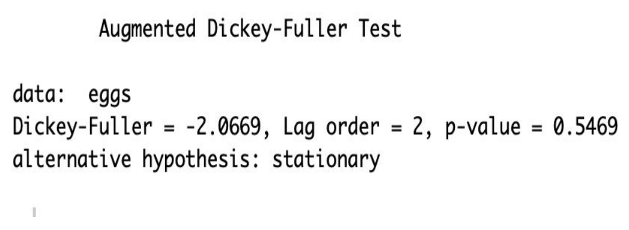
From the ACF and PACF plot (Fig: 1.3) we can observe that there is a trend in the series and also series is in non-stationary form. So we will try multiple ARIMA models for the given time series data and choose the best-fitted model.

ARIMA Modelling

Before proceeding to the ARIMA modeling we need to make the series stationary. We will start by performing the ADF test to check whether the time series data is in a stationary or non-stationary format.

The assumption for the ADF test:

H0: The given series is non-stationary
Ha:: The given series is stationary



```
        Augmented Dickey-Fuller Test

data:  eggs
Dickey-Fuller = -2.0669, Lag order = 2, p-value = 0.5469
alternative hypothesis: stationary
```

Fig: 1.4

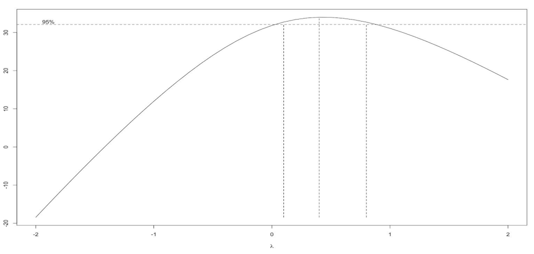
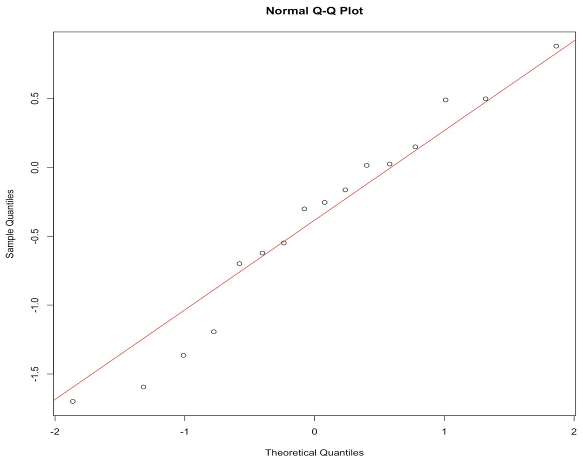


Fig: 1.5

From the ADF test(Fig: 1.4), we infer that as the p-value is greater than 0.05(0.5469>0.05) we failed to reject the null hypothesis, and hence the series is non-stationary.

To remove the trend in the series we applied Box-Cox transformation. The optimal value of λ is given within the range between 0.1 and 0.8. From the lambda values, we can say that the optimal value is between 0 and 1 so we have taken 0.45 as an optimal λ value.



:

Fig: 1.7

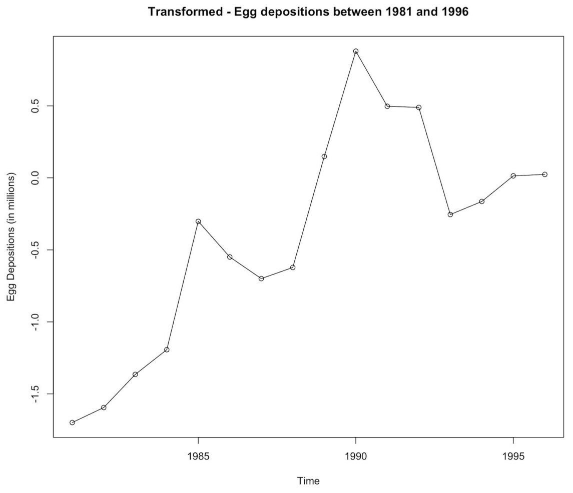


Fig: 1.8

From Fig: 1.7 (QQ-plots) and Fig: 1.8(transformed time series plot) we can claim that the BoxCox transformation using lambda value doesn't change the original time series that much. So we will apply the first difference to the original data time series directly.

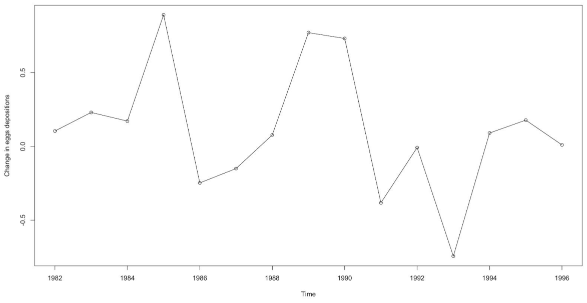
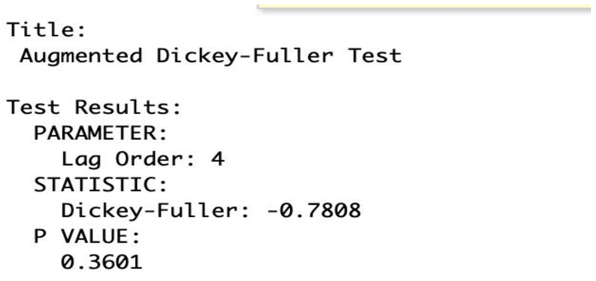


Fig: 1.9



```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 4
  STATISTIC:
    Dickey-Fuller: -0.7808
  P VALUE:
    0.3601
```

Fig: 1.10

After first differencing to the time series. From Fig: 1.9 we observe that there is a downward trend in the time series and also from Fig: 1.10, we observed that we cannot reject the null hypothesis as the p-value(0.3601) >

p-value(0.05). Hence, we conclude that the first difference series is still non-stationary. So we applied the second difference to the series.

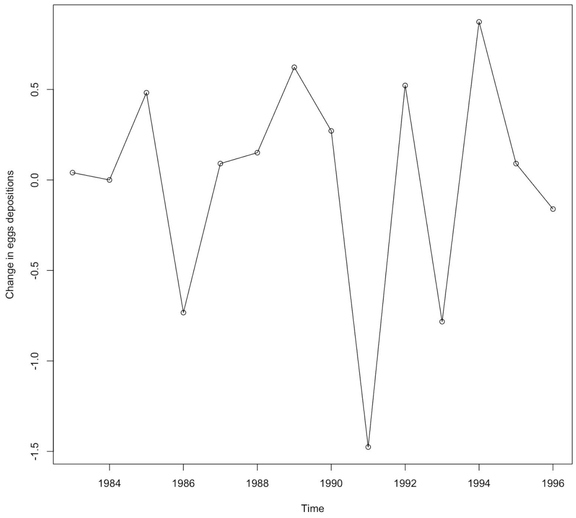
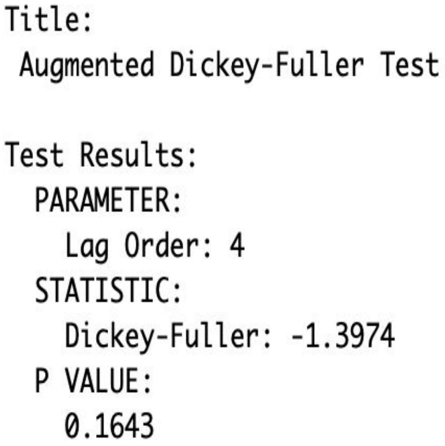


Fig: 1.11



```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 4
  STATISTIC:
    Dickey-Fuller: -1.3974
  P VALUE:
    0.1643
```

Fig: 1.12

After a second differencing to the time series. From Fig: 1.11 we observe that there is a loss in the trend in the time series and also from Fig: 1.12, we observed that we cannot reject the null hypothesis as the p-value(0.1643) > p-value(0.05). Hence, we conclude that the second difference series is still non-stationary. So we will apply the third difference to the series.

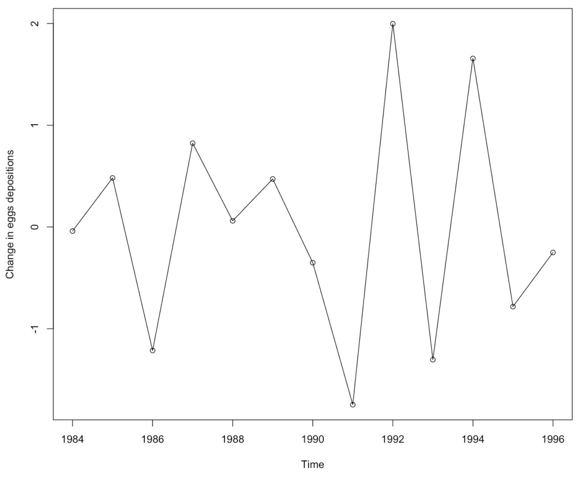
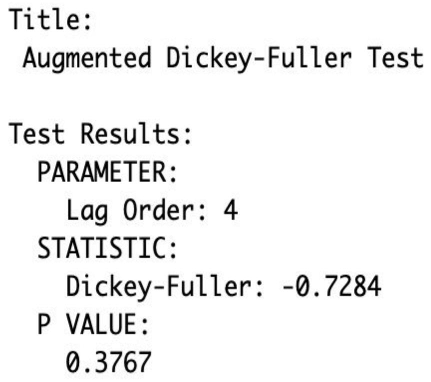


Fig: 1.13



```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 4
  STATISTIC:
    Dickey-Fuller: -0.7284
  P VALUE:
    0.3767
```

Fig: 1.14

After the third differencing to the time series, From Fig: 1.13 we observe that there is not much change in the trend in the time series but we still cannot reject the null hypothesis as the p-value(0.3767) > p-value(0.05). Hence, we conclude that the third difference series is still non-stationary. So we will apply fourth differencing to the series.

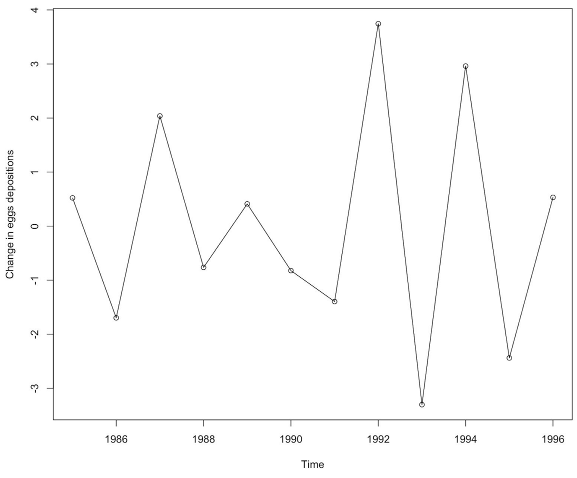
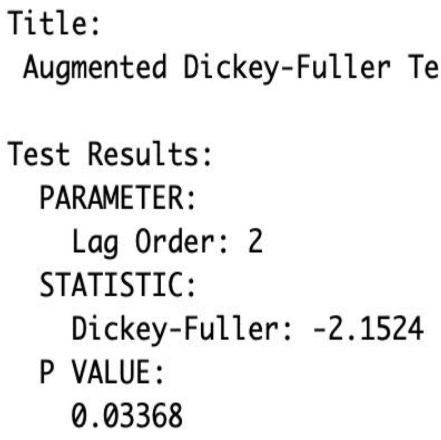


Fig: 1.15



```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 2
  STATISTIC:
    Dickey-Fuller: -2.1524
  P VALUE:
    0.03368
```

Fig: 1.16

Finally after applying fourth differencing to the time series, From Fig: 1.15 we observe that the trend remains almost constant as compared to the previous trend in the time series but here we rejected the null hypothesis as the p-value(0.03368) < p-value(0.05). Hence, we conclude that the fourth difference makes the series stationary.

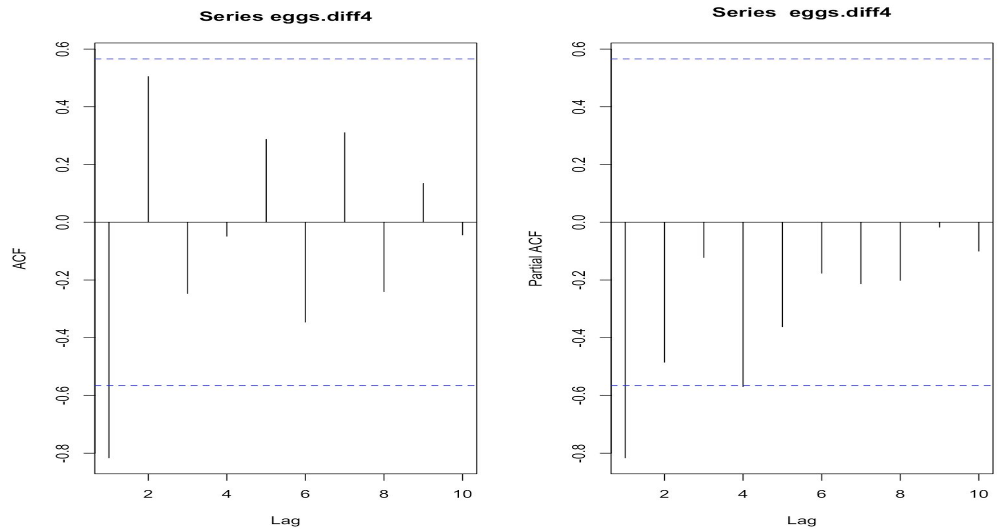
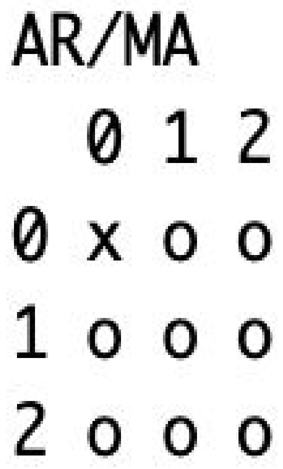


Fig: 1.17

From Fig: 1.17 we witness that the trend component is totally absent in the plots and hence differencing has made the series stationary. However, we are able to 1 significant lag in both ACF and PACF plots. So the candidate model is ARIMA(1,4,1).

We also used other tools like EACF - Extended Autocorrelation functions and BIC tables (armasubsets). to find candidate models.



```
AR/MA
  0 1 2
0 x o o
1 o o o
2 o o o
```

Fig: 1.18

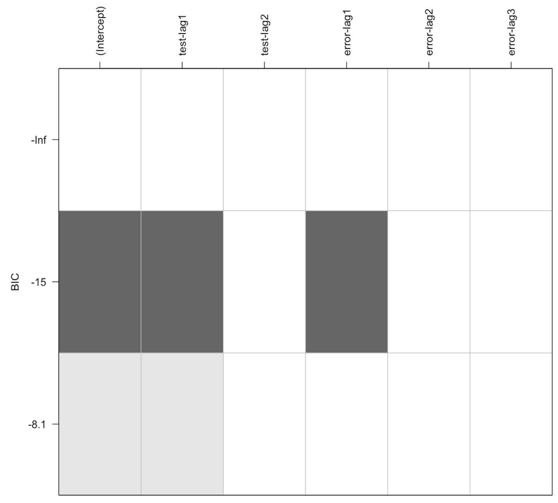
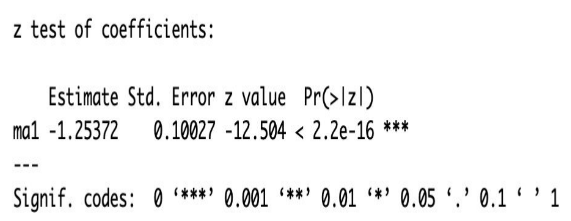


Fig: 1.19

From the Fig: 1.18 of EACF and Fig:1.19 of BIC table, the set of possible model values we observe are

ARIMA(0,4,1), ARIMA(1,4,1), ARIMA(2,4,1), ARIMA(0,4,2), and ARIMA(1,4,2).

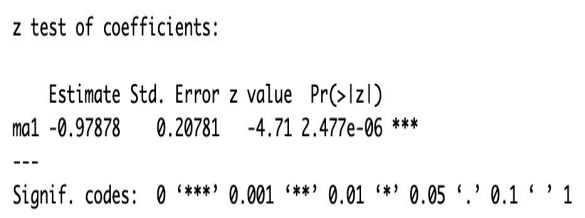
Parameter Estimation



```
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
```

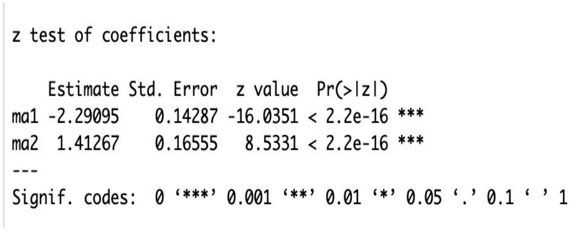
ARIMA(0,4,1) - CSS



```
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
```

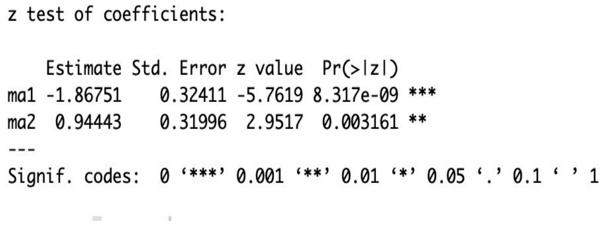
ARIMA(0,4,1) - ML



```
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
```

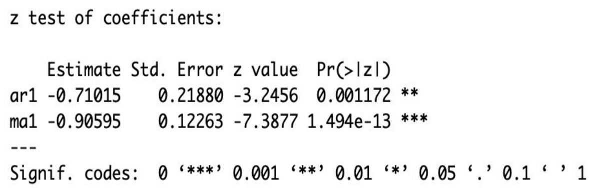
ARIMA(0,4,2) - CSS



```
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
```

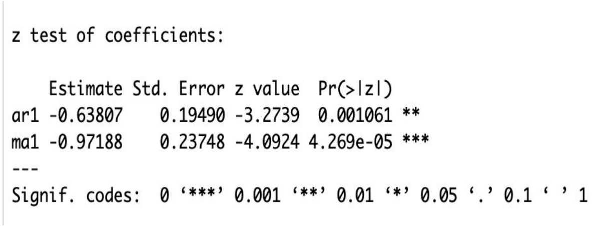
ARIMA(0,4,2) - ML



```
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
```

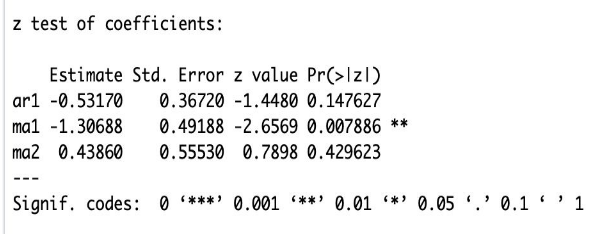
ARIMA(1,4,1) - CSS



```
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
```

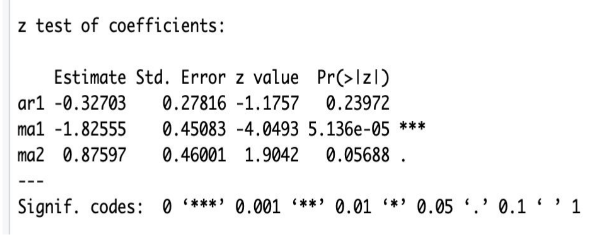
ARIMA(1,4,1) - ML



```
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
```

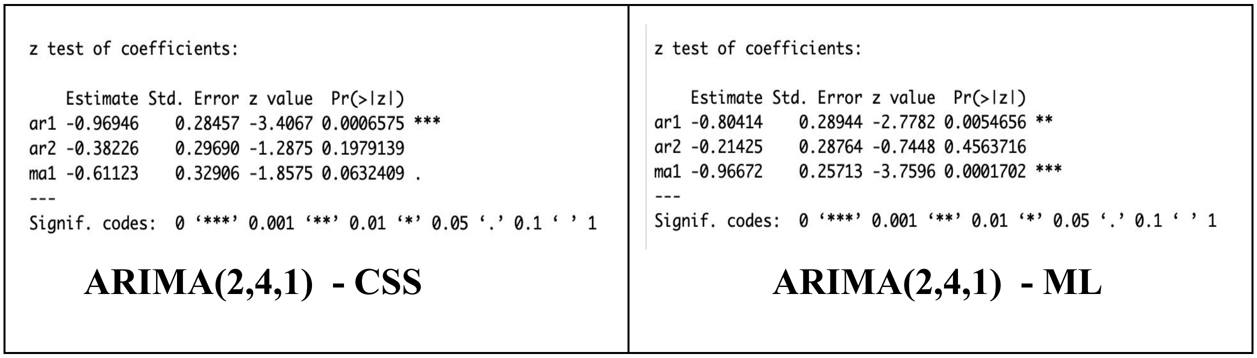
ARIMA(1,4,2) - CSS



```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ar1 -0.32703    0.27816 -1.1757   0.23972
ma1 -1.82555    0.45083 -4.0493 5.136e-05 ***
ma2  0.87597    0.46001  1.9042   0.05688 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ARIMA(1,4,2) - ML



```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ar1 -0.96946    0.28457 -3.4067 0.0006575 ***
ar2 -0.38226    0.29690 -1.2875 0.1979139
ma1 -0.61123    0.32906 -1.8575 0.0632409 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ARIMA(2,4,1) - CSS

```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ar1 -0.80414    0.28944 -2.7782 0.0054656 **
ar2 -0.21425    0.28764 -0.7448 0.4563716
ma1 -0.96672    0.25713 -3.7596 0.0001702 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

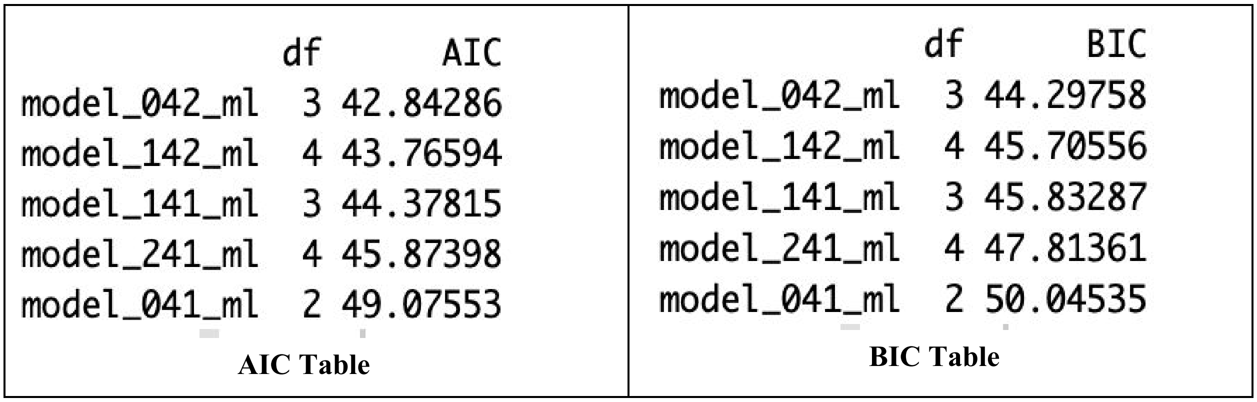
ARIMA(2,4,1) - ML

Fig: 1.20

By using parameter estimation we applied both maximum likelihood and least square estimates and then we selected the best model based on the AIC and BIC.

Analysis on parameters Estimation

From Fig 1.20 we can state that for the ARIMA(0,4,1) model, MA(1) component is significant by using both CSS and ML parameters. For the ARIMA(0,4,2) model, MA(1) and MA(2) components are significant using both CSS and ML methods. In model ARIMA(1,4,1), both AR(1) and ML(1) components are significant using both CSS and ML parameters. For the ARIMA(1,4,2) model only MA(1) component is significant for both CSS and ML parameters. Finally, for the ARIMA(2,4,1) model, the AR(1) component is significant for both CSS and ML parameters, and only the MA(1) is significant by using ML parameters.



```
             df      AIC
model_042_ml  3 42.84286
model_142_ml  4 43.76594
model_141_ml  3 44.37815
model_241_ml  4 45.87398
model_041_ml  2 49.07553
```

AIC Table

```
             df      BIC
model_042_ml  3 44.29758
model_142_ml  4 45.70556
model_141_ml  3 45.83287
model_241_ml  4 47.81361
model_041_ml  2 50.04535
```

BIC Table

Fig: 1.21

From the fig: 1.21 of AIC and BIC, we observe that the ARIMA(0,4,2) model has the smaller AIC and BIC value So this model is considered the best suitable model in the series for forecasting.

Model Diagnostic

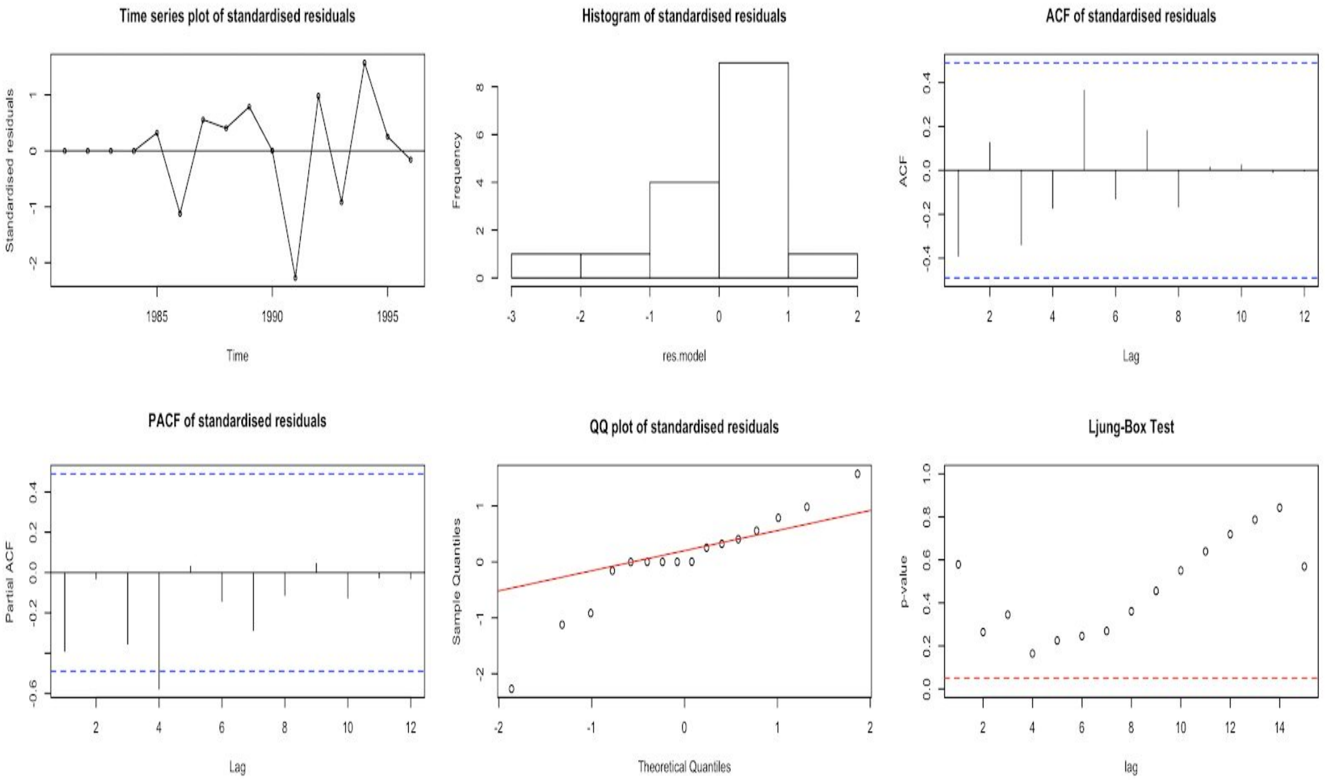
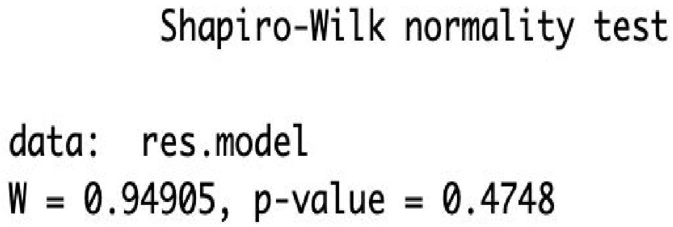


Fig: 1.22



```
Shapiro-Wilk normality test

data:  res.model
W = 0.94905, p-value = 0.4748
```

Fig: 1.23

From the Model Diagnostic plots (Fig: 1.22) we observe that there is no trend and changing variance in the time series. The histogram shown here is non-symmetric. Both ACF and PACF plots can be considered as the correlation of standard residuals and it doesn't contain any significant lags, so it can be considered as white noise. In the QQ-plot most of the points are falling nearby the straight line. The Ljung-Box plot shows that all the points lie above the red line at about 5% and the Shapiro-Wilk test(Fig: 1.23) also states that the p-values are greater than 0.05(i.e. p-value = 0.4748) which means we fail to reject normality. So from this, we conclude that as most of the model diagnostics are passing the normality test, ARIMA(0,4,2) can be considered as the best model for forecasting.

Forecasting

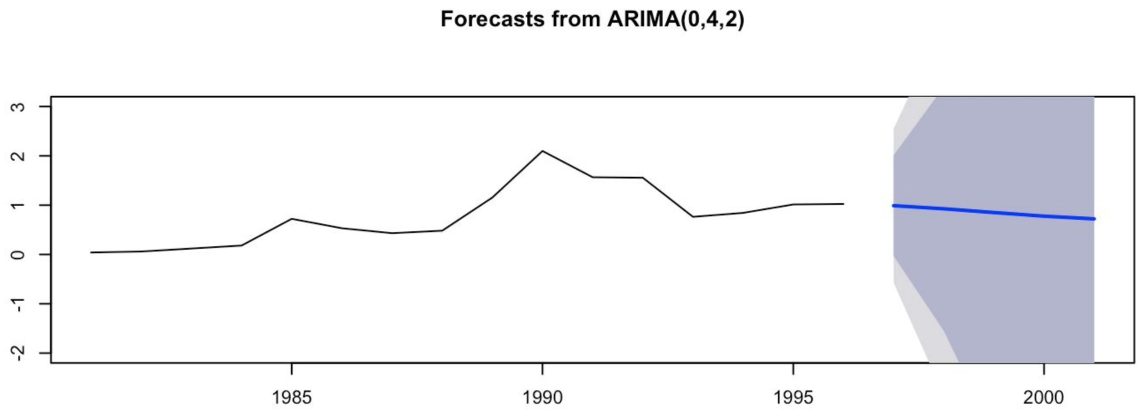


Fig: 1.24

From the Fig: 1.24 of Forecast we can observe that there will be a slight decrease in the trend of the egg decomposition for the next 5 years.

Conclusion

In this report we can conclude that ARIMA(0,4,2) is the best model as compared to other ARIMA models based on the aic and bic score. Forecast shows the decrease in the trend of egg depositions for the next 5 years.

Summary

In this report, Firstly, differencing is used to make the time series stationary. Then with the help of ACF, PACF, EACF, and BIC candidate models were selected. Then CSS and ML parameter estimation are applied on the models and based on the aic and bic score best ARIMA model (ARIMA(0,4,2)) was selected. Finally, that ARIMA(0,4,2) model is diagnosed with different model diagnostic techniques, and then forecasting is applied to that model showing the egg depositions for the next 5 years.

Appendix

```
#importing the libraries
library(TSA)
library(tseries)
library(fUnitRoots)
library(lmtest)
library(forecast)
library(FitAR)
source('./Documents/RMIT_Sem_3/Time Series Analysis/Assignment-2/sort.score.R')

#Task 1
#Read the data
eggs <- read.csv("./Documents/RMIT_Sem_3/Time Series Analysis/Assignment-2/eggs.csv")
head(eggs)
eggs$year <- NULL

#convert the data to timeseries format
eggs <- ts(as.vector(eggs), start=1981, end=1996)
#Checking whether the dataset is in timeseries format
is.ts(eggs)

#Plotting the time series plot
plot(eggs,type='o', xlab="Time in Years",ylab='eggs depositions in millions' ,main='Egg depositions from
1981-1996')

#To check the thickness of the ozone layer thickness over the year shown using correlation value
x = eggs
y = zlag(eggs)
index = 2:length(y)
cor(x[index], y[index])

#Scatter plot
plot(y=eggs,x=zlag(eggs),col=c("red"),xlab = "previous years egg depositions",main = "Scatter Plot of
Eggs Depositions")

#Plot the acf and pacf plot
par(mfrow = c(1,2))
acf(eggs)
pacf(eggs)
par(mfrow = c(1,1))

#adf test
adf.test(eggs)

#### From the acf and pacf we can conclude that the TimeSeries is Non stationary
# and there is a trend in the series

#Due to lot of variation in the data we perform Box-Cox Transformations
eggs_box = BoxCox.ar(eggs, method = 'yule-walker')

#Confidence Interval
eggs_box$ci

#drawing the plot with box-cox
```

```
lambda = 0.45
eggs.bc = ((eggs^lambda)-1)/lambda
plot(eggs.bc, type = "o", ylab = "Egg Depositions (in millions)", main = "Transformed - Egg depositions
between 1981 and 1996")

qqnorm(eggs.bc)
qqline(eggs.bc, col = 2)
shapiro.test(eggs.bc)

#For making the non stationary time series stationary, we are using 1st differencing.
eggs.diff1 <- diff(eggs, differences=1)
plot(eggs.diff1, ylab='Change in eggs depositions', type='o')

# -------- ADF Test --------
#We are using an ADF test to check whether the timeseries is stationary or non-stationary.
#Hypothesis (Null value) -- Time Series is Non-stationary.

#if the p-value of the test is less than the significance level (0.05)
#then you reject the null hypothesis and infer that the time series is indeed stationary.

order  = ar(diff(eggs.diff1))$order # To pass the order to adfTest function
adfTest(eggs.diff1, lags = order,  title = NULL,description = NULL)

#differencing 2
eggs.diff2 <- diff(eggs, differences=2)
plot(eggs.diff2, ylab='Change in eggs depositions', type='o')

order  = ar(diff(eggs.diff2))$order # To pass the order to adfTest function
adfTest(eggs.diff2, lags = order,  title = NULL,description = NULL)

#differencing 3
eggs.diff3 <- diff(eggs, differences=3)
plot(eggs.diff3, ylab='Change in eggs depositions', type='o')

order  = ar(diff(eggs.diff3))$order # To pass the order to adfTest function
adfTest(eggs.diff3, lags = order,  title = NULL,description = NULL)

#differencing 4
eggs.diff4 <- diff(eggs, differences=4)
plot(eggs.diff4, ylab='Change in eggs depositions', type='o')

order  = ar(diff(eggs.diff4))$order # To pass the order to adfTest function
adfTest(eggs.diff4, lags = order,  title = NULL,description = NULL)

#For making the non stationary timeseries stationary, we are using 1st differencing.
#eggs.bc.diff2 <- diff(eggs.bc, differences=2)
#plot(eggs.bc.diff2, ylab='Change in eggs depositions', type='l')
```

```
#adf.test(eggs.bc.diff2)

#Plot the acf and pacf plot
par(mfrow = c(1,2))
acf(eggs.diff4)
pacf(eggs.diff4)
par(mfrow = c(1,1))

#eacf
eacf(eggs.diff4,ar.max = 2, ma.max = 2)

#Answers - (0,4,1), (0,4,2), (1,4,1), (1,4,2), (2,4,1), (2,4,2)

#BIC Table
par(mfrow=c(1,1))
res3 = armasubsets(y=eggs.diff4,nar=2,nma=3,y.name='test',ar.method='ols')
plot(res3)

#Answers - (1,4,1)
-----------------------------------------------------------
#ARIMA(0,4,1)
model_041_css = arima(eggs,order=c(0,4,1),method='CSS')
coeftest(model_041_css)
model_041_ml = arima(eggs,order=c(0,4,1),method='ML')
coeftest(model_041_ml)

#ARIMA(0,4,2)
model_042_css = arima(eggs,order=c(0,4,2),method='CSS')
coeftest(model_042_css)

model_042_ml = arima(eggs,order=c(0,4,2),method='ML')
coeftest(model_042_ml)

#ARIMA(1,4,1)
model_141_css = arima(eggs,order=c(1,4,1),method='CSS')
coeftest(model_141_css)

model_141_ml = arima(eggs,order=c(1,4,1),method='ML')
coeftest(model_141_ml)

#ARIMA(1,4, 2)
model_142_css = arima(eggs,order=c(1,4,2),method='CSS')
coeftest(model_142_css)

model_142_ml = arima(eggs,order=c(1,4,2),method='ML')
coeftest(model_142_ml)
```

```
#ARIMA(2,4,1)
model_241_css = arima(eggs.bc,order=c(2,4,1),method='CSS')
coeftest(model_241_css)

model_241_ml = arima(eggs,order=c(2,4,1),method='ML')
coeftest(model_241_ml)

# AIC and BIC values

sort.score(AIC(model_041_ml,model_042_ml,model_141_ml,model_142_ml,model_241_ml), score =
"aic")
sort.score(BIC(model_041_ml,model_042_ml,model_141_ml,model_142_ml,model_241_ml), score =
"bic")

#Model Diagnostic

residual.analysis <- function(model, std = TRUE,start = 2, class = c("ARIMA","GARCH","ARMA-GA
RCH")[1]){
  # If you have an output from arima() function use class = "ARIMA"
  # If you have an output from garch() function use class = "GARCH"
  # If you have an output from ugarchfit() function use class = "ARMA-GARCH"
  library(TSA)
  library(FitAR)
 if (class == "ARIMA"){
    if (std == TRUE){
      res.model = rstandard(model)
    }else{
      res.model = residuals(model)
    }
  }else if (class == "GARCH"){
    res.model = model$residuals[start:model$n.used]
  }else if (class == "ARMA-GARCH"){
    res.model = model@fit$residuals
  }else {
    stop("The argument 'class' must be either 'ARIMA' or 'GARCH' ")
  }
  par(mfrow=c(3,2))
  plot(res.model,type='o',ylab='Standardised residuals', main="Time series plot of standardis
      ed residuals")
  abline(h=0)
  hist(res.model,main="Histogram of standardised residuals")
  acf(res.model,main="ACF of standardised residuals")
  pacf(res.model,main="PACF of standardised residuals")
  qqnorm(res.model,main="QQ plot of standardised residuals")
  qqline(res.model, col = 2)
  print(shapiro.test(res.model))
  k=0
  LBQPlot(res.model, lag.max = 10, StartLag = k + 1, k = 0, SquaredQ = FALSE)
}
residual.analysis(model = model_042_css)

#Forecasting

forecasting = Arima(eggs,c(0,4,2))
plot(forecast(forecasting, h=5), ylim =c(-2,3))
```

References

1. https://en.wikipedia.org/wiki/Coregonus_hoyi
2. MATH1318 Time Series Analysis notes and tutorials by Dr. Haydar Demirhan.