Time Series Analysis - Assigment 2

Loan Huynh - s3655461

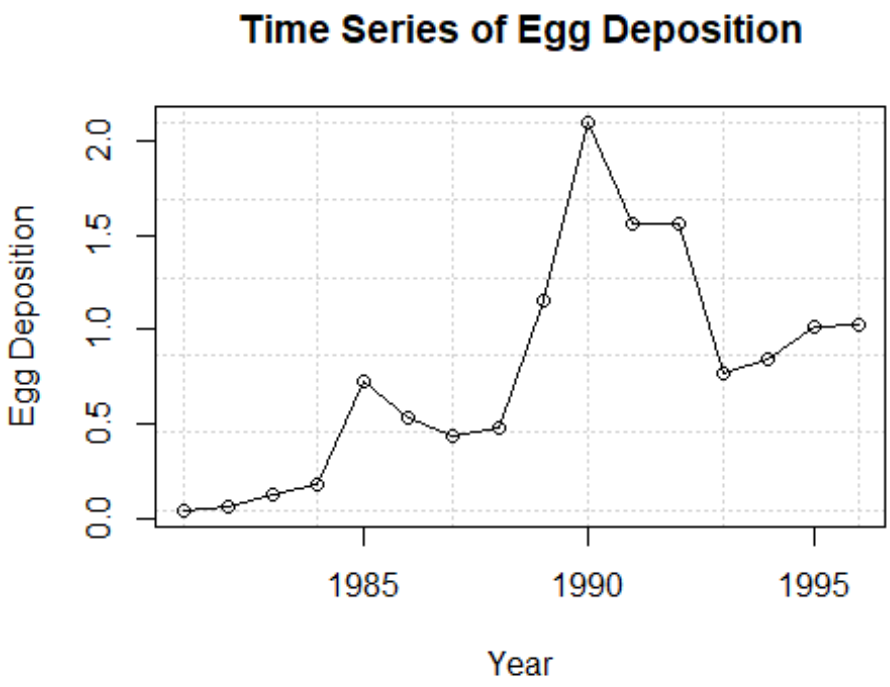
Introduction

This report tries to find the suitable model which used to predict Egg deposition in next 5 years. The dataset is Egg depositions of age-3 Lake Huron Bloasters between 1981 and 1996 in **BloaterLH** dataset of **FSAdata** package

```
# Loadpackage
library(FSAdata)
library(TSA)
library(fUnitRoots) # package used to test Non-stationary series
library(tseries)
library(lmtest)
library(forecast) # package used for forecasting
```

Data Description

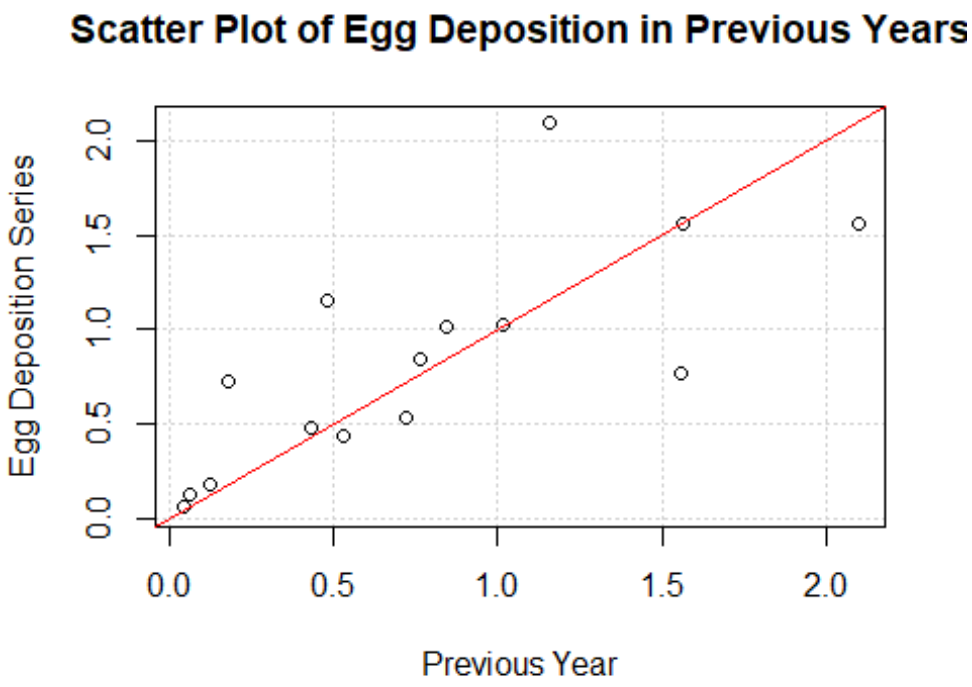
```
# Load dataset
data(BloaterLH)

# Only get Eggs column data in BloaterLH
egg <- BloaterLH$eggs

# Create Time Series objects for Egg data
egg_ts <- ts(egg, start = 1981, end = 1996)

# Plot Egg Deposition Series data
plot(egg_ts, type = 'o', xlab= 'Year',ylab="Egg Deposition",
panel.first=grid(), main = "Time Series of Egg Deposition")
```



This series has an upward trend and changes of variation, but no repeating pattern in seasonality. There are some suceeding points but no fluctuation. It looks like the autoregressive

Using Scatter Plot to see the relationship between Egg Deposition of previous years

```
# Scatter Plot
plot(y= egg_ts , x = zlag(egg_ts ), ylab = 'Egg Deposition Series', xlab =
'Previous Year',
     panel.first=grid(), main = "Scatter Plot of Egg Deposition in Previous
Years")
abline(coef = c(0,1), col= "red")
```



It is obviously to see that there is a strong correlation between Egg Depositions of previous Years in Scatter Plot. The following code is to calculate the value of correlation

```
y <- egg_ts # Load data into y

x <- zlag(egg_ts) # Generate first lag of the abundance series

index <- 2:length(x)# Create index to get rid of the first NA value in x

cor(y[index], x[index])# Calculate the correlation between numerical values
in x and y

## [1] 0.7445657
```

The correlation is 0.7445657. This is a strong positive correlation

Apply ADF Test to check whether series in stationary or non-stationary

```
ar(diff(egg_ts))

##
## Call:
## ar(x = diff(egg_ts))
##
##
## Order selected 0  sigma^2 estimated as  0.1841
```

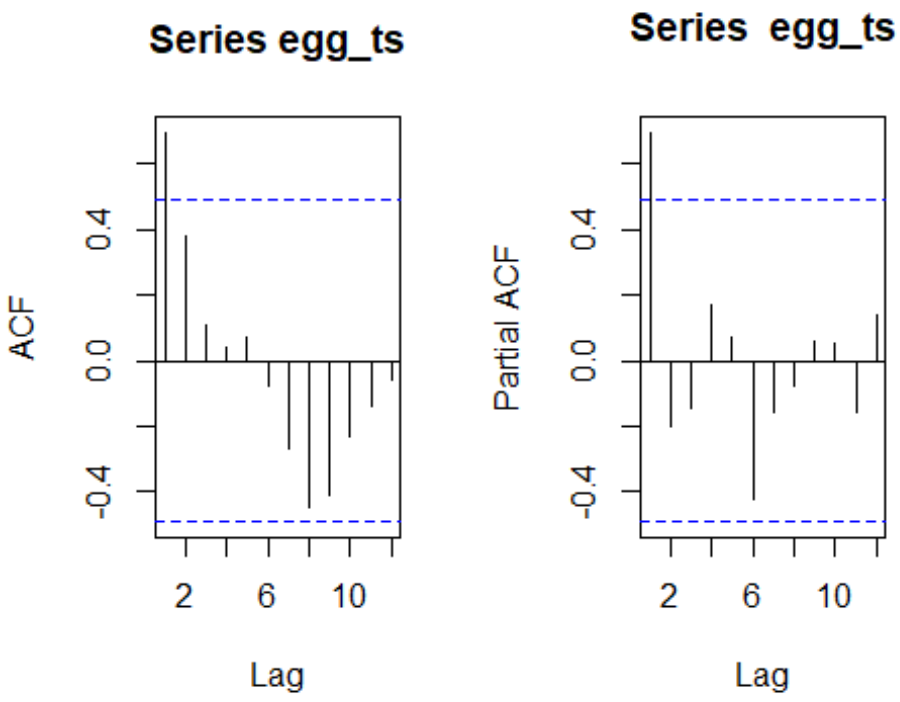
Order of this test is 0

```
adfTest(egg_ts, lags = 0)
```

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 0
##   STATISTIC:
##     Dickey-Fuller: -0.4911
##   P VALUE:
##     0.452
##
## Description:
##  Sun May 12 23:46:41 2019 by user: loanh
```

The p_value is 0.452 larger than 5% level of significant. Therefore we cannot reject null hypothesis about non-stationary. It implies that this series is non-stationary. It is clearly shown in ACF and PACF plots as below:

```
par(mfrow=c(1,2))
acf(egg_ts)
pacf(egg_ts)
```



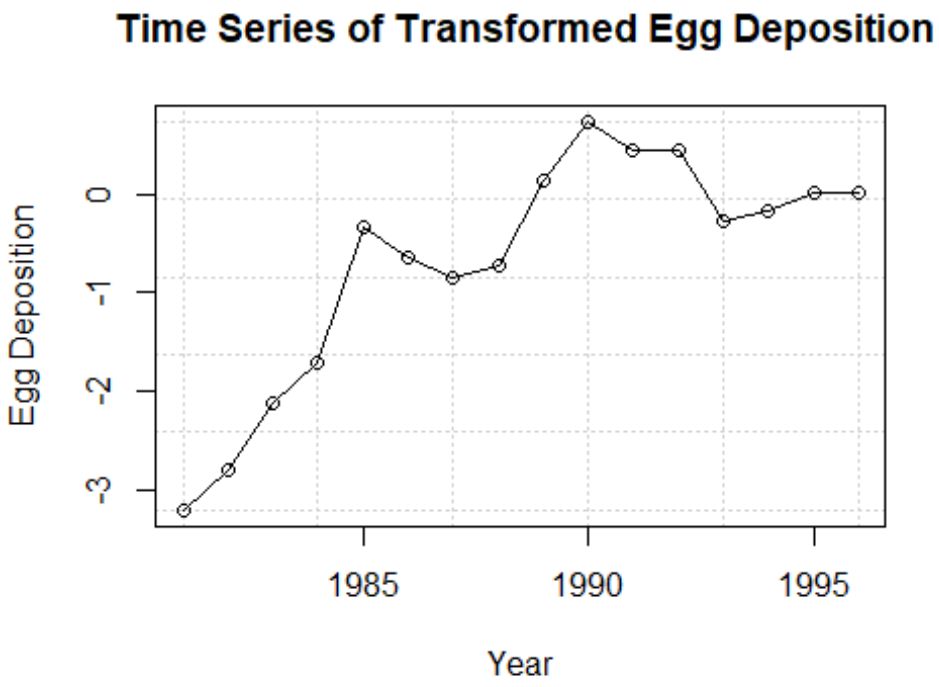
The ACF plot has pattern which has a large spike at lag 1 that decrease after a few lags

DATA MODELING

DATA TRANSFORMATION AND DIFFERENCE

Apply Log transformation for this series

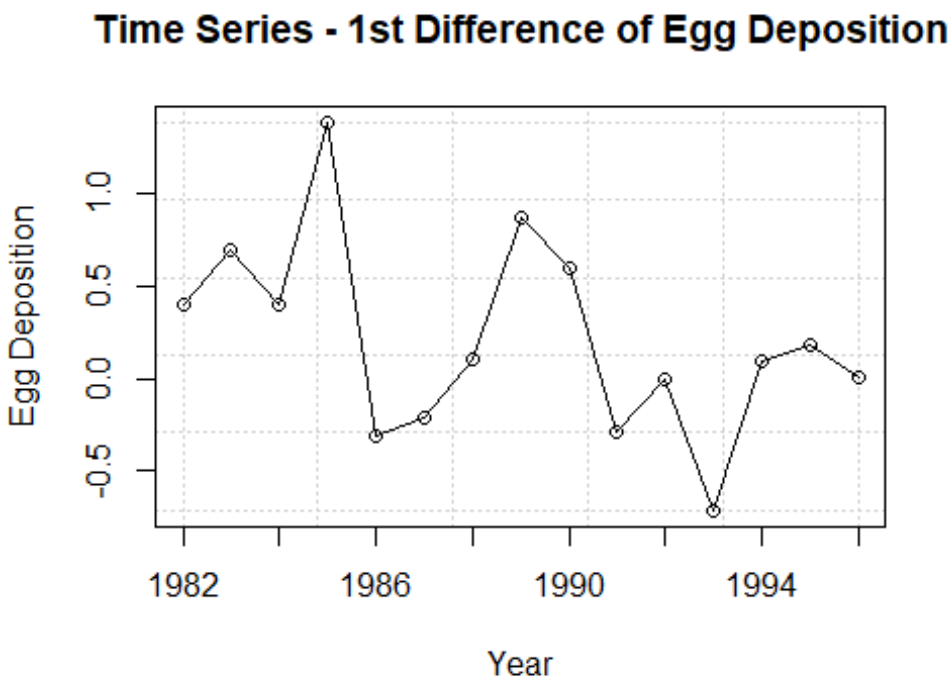
```
# Using Log transform
log.egg <- log(egg_ts)
plot(log.egg , type = 'o', xlab="Year", ylab= "Egg Deposition", main =" Time
Series of Transformed Egg Deposition", panel.first=grid())
```



This series still has a trend.

Applying the first difference for transformed data

```
diff.log.egg = diff(log.egg)
plot(diff.log.egg ,type='o',ylab='Egg Deposition', xlab= "Year", main = "Time
Series - 1st Difference of Egg Deposition",
     panel.first=grid())
```



Applying ADF unit-root test to test the existence of non-stationary with this series

```
ar(diff(diff.log.egg))
```

```
##
## Call:
## ar(x = diff(diff.log.egg))
##
## Coefficients:
##       1        2        3        4
## -0.7962  -0.5984  -0.7310  -0.4869
##
## Order selected 4  sigma^2 estimated as  0.3675
```

Order for this test is 4

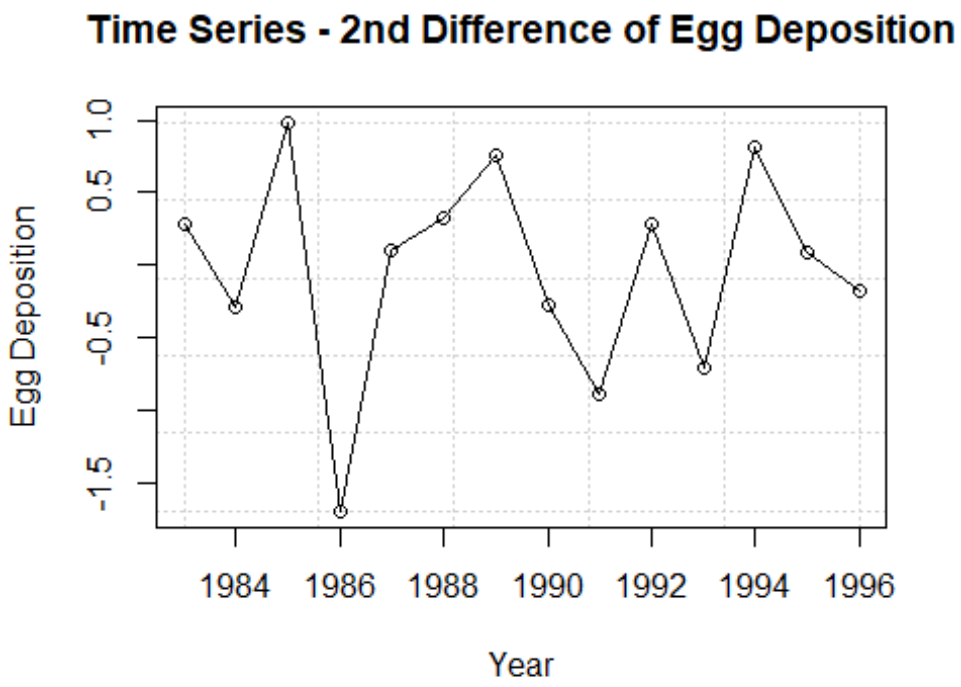
```
adfTest(diff.log.egg, lags = 4)
```

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 4
##   STATISTIC:
##     Dickey-Fuller: -1.2696
##   P VALUE:
```

```
##     0.2049
##
## Description:
##  Sun May 12 23:46:41 2019 by user: loanh
```

The p-value is 0.2049 larger than 5% level of significance. Therefore, this series is still non-stationary. Hence, we apply the second difference

```
diff2.log.egg <- diff(diff.log.egg, diference = 2)
plot(diff2.log.egg ,type='o',ylab='Egg Deposition', xlab= "Year", main =
"Time Series - 2nd Difference of Egg Deposition",
     panel.first=grid())
```



Applying ADF unit-root test for second difference

```
ar(diff(diff2.log.egg))

##
## Call:
## ar(x = diff(diff2.log.egg))
##
## Coefficients:
##       1
## -0.6514
##
## Order selected 1  sigma^2 estimated as  1.008
```

Order for this test is 1

```
adfTest(diff2.log.egg, lags = 1)
```

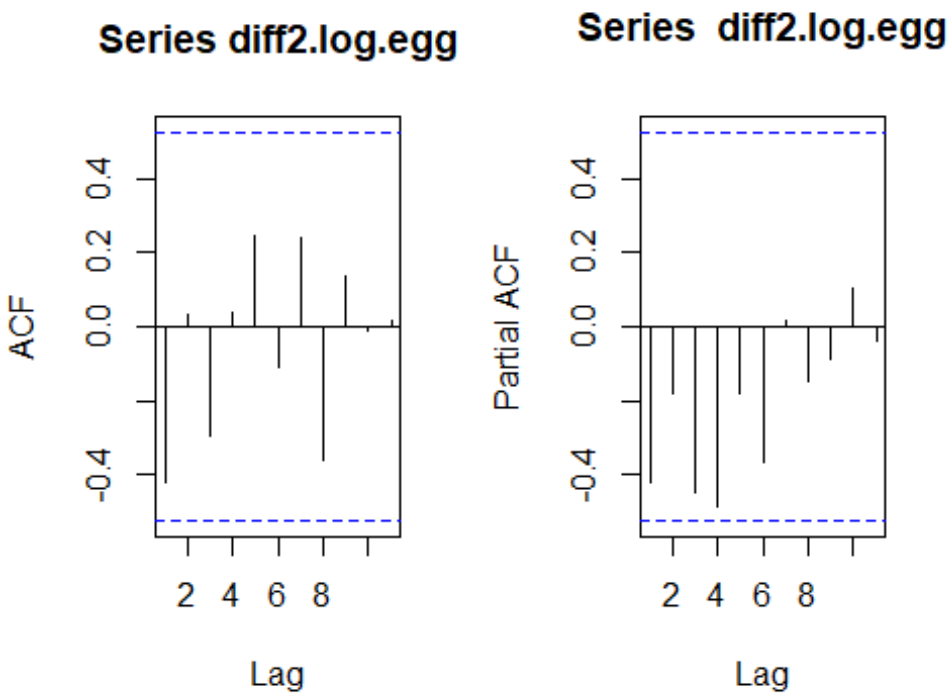
```
## Warning in adfTest(diff2.log.egg, lags = 1): p-value smaller than printed
## p-value

##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 1
##   STATISTIC:
##     Dickey-Fuller: -3.167
##   P VALUE:
##     0.01
##
## Description:
##  Sun May 12 23:46:41 2019 by user: loanh
```

p-value = 0.01

The p-value equals 0.01 less than 5% level of signigicance. Therefore, we can reject null hypothesis for non-stationary. It means that this series is stationary.

Displaying ACF and PACF plots

```
par(mfrow=c(1,2))
acf(diff2.log.egg)
pacf(diff2.log.egg)
```



There is no pattern in ACF and PACF plots.

Using EACF

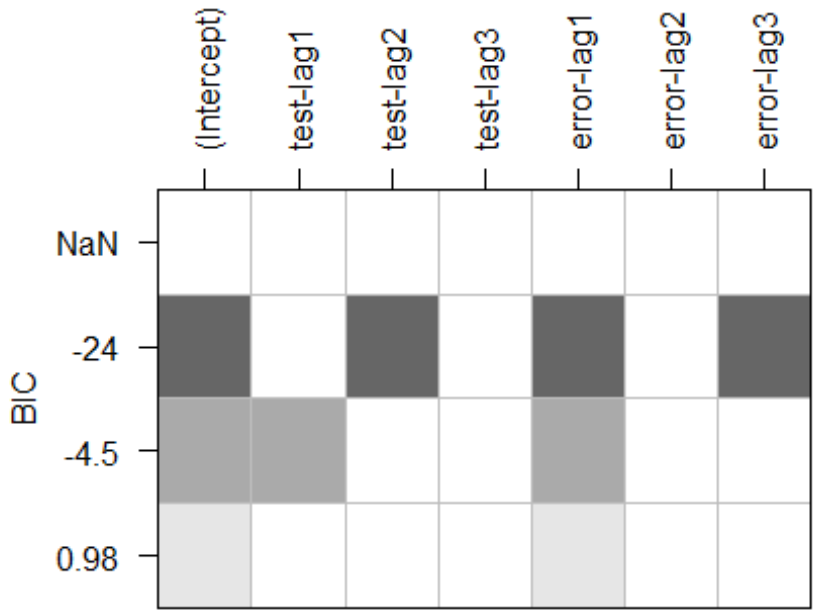
```
# Applying EACF
eacf(diff2.log.egg, ar.max = 3, ma.max = 3)
```

```
## AR/MA
##   0 1 2 3
## 0 o o o o
## 1 o o o o
## 2 o o o o
## 3 o o o o
```

From the above matrix, the possible models are ARIMA(0,2,1), ARIMA(1,2,0), ARIMA(1,2,1)

Using BIC table

```
# Using BIC table
res = armasubsets(y=diff2.log.egg,nar=3,nma=3,y.name='test',ar.method='ols')
plot(res)
```



From BIC table, possible models include ARIMA(2,2,1) , ARIMA(2,2,3) and ARIMA(1,2,1)

Hence the set of possible models will be {ARIMA(2,2,1), ARIMA(1,2,1), ARIMA(2,2,3), ARIMA(0,2,1), ARIMA(1,2,0), ARIMA(1,2,1)}

PARAMETER ESTIMATION

In order to estimate parameters of possible models, I use Maximum Likelihood method

ARIMA(0,2,1)

```
# Maximum likelihood estimates of the coefficients for ARIMA(0,2,1)
model.021 <- arima(egg_ts,order=c(0,2,1),method='ML')
model.021

##
## Call:
## arima(x = egg_ts, order = c(0, 2, 1), method = "ML")
##
## Coefficients:
##           ma1
##       -1.0000
## s.e.   0.2582
##
## sigma^2 estimated as 0.1841:  log likelihood = -9.37,  aic = 20.75
```

```
# Maximum likelihood estimates of the coefficients with significance tests
for ARIMA(0,2,1) model
coeftest(model.021)

##
## z test of coefficients:
##
##     Estimate Std. Error z value  Pr(>|z|)
## ma1 -1.00000    0.25823 -3.8725 0.0001077 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is less than 0.05. Therefore the estimated coefficients of ARIMA(0,2,1) model is significant

ARIMA(2,2,1)

```
# Maximum likelihood estimates of the coefficients for ARIMA(2,2,1)
model.221 <- arima(egg_ts, order = c(2,2,1), method = 'ML')
model.221

##
## Call:
## arima(x = egg_ts, order = c(2, 2, 1), method = "ML")
##
## Coefficients:
##          ar1      ar2      ma1
##       0.0711  -0.0106  -1.0000
## s.e.  0.2694   0.2595   0.2397
##
## sigma^2 estimated as 0.1846:  log likelihood = -9.34,  aic = 24.67

# Maximum likelihood estimates of the coefficients with significance tests
for ARIMA(2,2,1) model
coeftest(model.221)

##
## z test of coefficients:
##
##      Estimate Std. Error z value  Pr(>|z|)
## ar1  0.071133   0.269440  0.2640    0.7918
## ar2 -0.010595   0.259510 -0.0408    0.9674
## ma1 -0.999992   0.239725 -4.1714 3.027e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated ar1 and ar2 coefficients are larger than 0.05. This means that they are are not significant

ARIMA(1,2,1)

```r
# Maximum likelihood estimates of the coefficients for ARIMA(1,2,1)
model.121 <- arima(egg_ts, order = c(1,2,1), method = 'ML')
model.121
```

```
##
## Call:
## arima(x = egg_ts, order = c(1, 2, 1), method = "ML")
##
## Coefficients:
##          ar1      ma1
##       0.0718  -1.0000
## s.e.  0.2693   0.2369
##
## sigma^2 estimated as 0.1849:  log likelihood = -9.34,  aic = 22.67
```

```r
# Maximum likelihood estimates of the coefficients with significance tests
for ARIMA(1,2,1) model
coeftest(model.121)
```

```
##
## z test of coefficients:
##
##      Estimate Std. Error z value  Pr(>|z|)
## ar1  0.071764   0.269251  0.2665    0.7898
## ma1 -0.999999   0.236872 -4.2217 2.425e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

AR1 is insignificant in model ARIMA(1,2,1)

ARIMA(2,2,3)

```r
model.223 <- arima(egg_ts, order = c(2,2,3), method = 'ML')
coeftest(model.223)
```

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value Pr(>|z|)
## ar1 -0.58579    0.52391 -1.1181  0.26352
## ar2 -0.38546    0.29026 -1.3280  0.18418
## ma1 -0.13925    0.48815 -0.2853  0.77545
## ma2  0.13920    0.49983  0.2785  0.78063
## ma3 -0.99992    0.46140 -2.1671  0.03022 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

AR1, AR2, MA1 and MA2 are insignificant

ARIMA(1,2,0)

```
model.120 <- arima(egg_ts, order = c(1,2,0), method = 'ML')
coeftest(model.120)

##
## z test of coefficients:
##
##     Estimate Std. Error z value Pr(>|z|)
## ar1 -0.42966    0.22743 -1.8892  0.05886 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

AR1 is not significant as well in model ARIMA(1,2,0)

APPLYING AIC and BIC to decide the best model within the subset of possible models

__ AIC__

```
AIC(model.021)

## [1] 22.74602

AIC(model.120)

## [1] 26.57611

AIC(model.121)

## [1] 24.67428

AIC(model.221)

## [1] 26.67261

AIC(model.223)

## [1] 27.68002
```

BIC

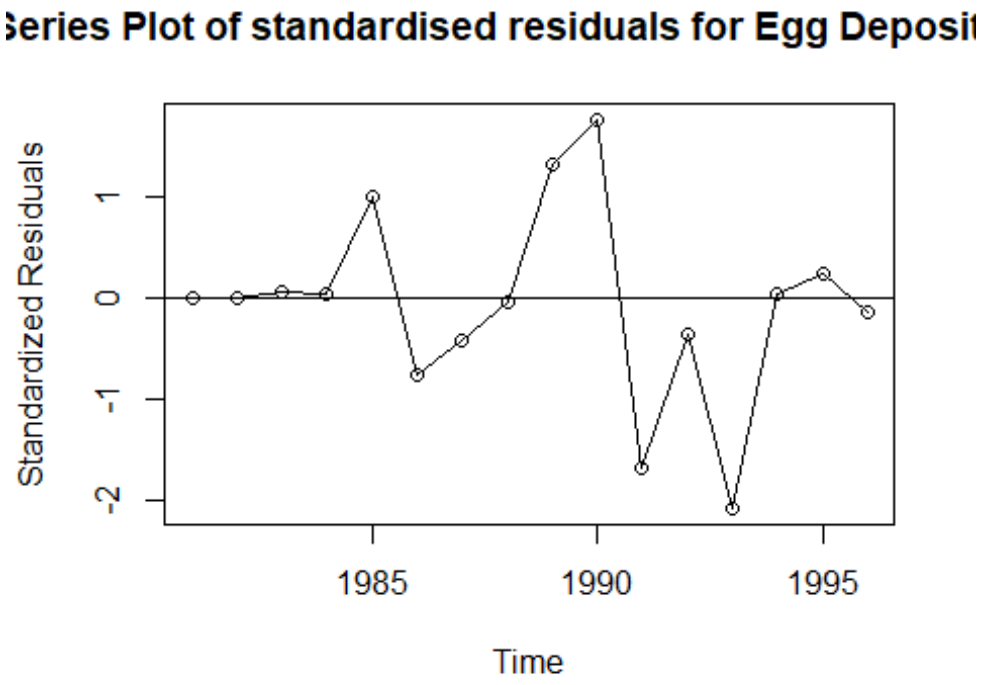
```
AIC(model.021, k = log(28))

## [1] 25.41043

AIC(model.121, k = log(28))

## [1] 28.67089

AIC(model.221, k = log(28))

## [1] 32.00143

AIC(model.223, k = log(28))

## [1] 35.67324
```

The smallest AIC is model ARIMA(0,2,1) with 22.74602. According to BIC, model ARIMA(0,2,1) with value 25.41043 is the best. Therefore ARIMA(0,2,1) is the best model.

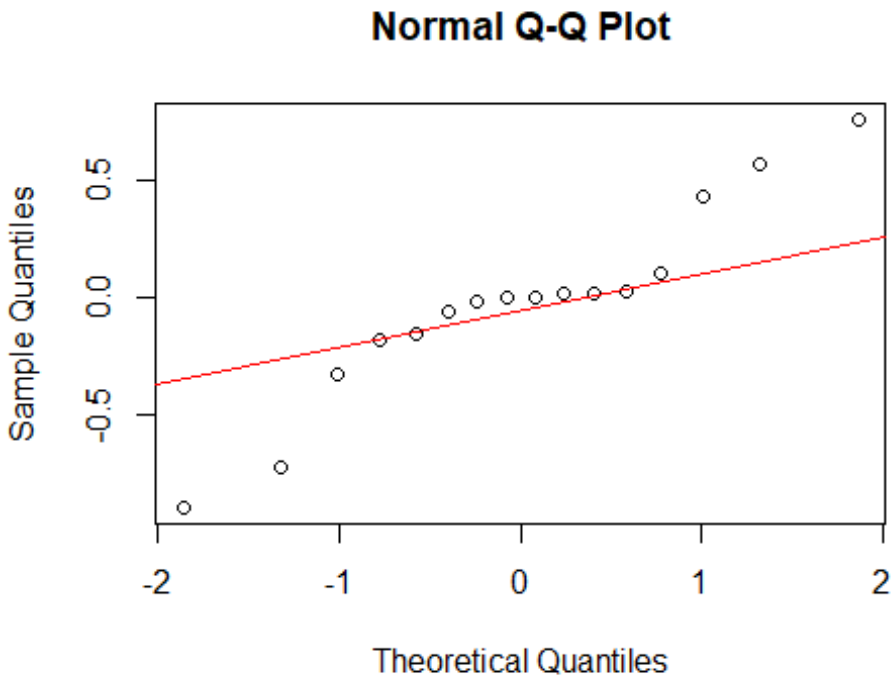
MODEL DIAGNOTICS

Normality Testing

```
plot(rstandard(model.021), type= 'o', ylab="Standardized Residuals", main =
"Time Series Plot of standardised residuals for Egg Deposition Series")
abline(h=0)
```




```
# Using QQplot
qqnorm(residuals(model.021))
qqline(residuals(model.021), col = 2)
```



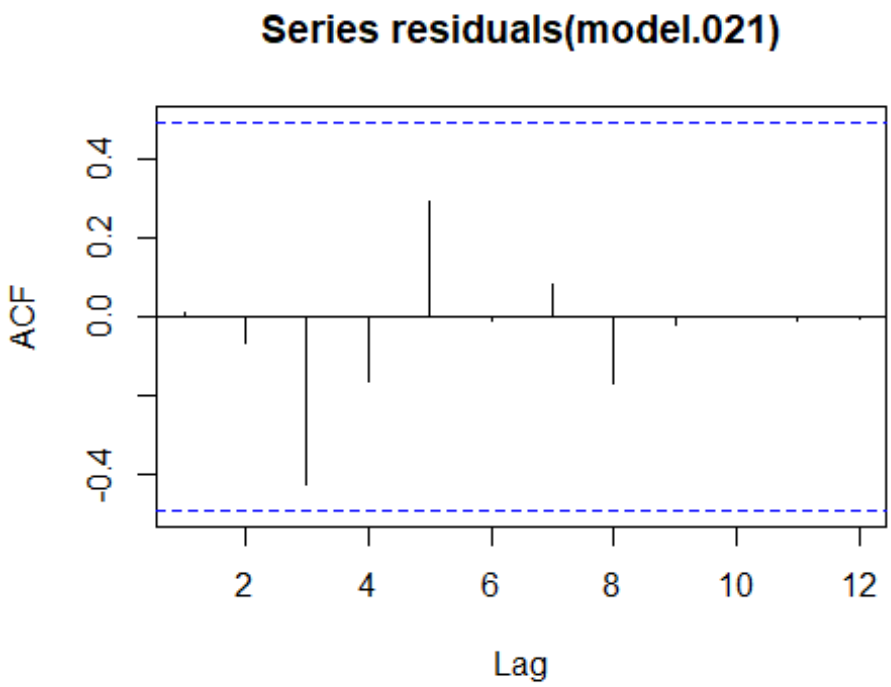
Most of values lie on the red line. However, there are still some outliers at the top right and left bottom

```
# Using Shapiro test
shapiro.test(diff2.log.egg)

##
##  Shapiro-Wilk normality test
##
## data:  diff2.log.egg
## W = 0.94806, p-value = 0.531
```

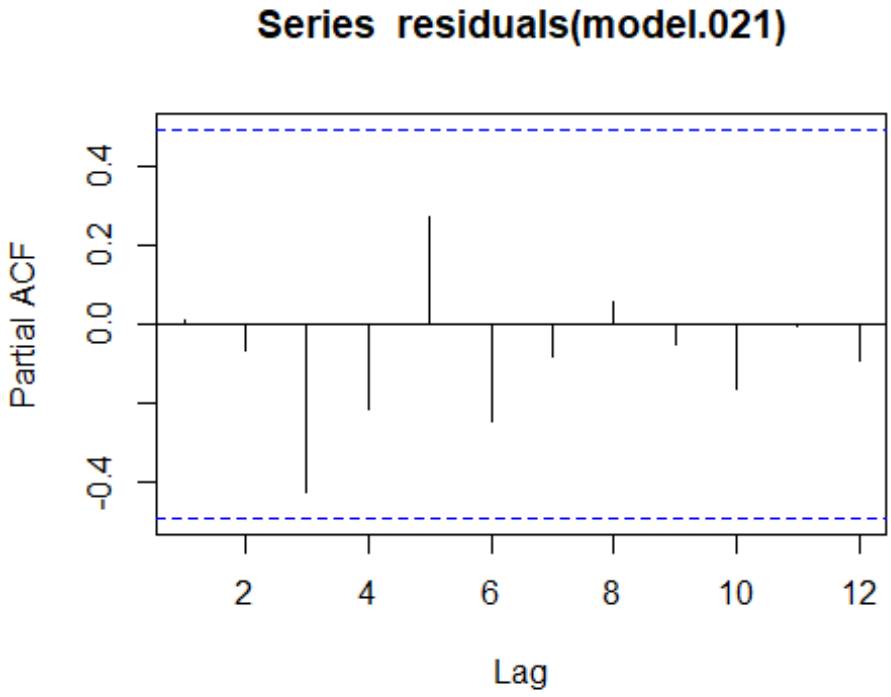
The p-value is 0.531 > 0.05. Therefore, we cannot reject H0 hypothesis which is normality error assumption. This implies that normality error assumption is not violated

AutoCorrelation of residuals testing

```
acf(residuals(model.021))
```




```
pacf(residuals(model.021))
```



From ACF and PACF plots of residuals, we can conclude that the residuals constitutes a white noise series as there is no highly significant correlation.

LJUNG - BOX TEST

```
Box.test(residuals(model.021), lag = length(model.021$residuals) -1, type =
"Box-Pierce", fitdf = 0)
```

```
##
##  Box-Pierce test
##
## data:  residuals(model.021)
## X-squared = 5.3017, df = 15, p-value = 0.9892
```

The p-value in LjungBox test is 0.9892 > 0.05. Therefore, we cannot reject null hypothesis that the error terms are uncorrelated.

In conclusion, there is no problem in the residuals of ARIMA(0,2,1) model.

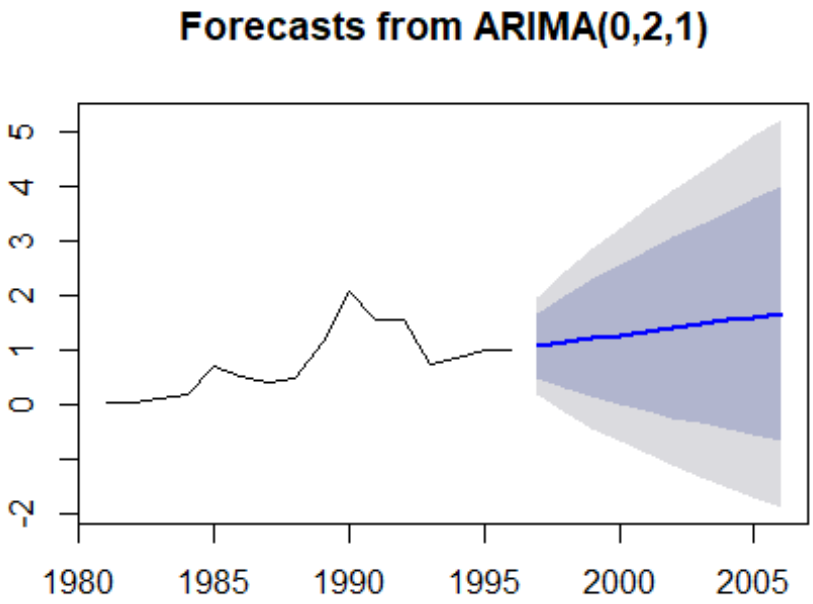
FORECASTING

```
# Create a matrix of the covariance for prediction
xreg=data.frame (constant=seq(egg_ts))
n = length(egg_ts)
n.ahead = 5
newxreq = data.frame(constant = (n+1): (n+n.ahead))
# Predict Egg Deposition for next 5 years
predict(model.021, n.ahead = 5, newxreg = NULL, se.fit = TRUE)
```

```
## $pred
## Time Series:
## Start = 1997
## End = 2001
## Frequency = 1
## [1] 1.089693 1.155287 1.220880 1.286473 1.352067
##
## $se
## Time Series:
## Start = 1997
## End = 2001
## Frequency = 1
## [1] 0.4431333 0.6459723 0.8140879 0.9657867 1.1078333
```

The predict values of Egg depositions in next five years are {1.089693 ; 1.155287; 1.220880; 1.286473 ; 1.352067}

```
#Plot forecasts over time series
fit = Arima(egg_ts,c(0,2,1))
plot(forecast(fit,h=10))
```


Forecasts from ARIMA(0,2,1)
-2
0
1
2
3
4
5
1980
1985
1990
1995
2000
2005