MATH1318 Assignment 2

Jeevan Hemmannu Tharanatha(s3755598)

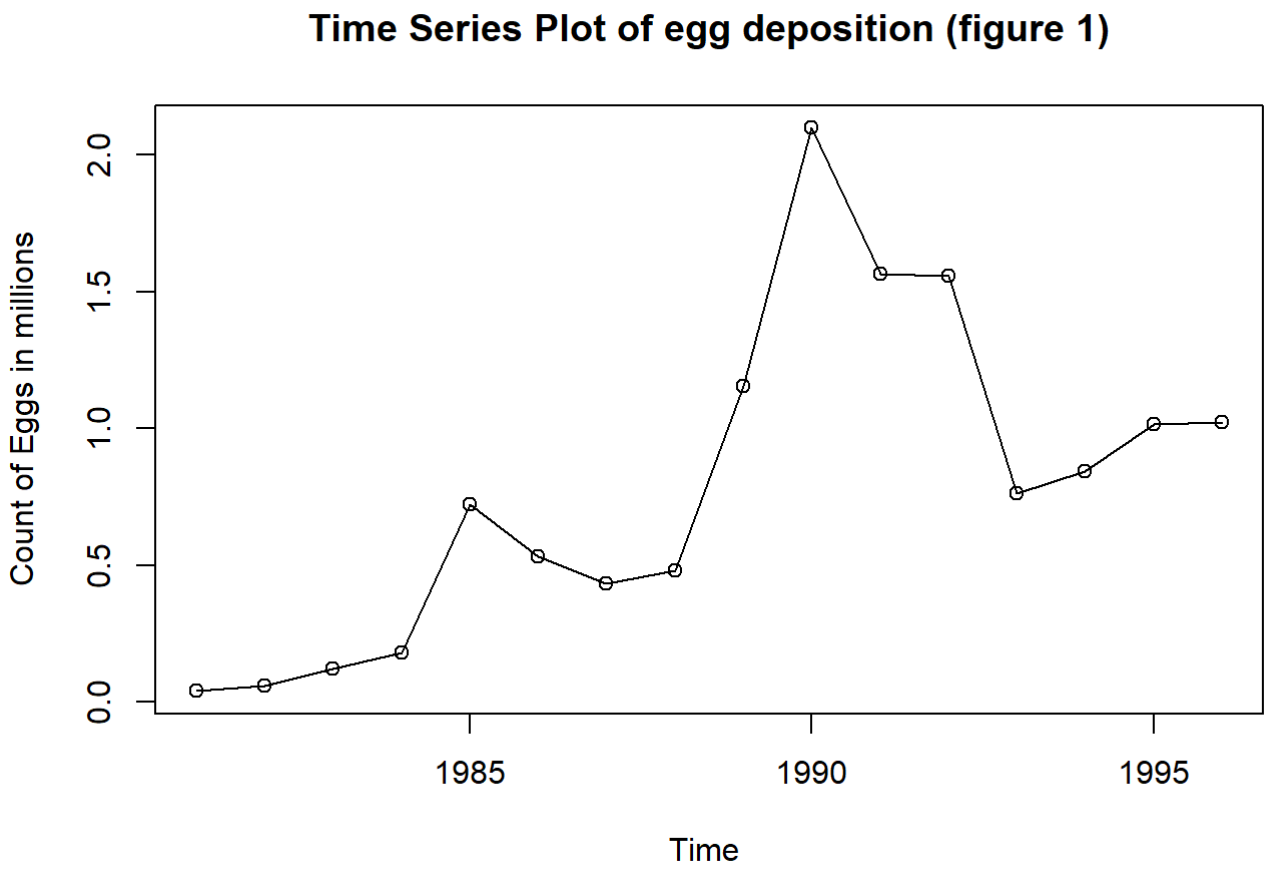
10 May 2020

Introduction

Coregonus hoyi, otherwise called the bloater, is a species or type of freshwater whitefish in the family Salmonidae. It is a shimmering shaded herring-like fish, 25.5 centimeters (10.0 in) long. It is found in the greater part of the Great Lakes and in Lake Nipigon, and inhabits underwater slopes.

The aim of this experiment is to analyse how the egg deposition of age-3 lake huron Bloaters has changed from 1981 to 1996 and to predict the changes in egg depositions for next 5 years using best model for the given data.

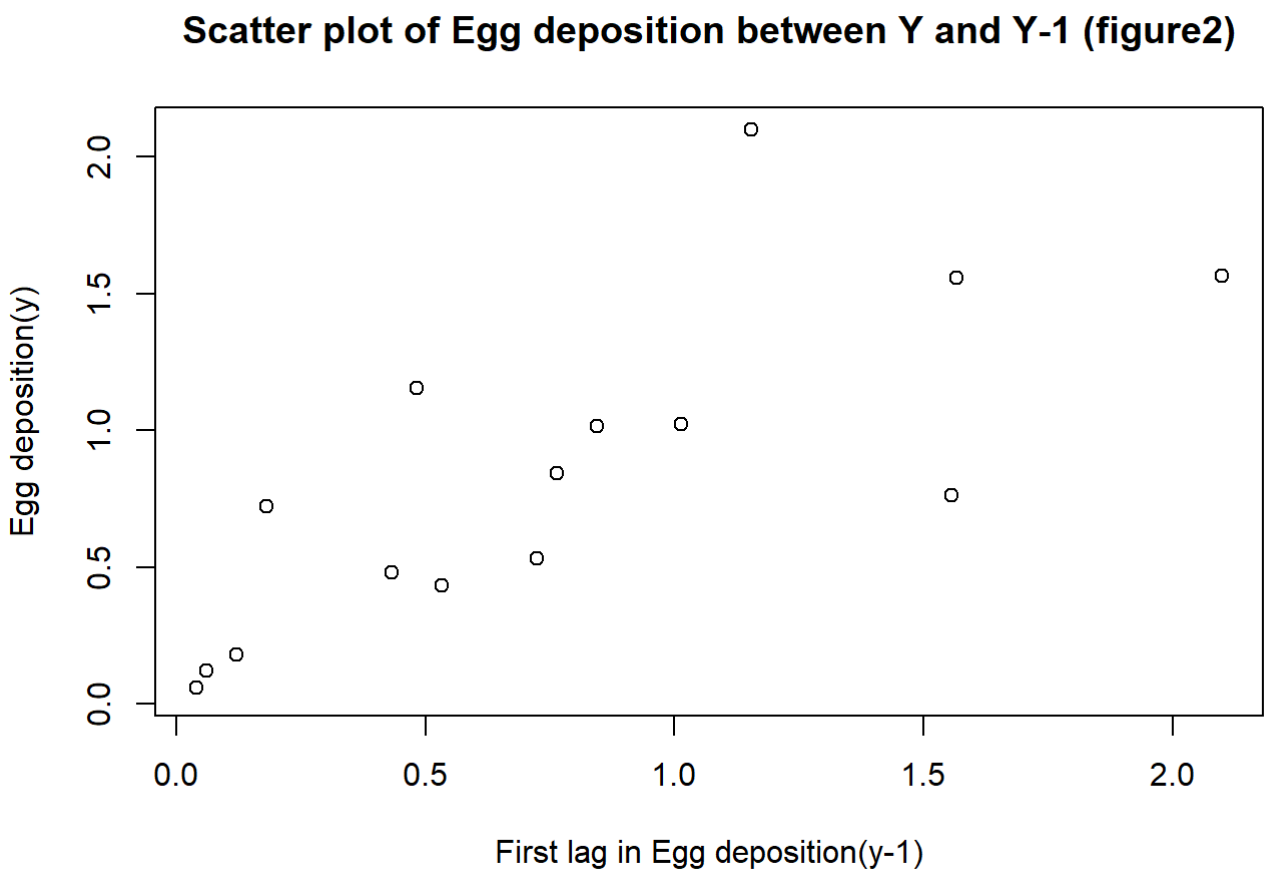
The data for our analysis is taken from FSAdata library which has the Egg count in millions from 1981 to 1996.



Lets talk about 5 charasteristics of the above Time series graph (figure 1)

1. **Trend** : It is obvious from the above graph that there is a slight upwords trend . This suggests that graph is non-stationary.
2. **Changing variation** : We can see changing variation among the circles over time, we can say that there is variance.
3. **Seasonality** : There are no repeating patterns. Since this is also a yearly observation, we can say that there is no seasonality.

4. **Autocorrelation structure** : We can see many observations hanging around. Many data points are following the trend from previous points. Hence we can say there is autocorrelation.
5. **Intervention point** : No intervention point since there is no sudden change in the patterns in the graph.

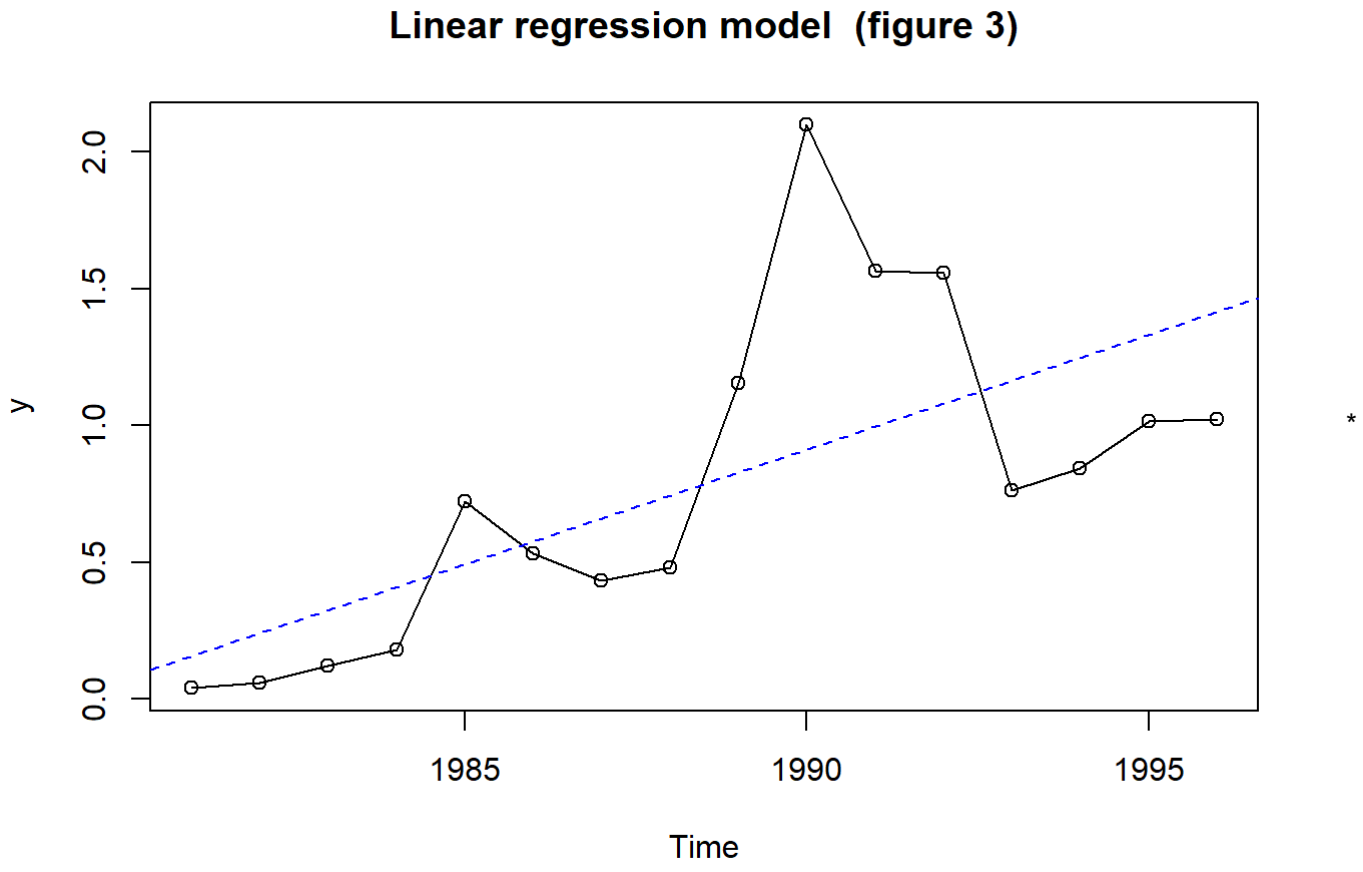



```
## [1] 0.7445657
```

- Strong correlation can be seen from scatter plot between egg depositions and its first lag (figure 2).
- This is proved by covariance test with the result of 0.74 which suggest positive autocorrelation between the same.

Building the Linear Model

```
##
## Call:
## lm(formula = eggs ~ time(eggs))
##
## 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(eggs)     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
```

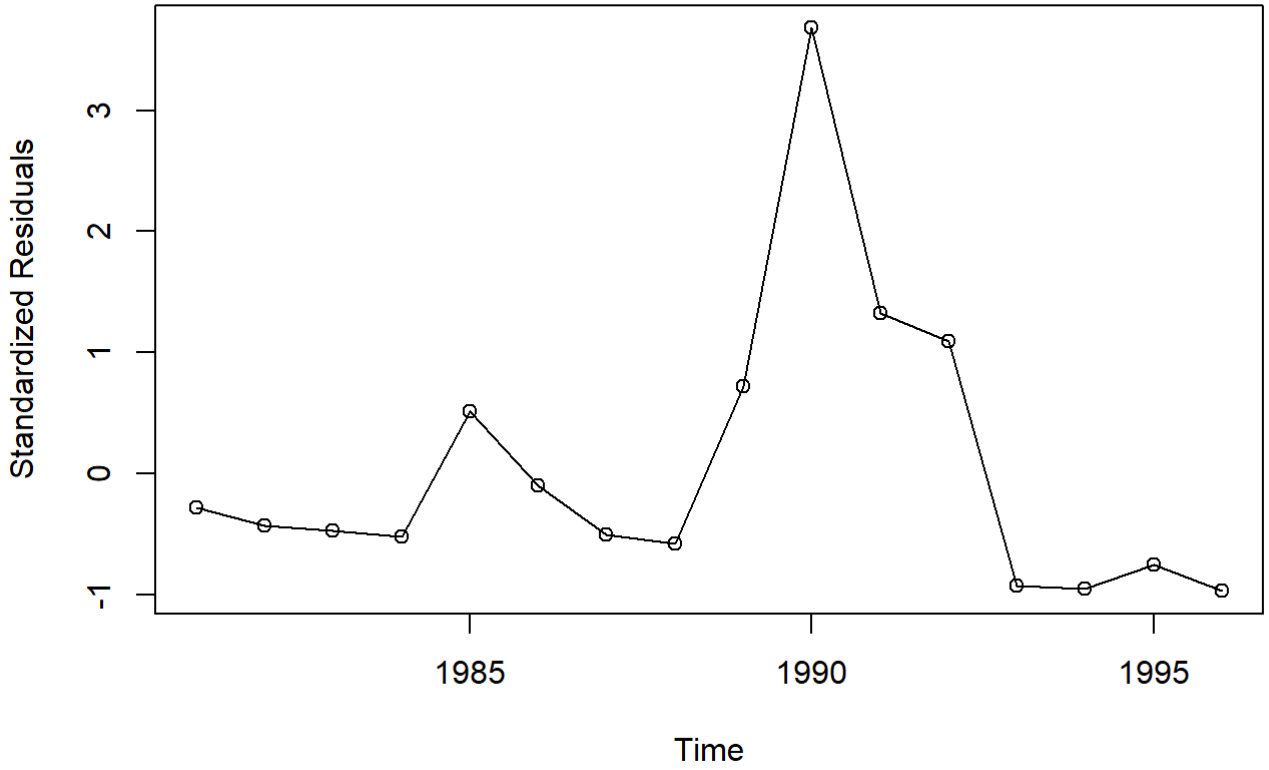


The plot(figure 3) shows the original time series graph with linear regression line fitted to it. We can see there are lot of data points which are away from our linear regression line.

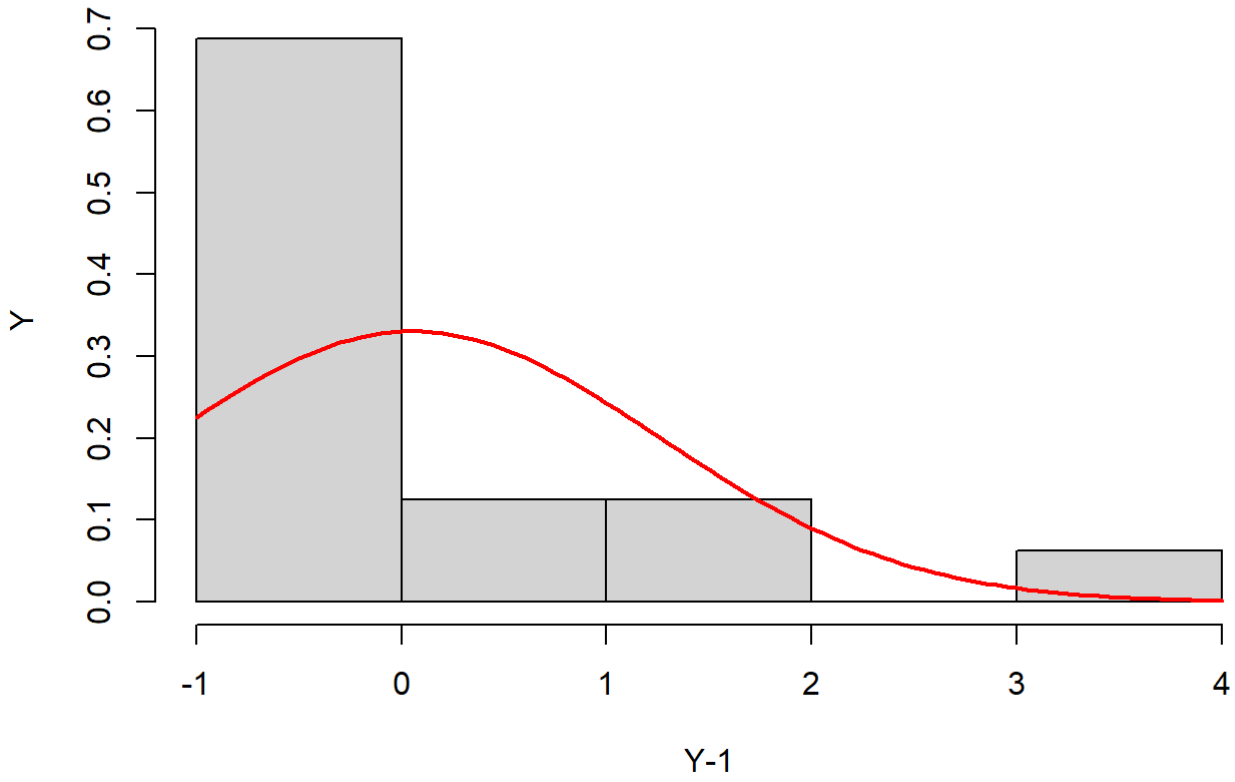
- p estimation of co-efficients is essentially lesser than 0.05. This suggest it is a great idea to consider this co-proficient.
- R square value is 0.4469 which suggest that 44.69% of the variation in the time series is explained by the linear model but analysing the other models and comparing their R square values gives the best model to use for this case.

Analysis of residuals for Linear regression model

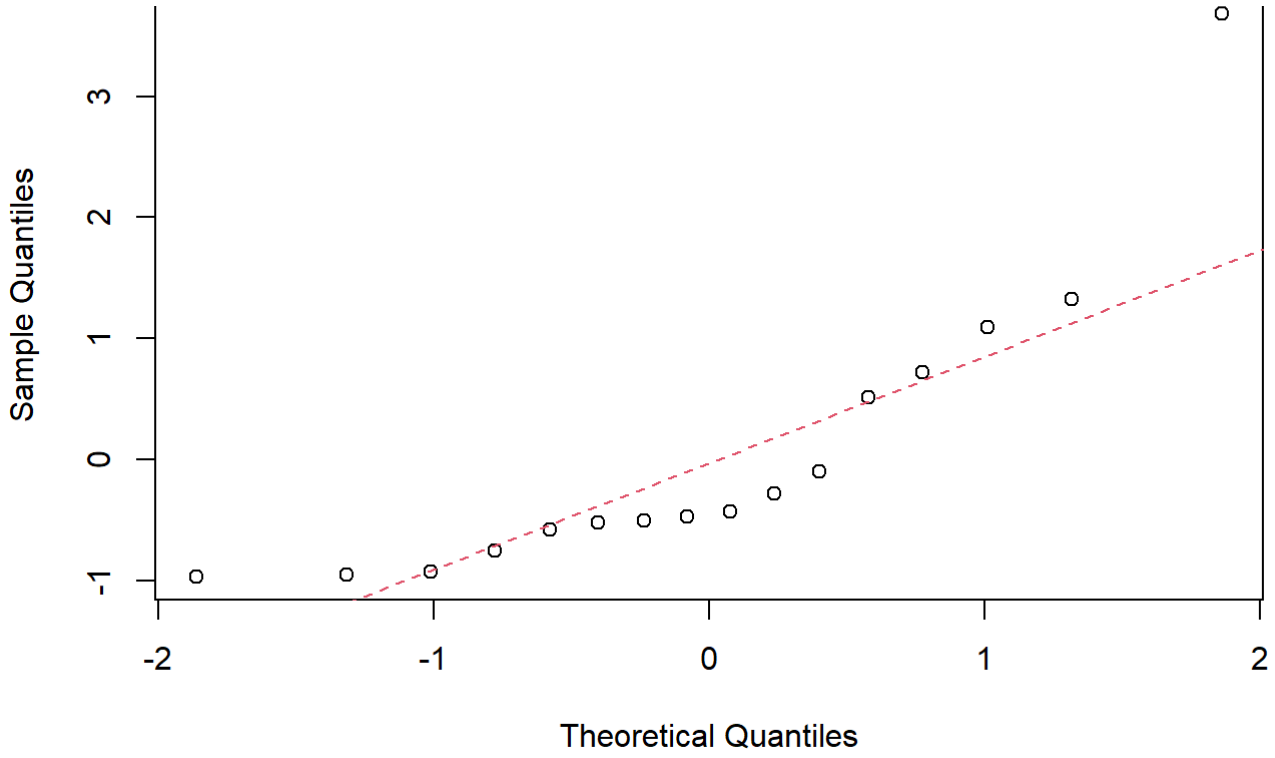
Residual VS fitted trend plot (figure 4)



normal curve over histogram figure(5)



Normal Q-Q Plot (figure 6)




```
##
##  Shapiro-Wilk normality test
##
## data:  res.model.eggs.ln
## W = 0.7726, p-value = 0.001205
```

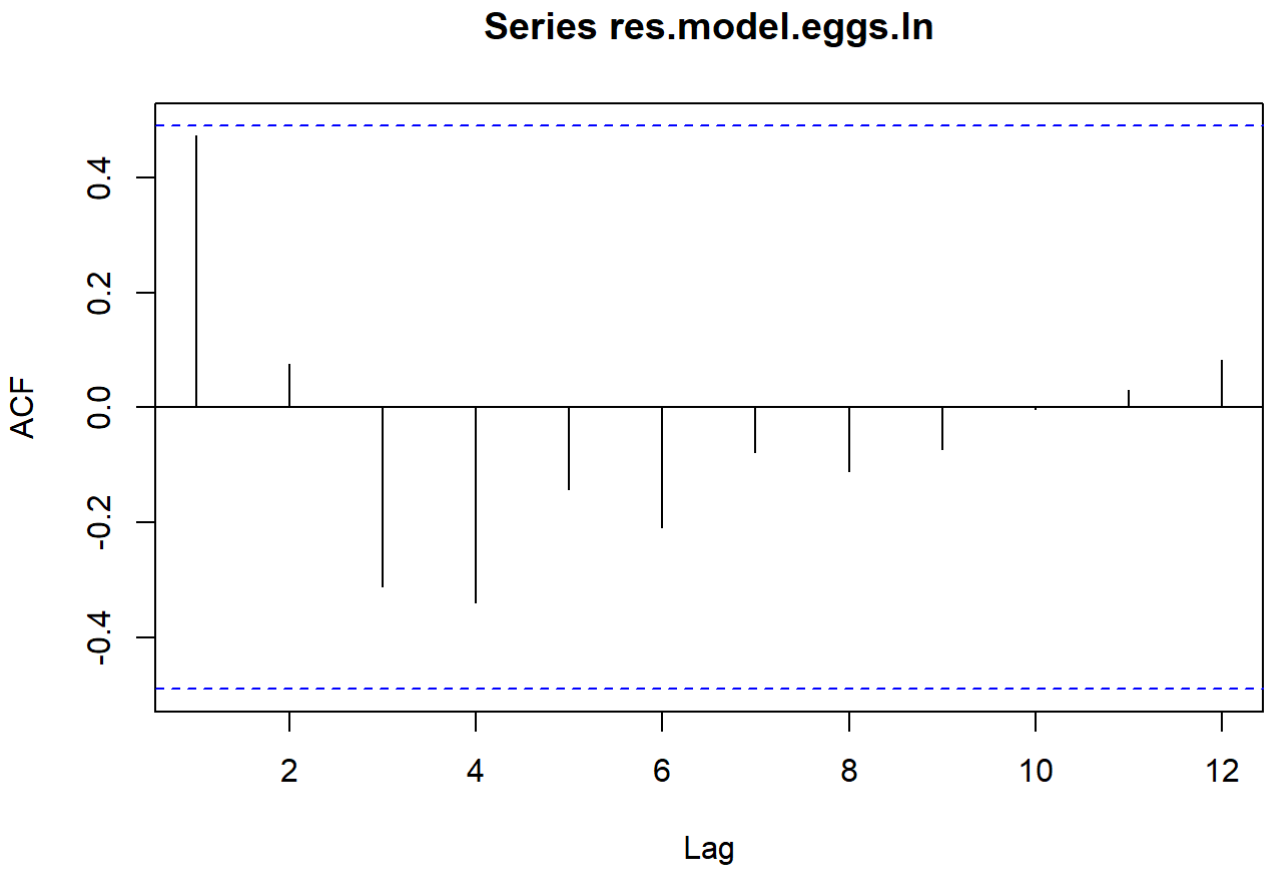
```
## $pvalue
## [1] 0.154
##
## $observed.runs
## [1] 5
##
## $expected.runs
## [1] 7.875
##
## $n1
## [1] 11
##
## $n2
## [1] 5
##
## $k
## [1] 0
```

Examination of residuals are significant before considering any model for predections. For a model to be viewed as acceptable, its residuals should be a white noise with normal distribution. Investigation of the residuals are done underneath:

1. **Residual VS fitted pattern plot** :This plot(figure 4) shows that data points are not scattered around. Additionally, variance is not constant among the residuals.
2. **Histogram of residuals** :From histogram of residuals(figure 5) it is clear that residuals are not normally distributed.

3. **QQ Plots** : In QQ plots(figure 6) of residuals we can see that data points are not at all alligned to the red line. This says that residuals are not normally distributed. This can be ensured in shapiro test.
4. **Shapiro Test** : With the test result of 0.001 Shapiro test rejects the Null hypothesis and concludes that residuals are not normally distributed.
5. **Independence test** : p value for this test is > 0.05 which propose that the residuals are independent.

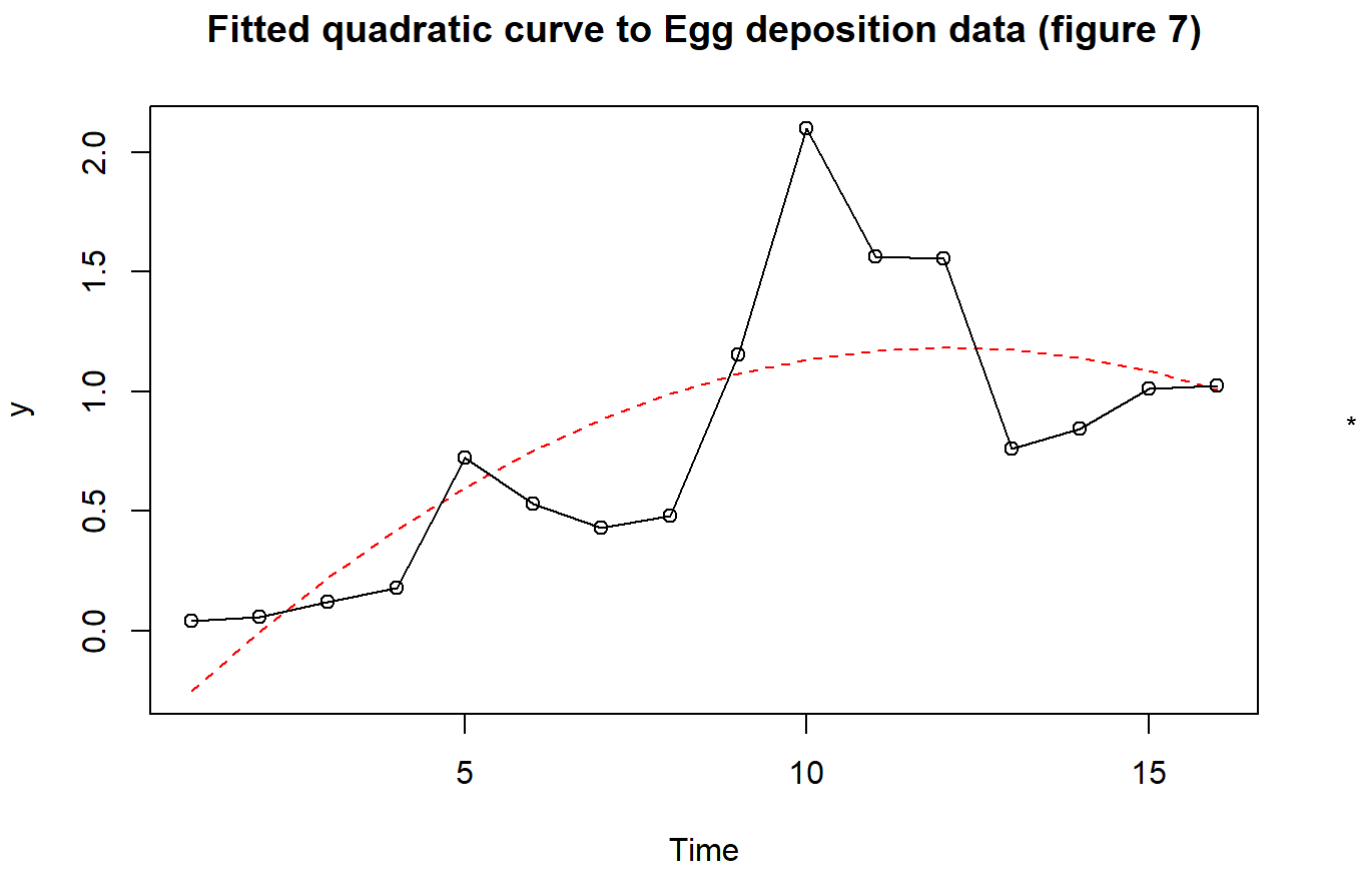
From all the above tests, it is clear that residuals of linear regression model is not white noise with normal distribution. This can likewise be seen from ACF chart.



- For the residuals to be white noise, all the lags should be below dotted blue lines in the acf plot. From the above plot we can see that one lag almost touches the blue dotted line. This is not significantly below blue dotted line. Because of this we can see that it lacks the white noise quality.
- Residual analysis of the linear regression model makes us to doubt about this model.

Building a Quadratic Model

```
##
## Call:
## lm(formula = eggs ~ 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
```

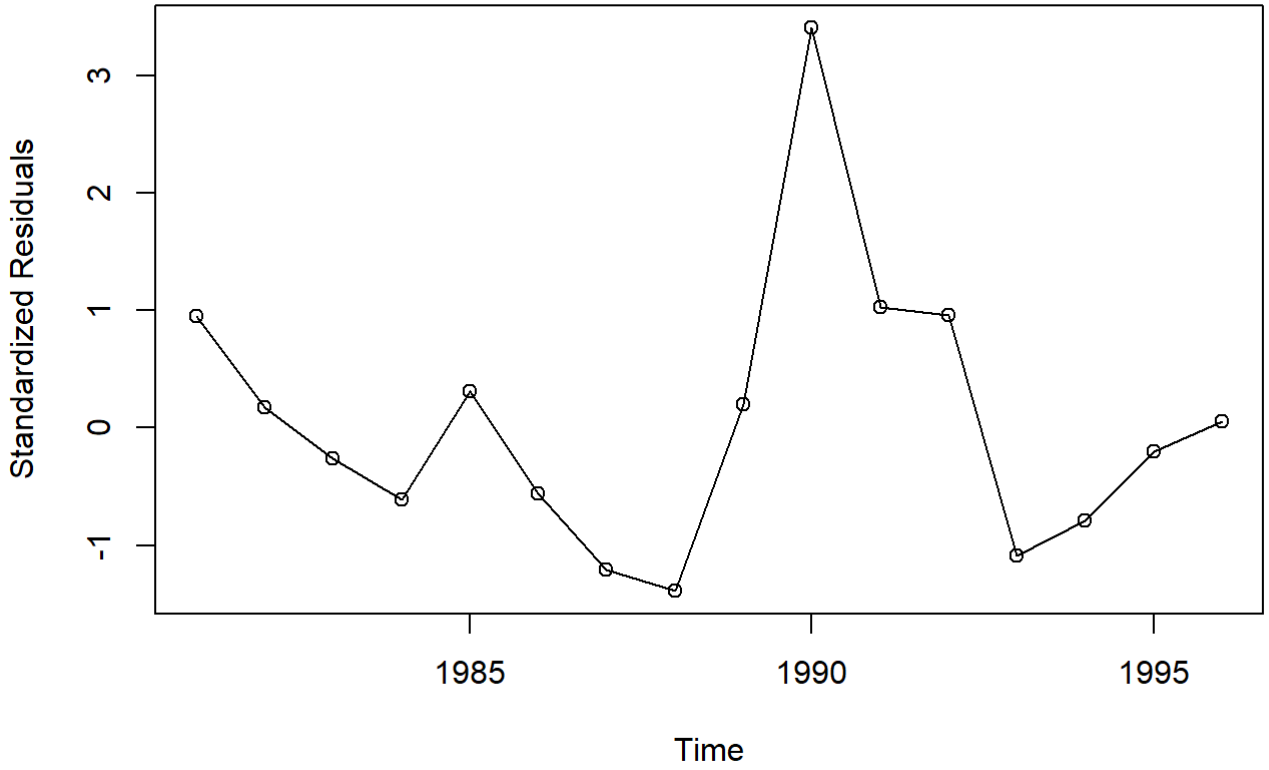


The plot(figure 7) shows the original time series graph with quadratic regression line fitted to it. We can see that there are lot of data points which are away from our quadratic regression line.

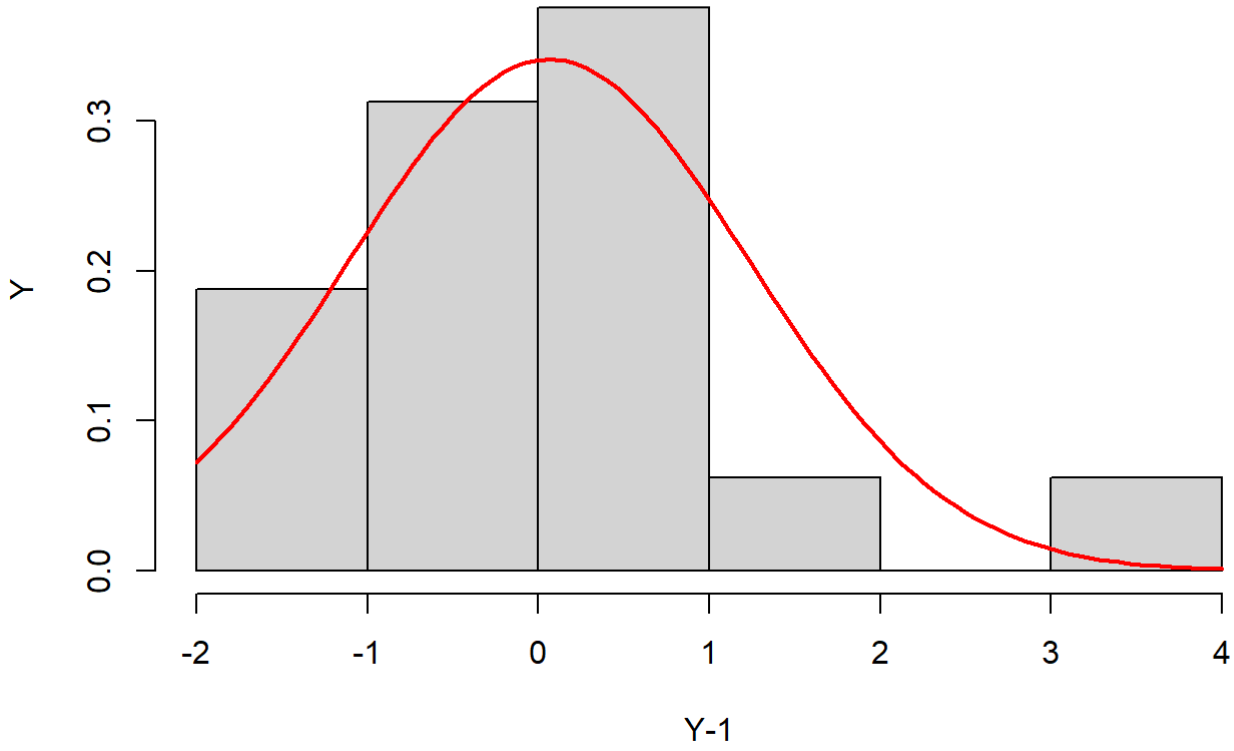
- p values of both co-efficients are just smaller than 0.05. So, it is good to consider both t and t2 co-efficient.
- R square value of the model is 0.5932 which suggest that 59.32% of the variation in the time series is explained by the quadratic model. This is greater than linear model which suggest there is an improvement.

Analysis of residuals for Quadratic regression model

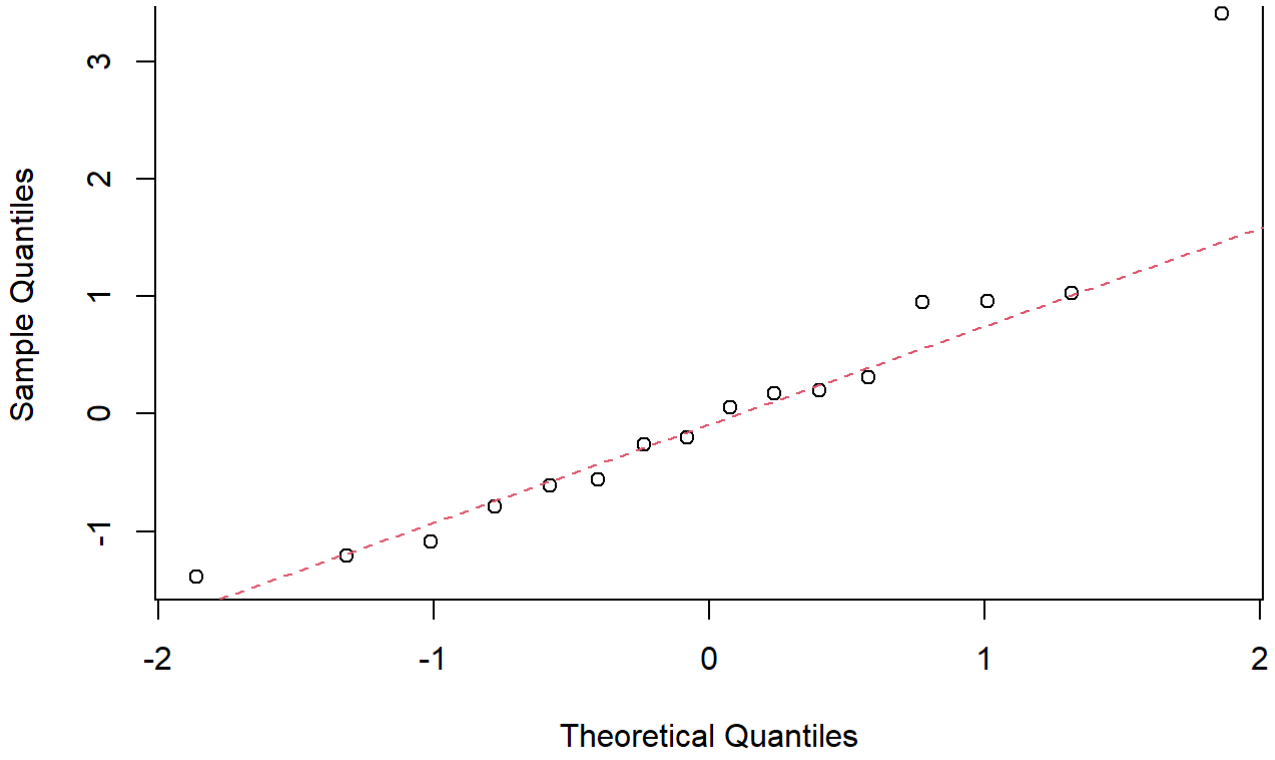
Residual VS fitted trend plot (figure 8)



normal curve over histogram (figure 9)



Normal Q-Q Plot (figure 10)




```
##
##  Shapiro-Wilk normality test
##
## data:  res.model.eggs.qa
## W = 0.87948, p-value = 0.03809
```

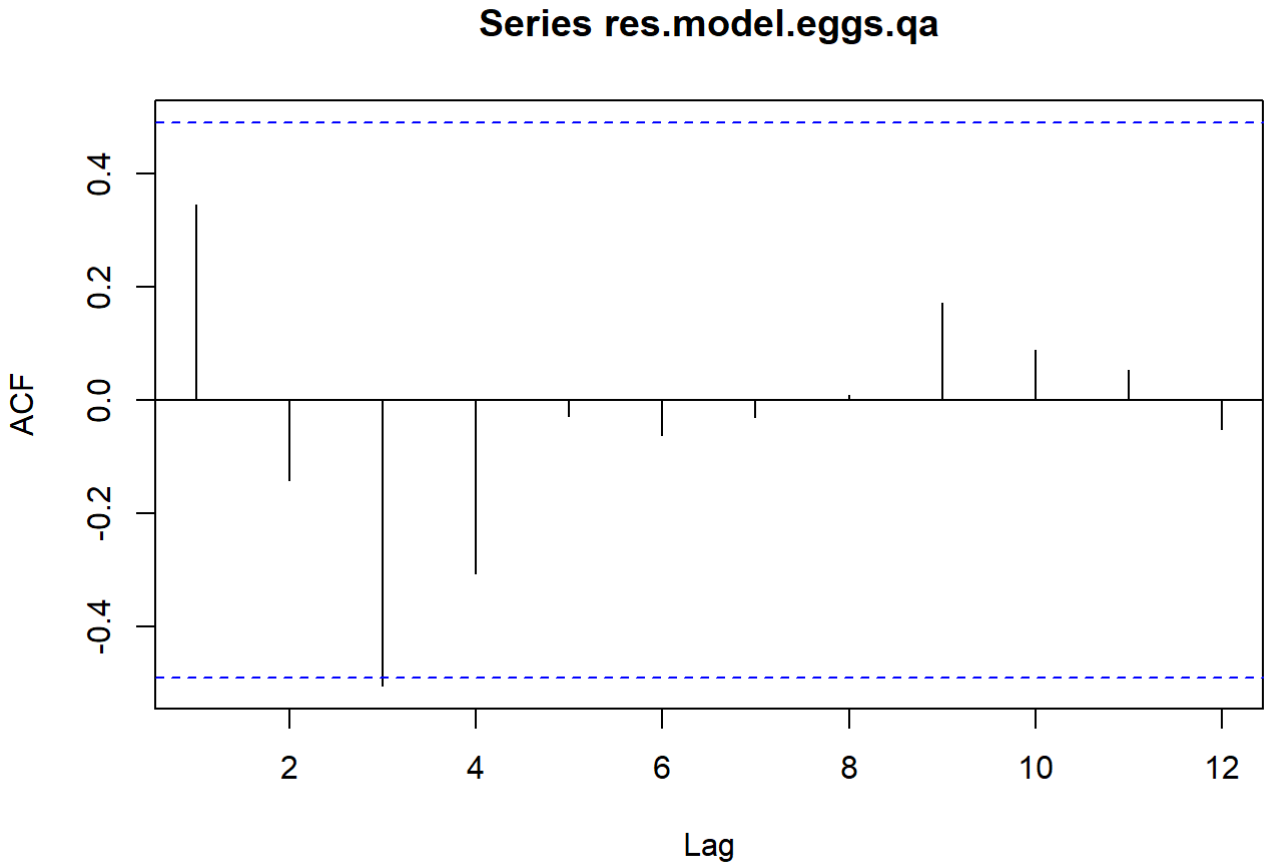
```
## $pvalue
## [1] 0.154
##
## $observed.runs
## [1] 5
##
## $expected.runs
## [1] 7.875
##
## $n1
## [1] 11
##
## $n2
## [1] 5
##
## $k
## [1] 0
```

Examination of residuals are significant before considering any model for predections. For a model to be viewed as acceptable, its residuals should be a white noise with normal distribution. Investigation of the residuals for quadratic linear model are done underneath:

1. **Residual VS fitted pattern plot** :This plot(figure 8) shows that data points are not scattered around. Additionally, variance is not constant among the residuals.
2. **Histogram of residuals** :From histogram of residuals(figure 9) it is clear that residuals are not normally distributed.

3. **QQ Plots** : In QQ plots of residuals (figure 10) we can see that data points are somewhat alligned to the red line. This says that residuals are not normally distributed. This can be ensured in shapiro test.
4. **Shapiro Test** : With the test result of 0.03 Shapiro test rejects the Null hypothesis and concludes that residuals are not normally distributed.
5. **Independence test** : p value for this test is > 0.05 which propose that the residuals are independent.

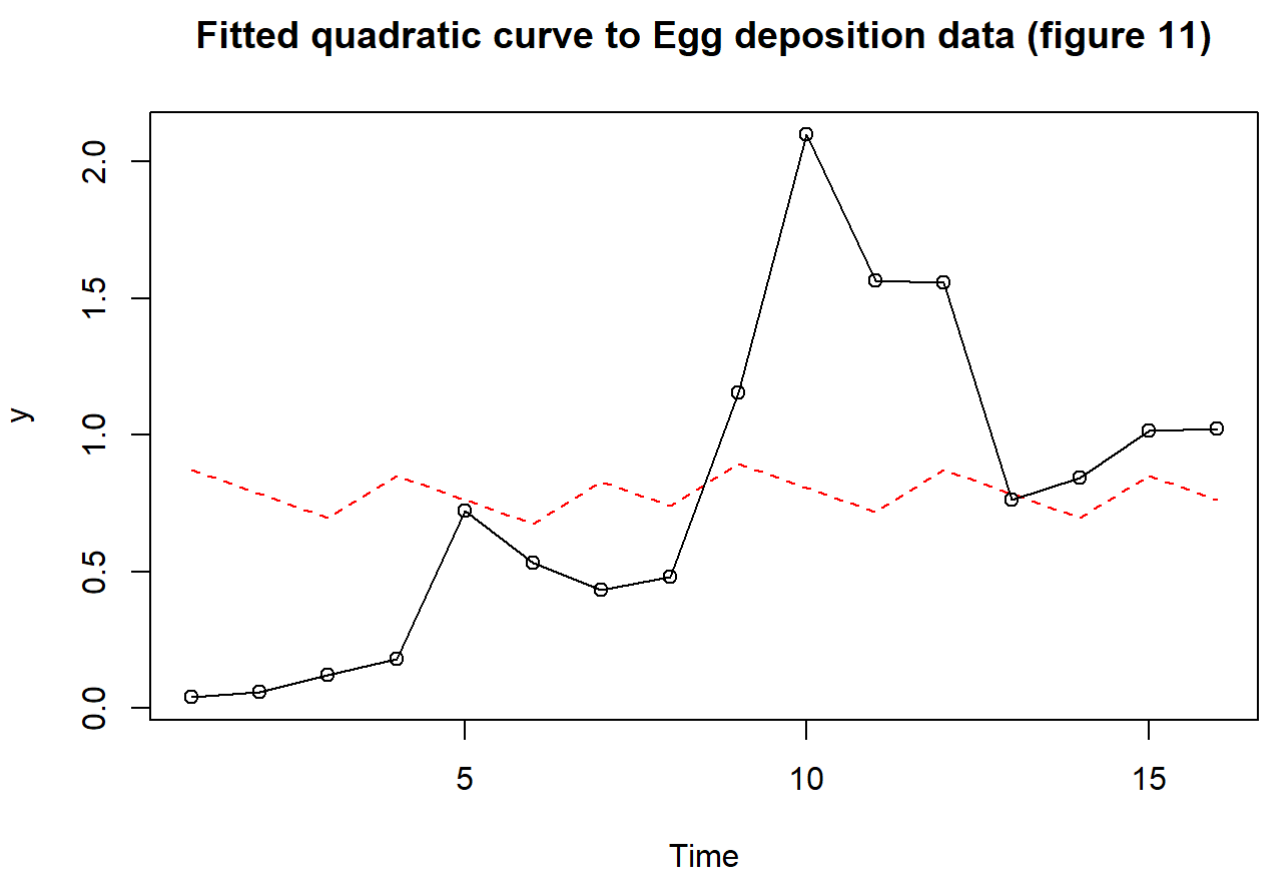
From all the above tests, it is clear that residuals of quadratic regression model is not white noise with normal distribution. This can likewise be seen from ACF chart.



- For the residuals to be white noise, all the lags should be below dotted blue lines in the acf plot. From the above plot we can see that one lag crosses blue dotted line. Because of this we can see that it lacks the white noise quality.
- Residual analysis of the quadratic regression model makes us to doubt about this model.

Building a Harmonic Model

```
##
## Call:
## lm(formula = eggs ~ har.)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.83161 -0.44101 -0.03012  0.26007  1.29195
##
## Coefficients: (1 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.241e-01  2.087e-01   3.469  0.00376 **
## har.cos(2*pi*t)         NA         NA      NA       NA
## har.sin(2*pi*t) -1.339e+11  3.015e+11  -0.444  0.66369
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.614 on 14 degrees of freedom
## Multiple R-squared:  0.0139, Adjusted R-squared:  -0.05654
## F-statistic: 0.1973 on 1 and 14 DF,  p-value: 0.6637
```

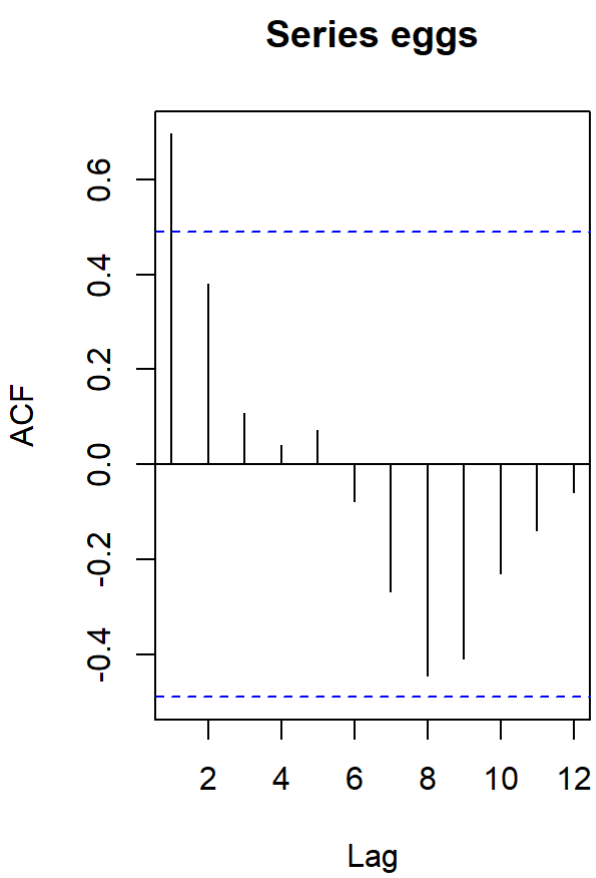


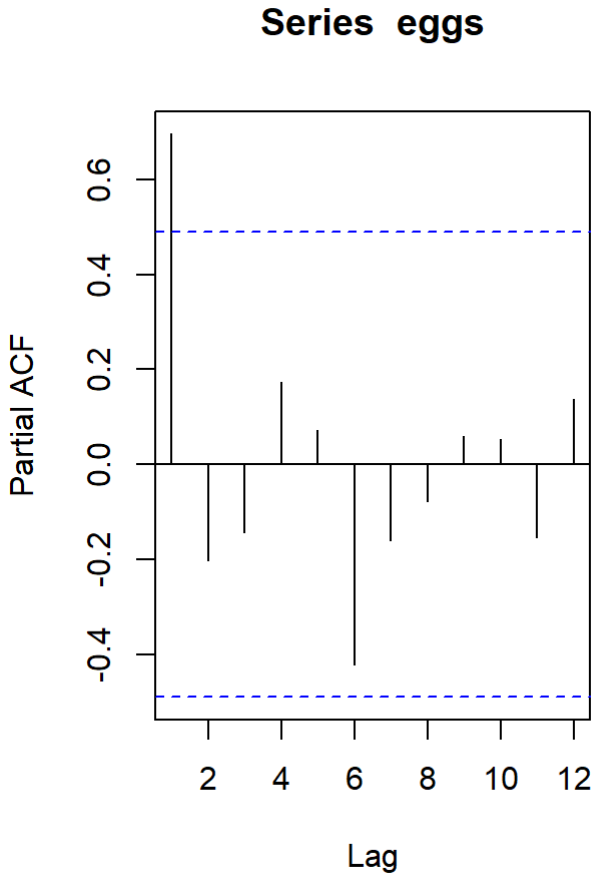
- A harmonic trend couldn't be fitted to the Eggs deposition data since there is no proof of seasonality in the data. This can be observed in figure 9 .
- Same can see from the R squared value of 0.0139 which says that 1.39% of variation in the time series data is explained by harmonic model. We can see that coefficients of this model are more than the significant level.
- Because of the above reasons its better to not analyse this model any further.

Trend Model Evaluation

Trend Model	R - Squared value
Linear Model	0.4469
Quadratic Model	0.5932
Harmonic Model	0.0139

- Above are the R squared test results for our trend models. Even though from the above table Quadratic model seems ot be the best fitted model for our predection, Analysis of residuals of above trend models makes us to doubt about these model.
- This is also accomodated by the fact that time series plot of Eggs depositions data is non-stationary. Lets make our data stationary and analyse further possible models.

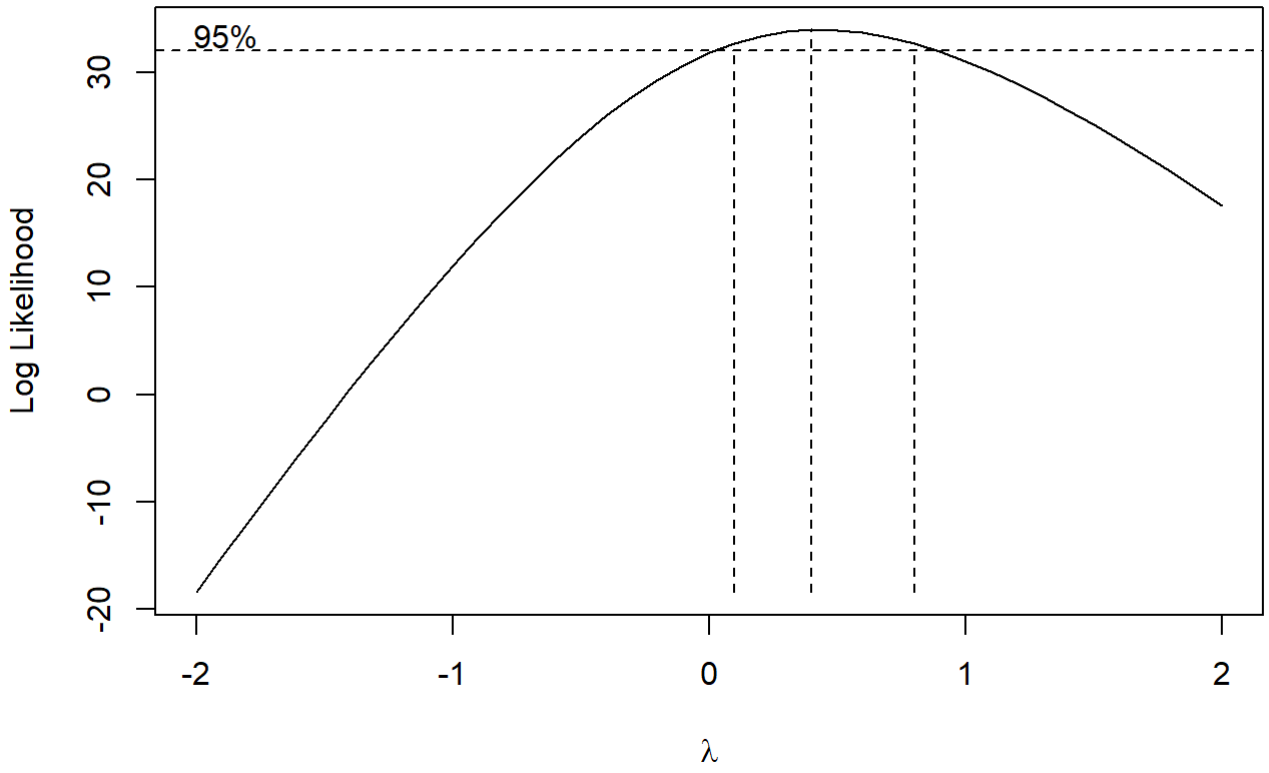




```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 0
##   STATISTIC:
##     Dickey-Fuller: -0.4911
##   P VALUE:
##     0.452
##
## Description:
##  Sun May 10 21:59:25 2020 by user: jeeva
```

- Continously decaying lags from ADF test and very high first correlation in PADF test suggest the trend and non-stationary in the data.
- This can be confirmed from Augmented Dickey-Fuller Test with the vale of 0.45. Test tells us that the data is non-stationary, becoasue p > 0.05, which fails to reject series in non-stationary.
- This non-stationarity can be fixed by differencing the data. But before that lets check if transformation is needed.

Applying Transformation

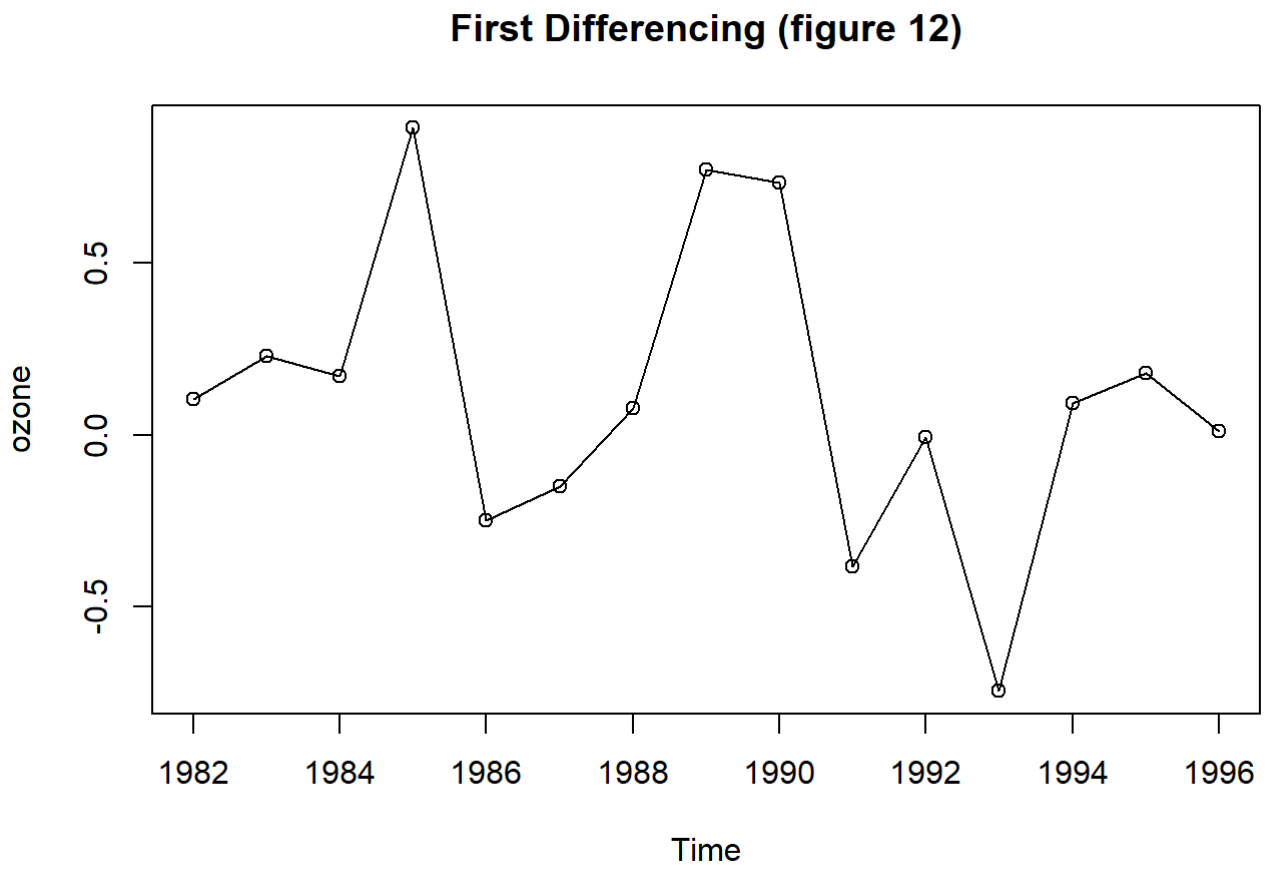



```
## [1] 0.1 0.8
```

- Changing variance can be corrected by taking transformation of the data. We shall take BoxCox transformation

- From the above test, We can see log likelyhood function lamda is in between 0.1 and 0.8. the middle value of 0.45 is taken for further analysis.

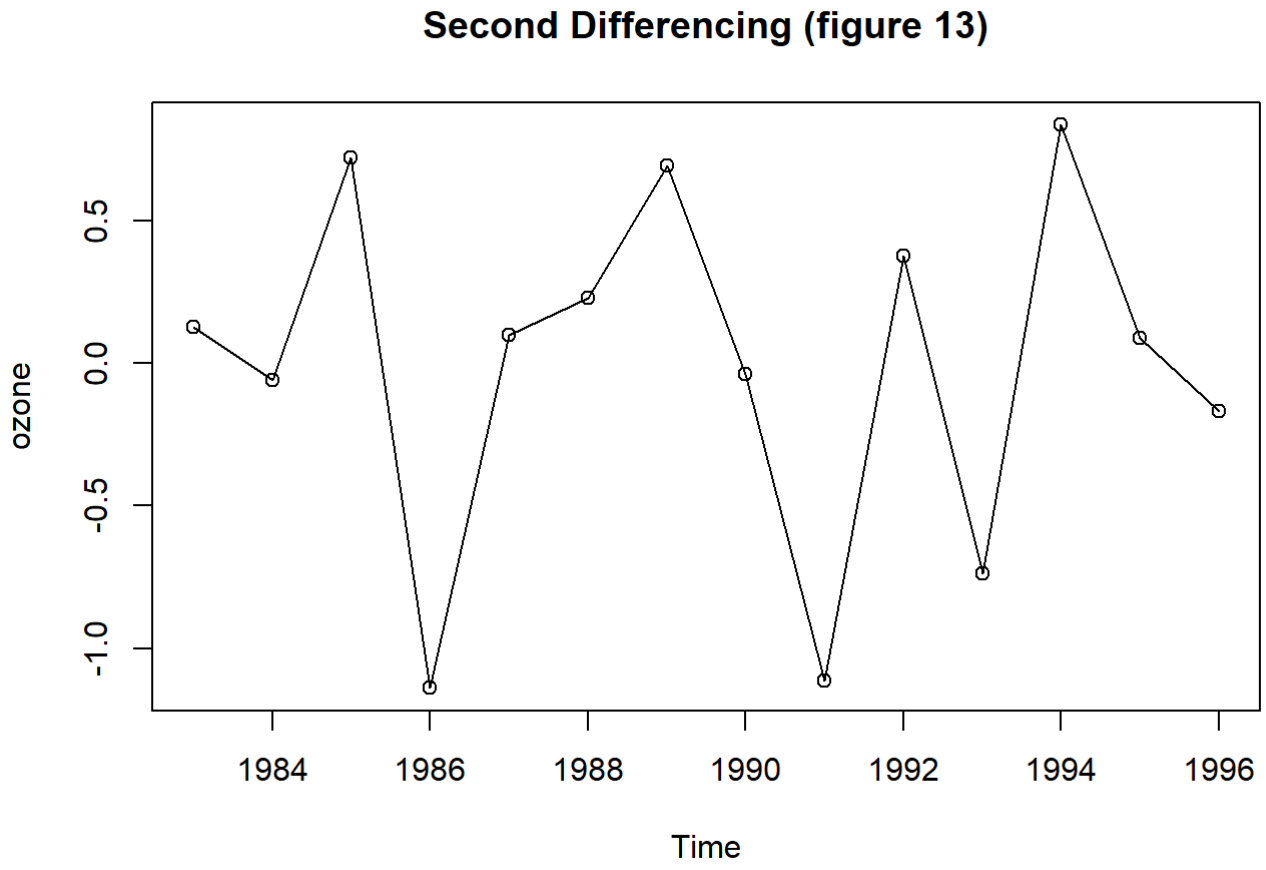
Applying First Differencing to Transformed Data




```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 4
##   STATISTIC:
##     Dickey-Fuller: -0.8222
##   P VALUE:
##     0.3469
##
## Description:
##  Sun May 10 21:59:26 2020 by user: jeeva
```

- Applying first differencing to our data did not stabilize stationarity of our dataset. we didn't detrended it yet.
- This can be made sure using ADF test. With p value of 0.35, where p>0.05 which fails to reject the null hypothesis. so our data still non-stationary.

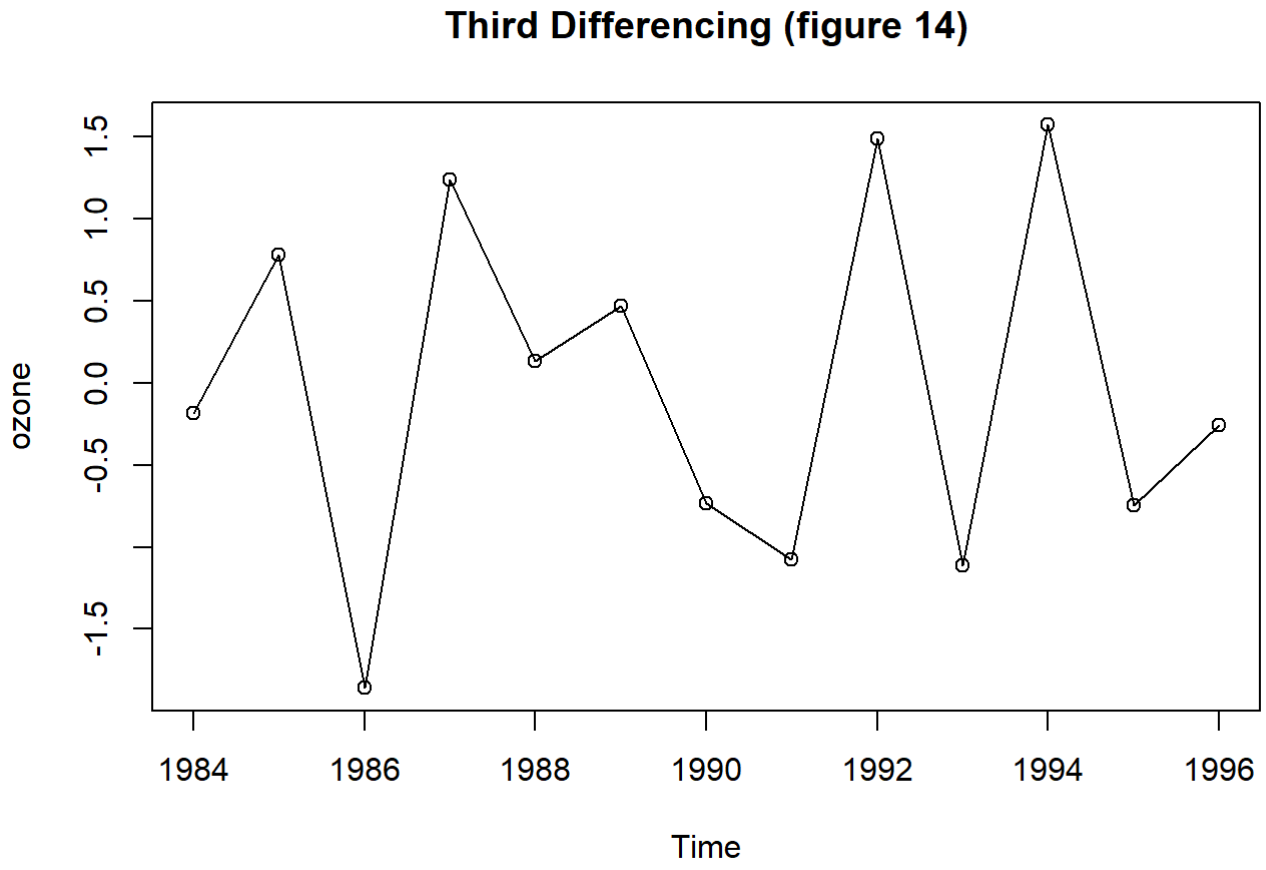
Applying Second Differencing to Transformed Data




```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 4
##   STATISTIC:
##     Dickey-Fuller: -1.5692
##   P VALUE:
##     0.1098
##
## Description:
##  Sun May 10 21:59:26 2020 by user: jeeva
```

- Applying Second differencing to our data did not stabilize stationarity of our dataset aswell. we didn't detrended it yet.
- This can be made sure using ADF test. With p value of 0.10, where p>0.05 which fails to reject the null hypothesis. so our data is still non-stationary.

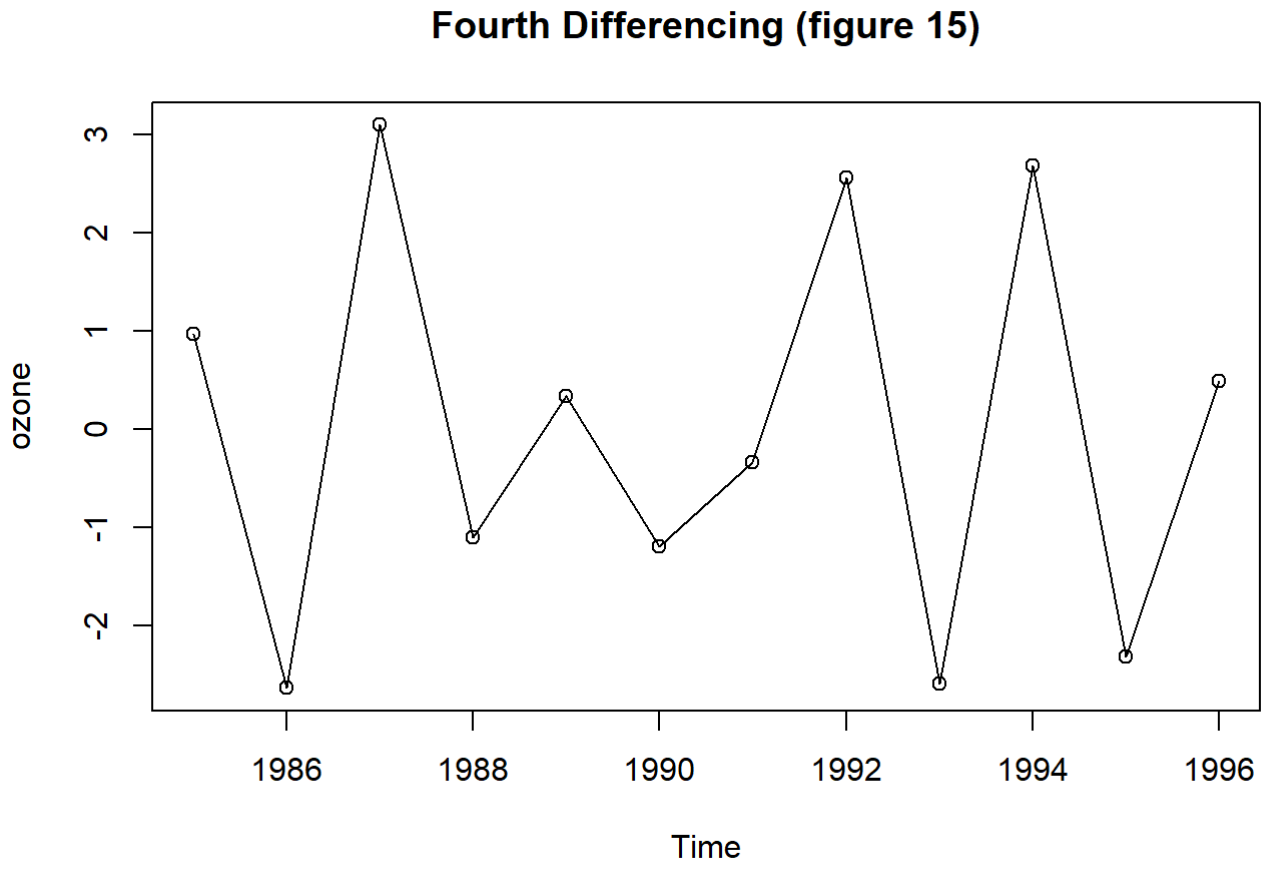
Applying Third Differencing to Transformed Data




```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 4
##   STATISTIC:
##     Dickey-Fuller: -1.3368
##   P VALUE:
##     0.1836
##
## Description:
##  Sun May 10 21:59:26 2020 by user: jeeva
```

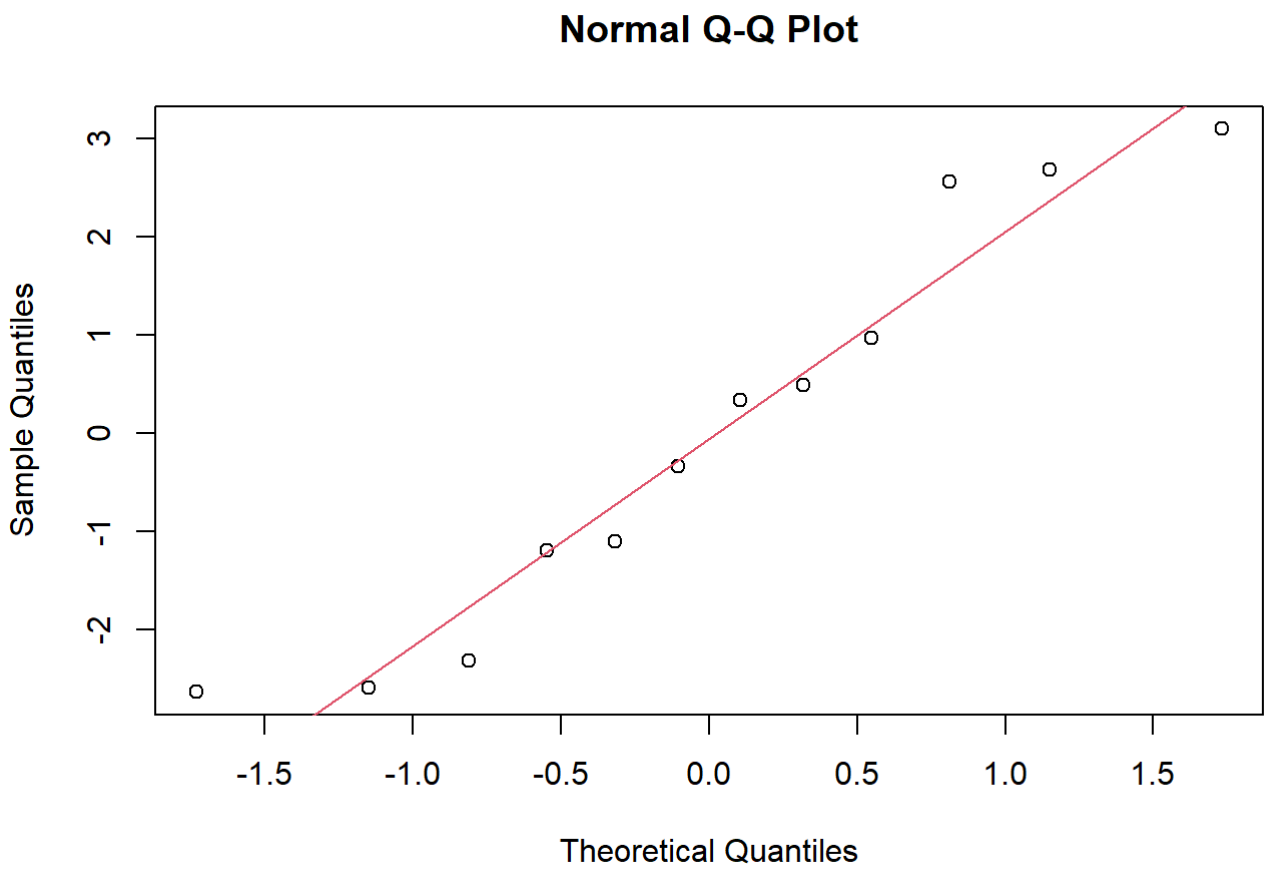
- Applying third differencing to our data did not stabilize stationarity of our dataset aswell. we didn't detrended it yet.
- This can be made sure using ADF test. With p value of 0.18, where p>0.05 which fails to reject the null hypothesis. so our data is still non-stationary.

Applying Fourth Differencing to Transformed Data




```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 2
##   STATISTIC:
##     Dickey-Fuller: -2.3228
##   P VALUE:
##     0.02265
##
## Description:
##  Sun May 10 21:59:26 2020 by user: jeeva
```

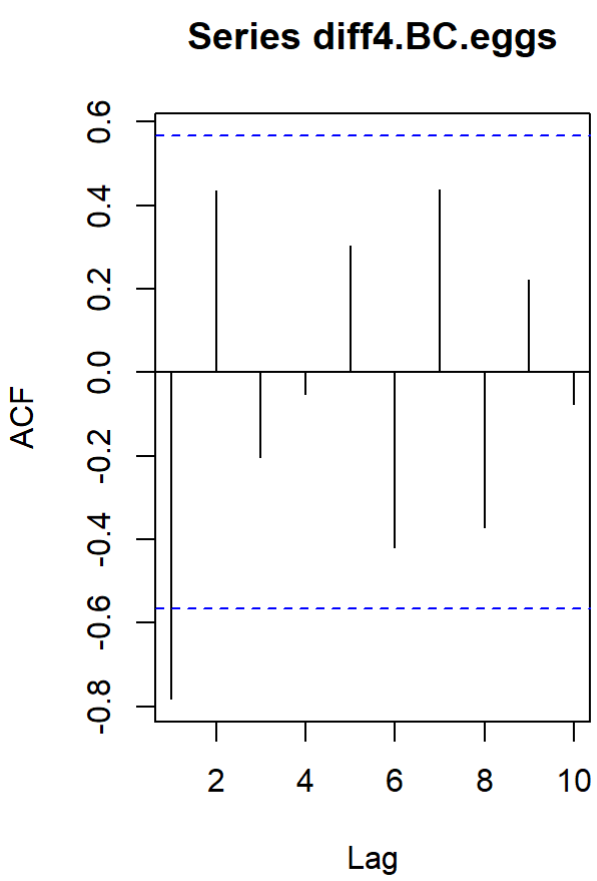
- Applying Fourth differencing on our series stabilized the non-stationarity of our dataset. we detrended our data by taking the fourth differencing successfully.
- This can be made sure using ADF test. With p value of 0.02, where p<0.05 which rejects the null hypothesis. so our data is now stationary.

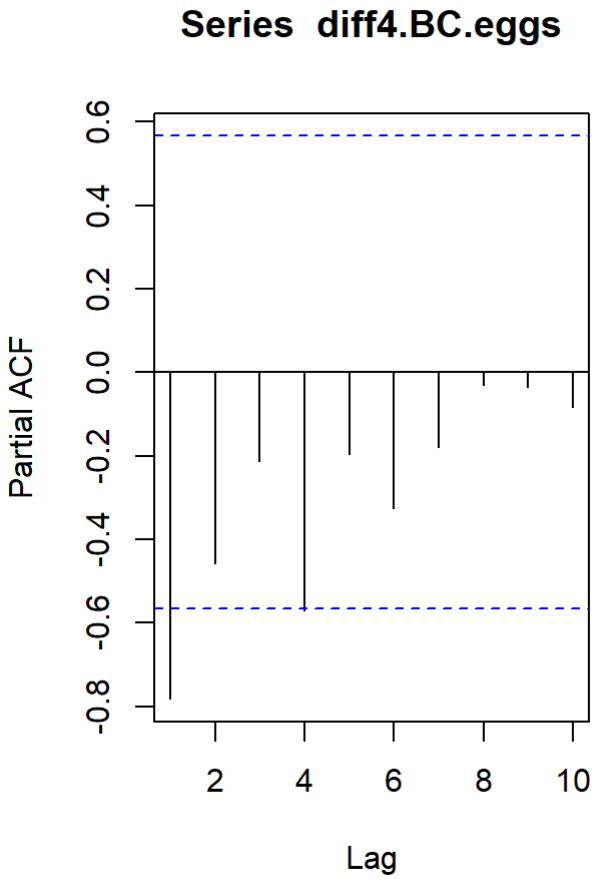



```
##
##  Shapiro-Wilk normality test
##
## data:  diff4.BC.eggs
## W = 0.9232, p-value = 0.3136
```

*Even dough from QQ-plots we can see that data points seems like they are going away from QQ Line, from shapiro test with p valve > 0.05 we can confirm that data after differencing 4 times is normally distributed.

ACF and PACF graph





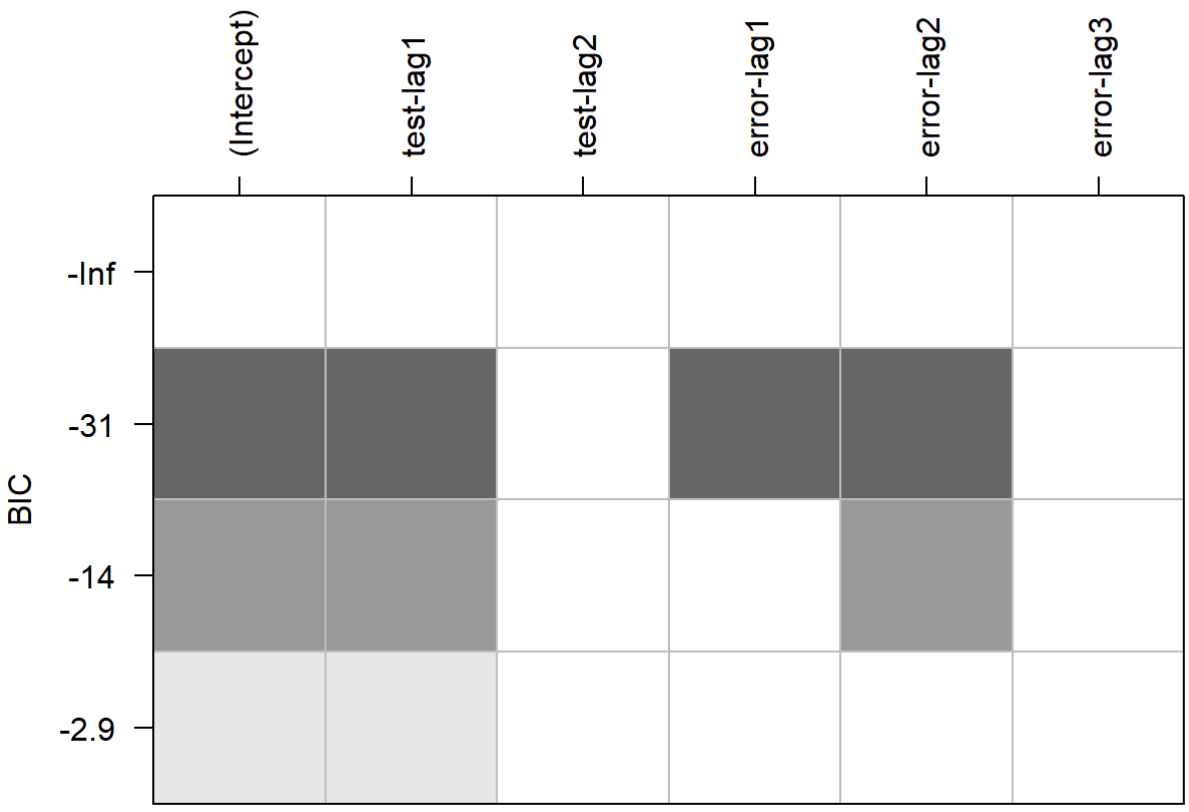
- We can see both ACF and PACF are tailing off. This sugges this is a ARIMA model.
- From ACF graph we can see 1 lag crossing the blue dotted line which suggest MA component is 1
- From PACF graph we can see there is 1 lag which is significantly away from blue dotted line and another lag almost crossed blue dotted line. So we can take AR component of either 2 or 1
- ARIMA(2,4,1) and ARIMA(1,4,1) are the possible models from ACF and PACF graphs.

EACF Graph

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

- The vertex in EACF matrix is starting from AR(0) and MA(1) taking 2 models around it lets say ARIMA(0,4,1), ARIMA(1,4,0) , ARIMA(0,4,2)and ARIMA(1,4,1) are the possible models from EACF graph.

BIC Graph



- AR(1),MA(1) and MA(2) components can be seen significant from the plot. therefore possible models from BIC graph are ARIMA(1,4,1) and ARIMA(1,4,2)

Parameter Estimation of the Probable Models

Below are the possible models we got from our tests. We have to perform parameter estimation on each model. We have check the conditional sum of squares (CSS) and the maximum likelihood estimation (ML) of each co-efficients of the model to find if they are significant. If the value of Pr(>|z|) comes under .05, it is good to say that the co-efficients are significant and we shall proceed further for the same.

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

ARIMA(1,4,1)

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value  Pr(>|z|)
## ar1 -0.68261    0.23574 -2.8956  0.003785 **
## ma1 -0.85228    0.13550 -6.2899 3.176e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value  Pr(>|z|)
## ar1 -0.59065    0.20848 -2.8332  0.004609 **
## ma1 -0.97233    0.23563 -4.1265 3.684e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For both CSS amd ML test we can see that first co-efficients of both AR and MA components are significant for model ARIMA(1,4,1)

ARIMA(1,4,2)

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value Pr(>|z|)
## ar1 -0.73180    0.27154 -2.6950 0.007038 **
## ma1 -0.72655    0.46791 -1.5528 0.120477
## ma2 -0.13349    0.47970 -0.2783 0.780797
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value  Pr(>|z|)
## ar1 -0.26282    0.27980 -0.9393   0.34758
## ma1 -1.84548    0.39310 -4.6946 2.671e-06 ***
## ma2  0.90979    0.39755  2.2885   0.02211 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above results we can see that co-efficients are not significant in either CSS or ML for model ARIMA(1,4,2).

ARIMA(0,4,1)

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value  Pr(>|z|)
## ma1 -1.07312    0.10255 -10.465 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value  Pr(>|z|)
## ma1 -0.97867    0.20858  -4.692 2.705e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From both CSS amd ML test we can see that co-efficients of MA is significant for model ARIMA(0,4,1)

ARIMA(1,4,0)

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value  Pr(>|z|)
## ar1 -0.78613    0.17493 -4.4941 6.988e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value  Pr(>|z|)
## ar1 -0.75117    0.16141 -4.6539 3.257e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From both CSS amd ML test we can see that co-efficients of AR is significant for model ARIMA(1,4,0)

ARIMA(2,4,1)

```
##
## 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
```

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value  Pr(>|z|)
## ar1 -0.74683    0.28782 -2.5948 0.0094657 **
## ar2 -0.21570    0.28852 -0.7476 0.4546922
## ma1 -0.96689    0.25620 -3.7740 0.0001607 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From CSS test 2nd coefficient of AR and MA co-efficients are not significant and from ML test 2nd co-efficient of AR is not significant for ARIMA(2,4,1) model.

ARIMA(0,4,2)

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value  Pr(>|z|)
## ma1 -1.86033    0.15221 -12.223 < 2.2e-16 ***
## ma2  0.98970    0.14778   6.697 2.127e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value Pr(>|z|)
## ma1 -1.86524    0.32284 -5.7775 7.58e-09 ***
## ma2  0.94502    0.31884  2.9640 0.003037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From both CSS amd ML test we can see that all co-efficients of AR and MA is significant for model ARIMA(0,4,2)

- Therefore below models are considered for further analysis after Parameter Estimation of the Probable Models

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

AIC and BIC values of selected models

```
##              df      AIC
## model_042_ml  3 42.61715
## model_141_ml  3 44.96153
## model_041_ml  2 48.39360
## model_140_ml  2 49.32783
```

```
##              df      BIC
## model_042_ml  3 44.07187
## model_141_ml  3 46.41625
## model_041_ml  2 49.36341
## model_140_ml  2 50.29765
```

From the above AIC and BIC table we can finalize ARIMA(0,4,2) as the best fit to predict on this particular data.

Model Overfitting

- Overfitting is made to the best model by increasing one component by one each at a time to find out if there is any other best model. So we will check for ARIMA(0,4,3) and ARIMA(1,4,2)
- Since analysis for ARIMA(1,4,2) is already performed we will do Parameter Estimation for model ARIMA(0,4,3) only

ARIMA(0,4,3)

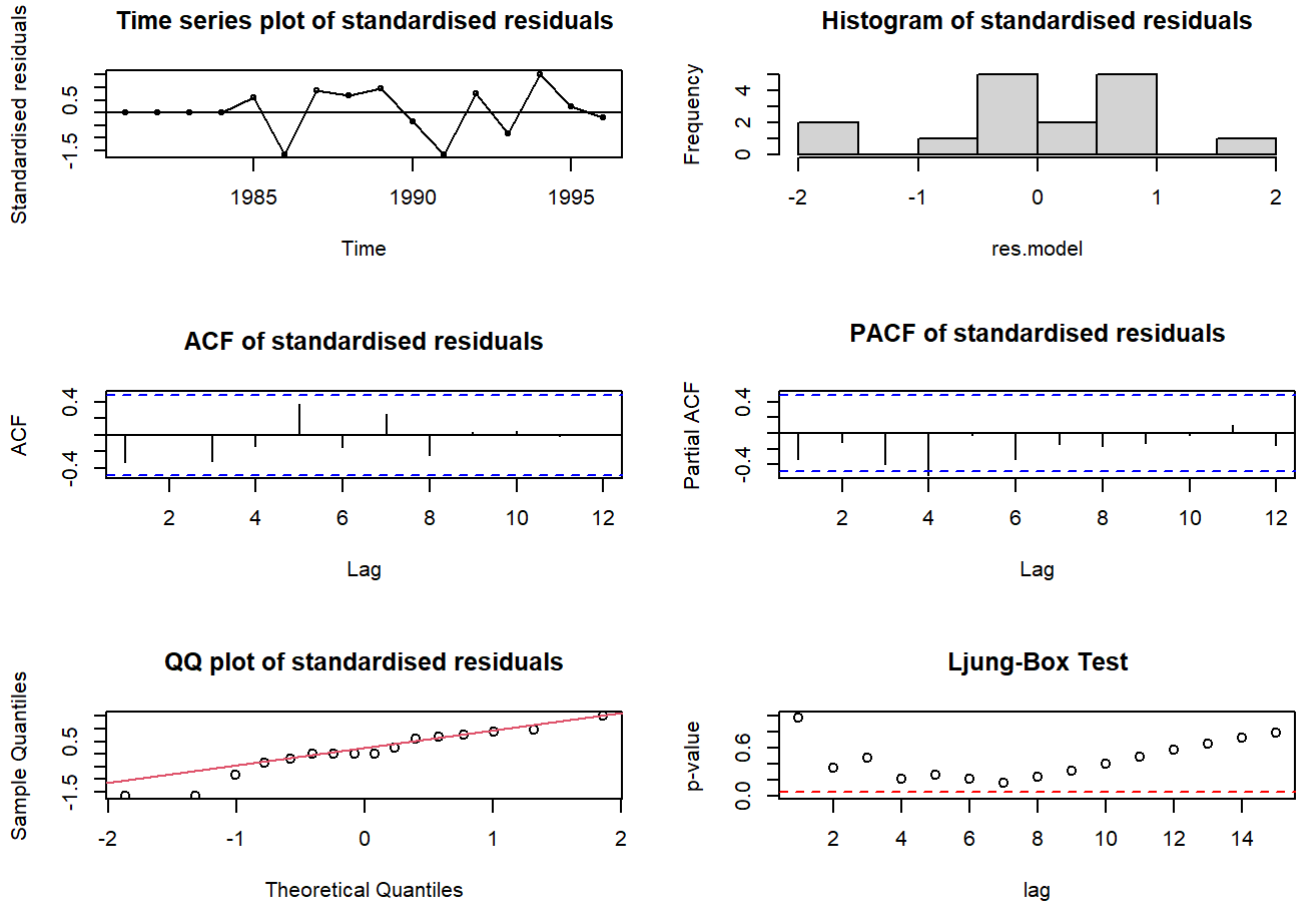
```
##
## z test of coefficients:
##
##     Estimate Std. Error z value  Pr(>|z|)
## ma1 -2.04563    0.30687 -6.6661 2.627e-11 ***
## ma2  1.38897    0.59467  2.3357   0.01951 *
## ma3 -0.24057    0.34033 -0.7069   0.47966
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value  Pr(>|z|)
## ma1 -1.93497    0.42185 -4.5869 4.499e-06 ***
## ma2  1.03498    0.62442  1.6575   0.09742 .
## ma3 -0.01929    0.31165 -0.0619   0.95064
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From CSS and ML test we can see that coefficients are not significant for MA for madel(0,4,3). We will reject this model

Residual Analysis of the best Model

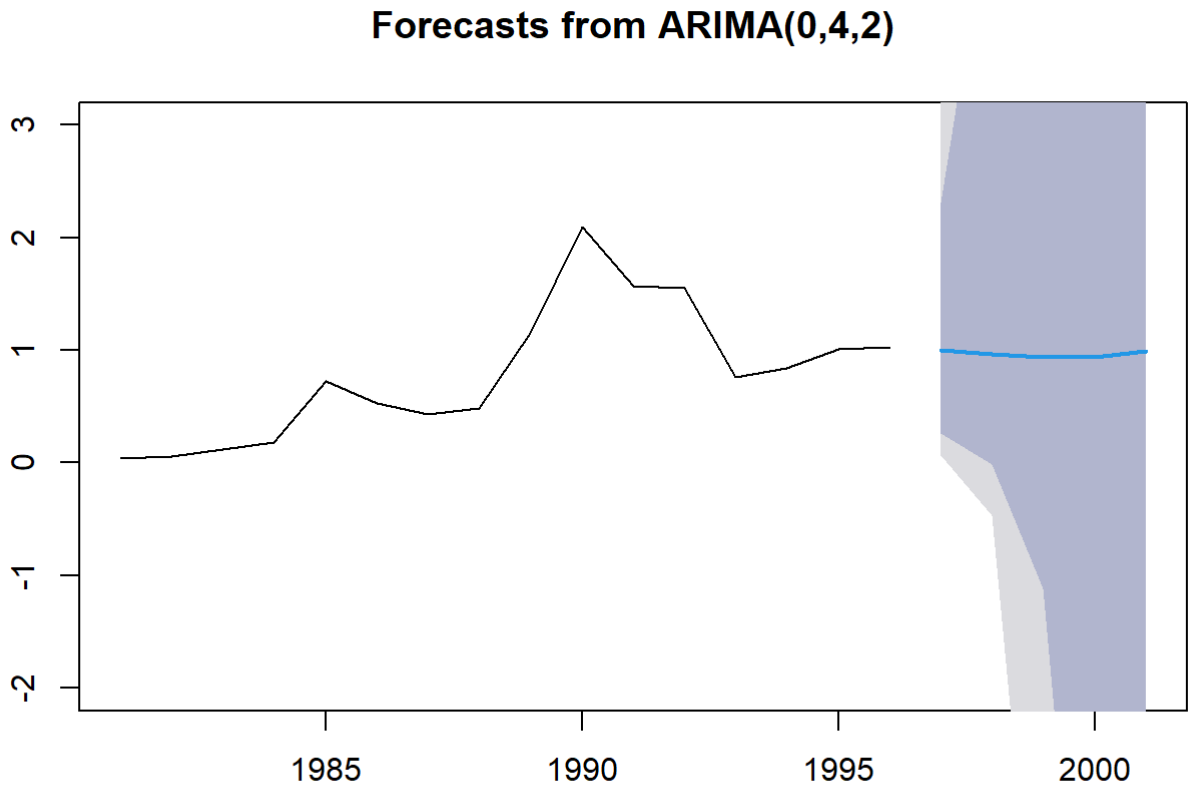
```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.9356, p-value = 0.2986
##
##
##  Box-Ljung test
##
## data:  res.model
## X-squared = 13.613, df = 10, p-value = 0.1914
```



1. **Residual VS fitted pattern plot** :This plot shows that data points are scattered around. Residuals are randomly distributed between limit -1.5 to 1.5
2. **Histogram of residuals** :From histogram of residuals it can be seen that residuals are normally distributed.
3. **QQ Plots** : In QQ plots of residuals we can see that data points are somewhat alligned to the red line. This says that residuals are almost normally distributed. This can be ensured in shapiro test.
4. **Shapiro Test** : With the test result of 0.29 Shapiro test fail to rejects the Null hypothesis and concludes that residuals are normally distributed.
5. **ACF and PACF of standard Residuals** : No Significant lags crossing the blue dotted lines. Even though one lag almost seems like it has crossed the blue dotted line it has not. This can be made sure in Ljung-Box Test. Residuals are uncorrelated and Ljug-Box test result supports it.
6. **Box-Ljung Test** : p-values is 0.19 which is greater than 0.05. so this confirms that ARIMA(0,4,2) model successfully deals with serial correlation.

From all the above tests, it is clear that ARIMA(0,4,2) model is best fit model when compared to all other models and it is good to go with for forecasting.

Forecasting




```
##      Point Forecast        Lo 80     Hi 80        Lo 95      Hi 95
## 1997      1.0000891   0.25823562  2.305648   0.07001039   3.242599
## 1998      0.9642666  -0.00953809  5.149759  -0.46606481   8.917579
## 1999      0.9379195  -1.12107483 11.453943  -5.33569895  22.598049
## 2000      0.9412374  -6.27436611 24.670209 -21.89263419  52.740443
## 2001      0.9956375 -20.26429525 51.039303 -62.98238264 114.652448
```

- Forecasting for next five years with upper and lower 5% and 20% forecast limit can be seen from Time series graph.
- Forecast for Egg count in millions from 1997 to 2001 with 80% and 95% confidance interval is shown in above chart

Conclusion

As per the forecasting based on ARIMA(0,4,2) model, There should not be much variation in the Egg depositions and it should maintain almost constant for the next 5 years.

Summary

In this project we managed the dataset speaking to egg depositions (in millions) of age-3 Lake Huron Bloaters (Coregonus hoyi) between years 1981 and 1996. The objective was to locate the best fitting model to the dataset and give forecasts of yearly changes for the following 5 years. Trend, seasonal and ARIMA models were applied and their outputs and the plots were obtained. In view of which the appropriate model was picked and taken forward for the diagnostic testing. As observed, pattern models were removed in the underlying stage simply because they fizzled the analyze testing. So we further put together our investigation with respect

to the ARIMA models. In the wake of utilizing different model detail techniques and applying different diagnostics tests we picked ARIMA(0,4,2) as the most reasonable model out of candidate models.Our diagnostic stage for picked ARIMA model included after tests-

1. **Residual analysis**
2. **Histogram of residuals**
3. **ACF and PACF plots**
4. **Shapiro Wilk test**
5. **Ljung-Box test**

ARIMA(0,4,2) successfuly clears all the diagnostic tests and comes up as the best fit model for the given dataset of egg depositions for age-3 Lake Huron Bloaters. At last the estimate is shown for the next 5 years which forecasts “There should not be much variation in the Egg depositions and it should maintain almost constant for the next 5 years” .

Appendix

```
#Load the requied libraries
library(TSA)
library(readr)
library(tseries)
library(fUnitRoots)
library(lmtest)
library(FSAdata)
library(forecast)


#Read the dataset
eggs <- BloaterLH$eggs

#Convert the dataframe to time series data
eggs <- ts(as.vector(eggs), start = 1981)
plot(eggs,type='o',ylab='Count of Eggs in millions',xlab='Time ',  main = 'Time Series Plot o
f egg deposition (figure 1)')
#Plot the scatter plot between egg count layer and its first lag
plot(y=eggs,x=zlag(eggs),ylab='Egg deposition(y)', xlab='First lag in Egg deposition(y-1)' ,
 main = "Scatter plot of Egg deposition between Y and Y-1 (figure2)")


#Find the Autocovariance between thickness of Ozone layer
y = eggs
x = zlag(eggs)        # Generate first lag
index = 2:length(x)          # Create an index to get rid of the first NA value and the last
cor(y[index],x[index])

#Building the Linear Model
model.eggs.ln = lm(eggs~time(eggs))
summary(model.eggs.ln)
plot(eggs,type='o',ylab='y', xlab='Time', main = "Linear regression model  (figure 3)")
abline(model.eggs.ln, col = "Blue", lty = 2)
#Residual VS fitted trend plot for Linear regression model
res.model.eggs.ln = rstudent(model.eggs.ln)
plot(y = res.model.eggs.ln, x = as.vector(time(eggs)),xlab = 'Time', ylab='Standardized Resid
uals', main = "Residual VS fitted trend plot (figure 4)",type='o')

#Histogram of residuals for Linear regression model
g = res.model.eggs.ln
m<-mean(g)
std<-sqrt(var(g))
hist(g, prob=TRUE, xlab="Y-1",ylab="Y", main="normal curve over histogram figure(5)")
curve(dnorm(x, mean=m, sd=std),col="Red", lwd=2, add=TRUE, yaxt="n")

#QQ Plots of residuals for Linear regression model
qqnorm(res.model.eggs.ln, main = "Normal Q-Q Plot (figure 6)")
qqline(res.model.eggs.ln, col = 2, lwd = 1, lty = 2)

#Shapiro Test of residuals for Linear regression model
shapiro.test(res.model.eggs.ln)

#Independence test of residuals for Linear regression model
y = rstudent(model.eggs.ln)
runs(y)
#acf plot of residuals for Linear regression model
acf(res.model.eggs.ln)
```

```
#Quadratic Model
t = time(eggs)
t2 = t^2
model.eggs.qa = lm(eggs~ t + t2)
summary(model.eggs.qa)
plot(ts(fitted(model.eggs.qa)), ylim = c(min(c(fitted(model.eggs.qa),
                                                as.vector(eggs))), max(c(fitted(mode
l.eggs.qa),as.vector(eggs)))),
     ylab='y', xlab = 'Time' , main = "Fitted quadratic curve to Egg deposition data (figure
 7)", type="l",lty=2,col="red")
lines(as.vector(eggs),type="o")
#Residual VS fitted trend plot for Quadratic regression model
res.model.eggs.qa = rstudent(model.eggs.qa)
plot(y = res.model.eggs.qa, x = as.vector(time(eggs)),xlab = 'Time', ylab='Standardized Resid
uals',main = "Residual VS fitted trend plot (figure 8)",type='o')


#Histogram of residuals of residuals for Quadratic regression model
g = res.model.eggs.qa
m<-mean(g)
std<-sqrt(var(g))
hist(g, prob=TRUE, xlab="Y-1",ylab="Y", main="normal curve over histogram (figure 9)")
curve(dnorm(x, mean=m, sd=std),col="Red", lwd=2, add=TRUE, yaxt="n")

#QQ Plots of residuals for Quadratic regression model
qqnorm(res.model.eggs.qa, main="Normal Q-Q Plot (figure 10)")
qqline(res.model.eggs.qa, col = 2, lwd = 1, lty = 2)

#Shapiro Test of residuals for Quadratic regression model
shapiro.test(res.model.eggs.qa)


#Independence test of residuals for Quadratic regression model
y = rstudent(model.eggs.ln)
runs(y)

#acf graph of residuals for Quadratic regression model
acf(res.model.eggs.qa)
# Harmonic Model
har.=harmonic(eggs, 0.45)
model.eggs.har=lm(eggs~har.)
summary(model.eggs.har)

#plot the time series graph with harmonic predection line
plot(ts(fitted(model.eggs.har)), ylim = c(min(c(fitted(model.eggs.har),
                                                 as.vector(eggs))), max(c(fitted(mode
l.eggs.har),as.vector(eggs)))),
     ylab='y' ,xlab='Time', main = "Fitted quadratic curve to Egg deposition data (figure 1
1)", type="l",lty=2,col="red")
lines(as.vector(eggs),type="o")

#Plot acf and pacf graph on original time series data
par(mfrow=c(1,2))
acf(eggs)
pacf(eggs)
par(mfrow=c(1,1))
```

```
#ADF test of original time series data
order <- ar(diff(eggs))$order
adfTest(eggs, lags = order, title = NULL, description = NULL)
#BoxCox Transformation on original time series data
eggs.transform1 = BoxCox.ar(eggs, method = "yule-walker")
eggs.transform1$ci
lambda =0.45
BC.eggs = (eggs^lambda-1)/lambda
#First Differencing the  data
diff1.BC.eggs = diff(BC.eggs, differences = 1)
plot(diff1.BC.eggs,type='o',ylab='ozone', main = "First Differencing (figure 12)")

#ADF test of First Differencing
order <- ar(diff(diff1.BC.eggs))$order
adfTest(diff1.BC.eggs, lags = order, title = NULL, description = NULL)

#Differencing data with d = 2
diff2.BC.eggs = diff(BC.eggs, differences = 2)
plot(diff2.BC.eggs,type='o',ylab='ozone', main = "Second Differencing (figure 13)")

#ADF test for Second differencing
order <- ar(diff(diff2.BC.eggs))$order
adfTest(diff2.BC.eggs, lags = order, title = NULL, description = NULL)

#Differencing data with d = 3
diff3.BC.eggs = diff(BC.eggs, differences = 3)
plot(diff3.BC.eggs,type='o',ylab='ozone', main = "Third Differencing (figure 14)")

#ADF test for third differencing
order <- ar(diff(diff3.BC.eggs))$order
adfTest(diff3.BC.eggs, lags = order, title = NULL, description = NULL)


#Differencing data with d = 4
diff4.BC.eggs = diff(BC.eggs, differences = 4)
plot(diff4.BC.eggs,type='o',ylab='ozone', main = "Fourth Differencing (figure 15)")

#ADF test for fourth differencing
order <- ar(diff(diff4.BC.eggs))$order
adfTest(diff4.BC.eggs, lags = order, title = NULL, description = NULL)

#QQ pot of differenced data
qqnorm(diff4.BC.eggs)
qqline(diff4.BC.eggs, col = 2)

#Shapiro test after differencing the data four times
shapiro.test(diff4.BC.eggs)

#Plot acf and pacf graph of stationary eggs data
par(mfrow=c(1,2))
acf(diff4.BC.eggs)
pacf(diff4.BC.eggs)
par(mfrow=c(1,1))


#EACF graph
eacf(diff4.BC.eggs, ar.max=2, ma.max=2)
#BIC graph
```

```
res = armasubsets(y = diff4.BC.eggs, nar = 2, nma = 3, y.name = 'test', ar.method = 'ols')
plot(res)

#Parameter Estimation of ARIMA(1,4,1)
model_141_css = arima(BC.eggs, order = c(1,4,1), method = 'CSS')
coeftest(model_141_css)
model_141_ml = arima(BC.eggs, order = c(1,4,1), method = 'ML')
coeftest(model_141_ml)

#Parameter Estimation of ARIMA(1,4,2)
model_142_css = arima(BC.eggs, order = c(1,4,2), method = 'CSS')
coeftest(model_142_css)
model_142_ml = arima(BC.eggs, order = c(1,4,2), method = 'ML')
coeftest(model_142_ml)


#Parameter Estimation of ARIMA(0,4,1)
model_041_css = arima(BC.eggs, order = c(0,4,1), method = 'CSS')
coeftest(model_041_css)
model_041_ml = arima(BC.eggs, order = c(0,4,1), method = 'ML')
coeftest(model_041_ml)

#Parameter Estimation of ARIMA(1,4,0)
model_140_css = arima(BC.eggs, order = c(1,4,0), method = 'CSS')
coeftest(model_140_css)
model_140_ml = arima(BC.eggs, order = c(1,4,0), method = 'ML')
coeftest(model_140_ml)

#Parameter Estimation of  ARIMA(2,4,1)
model_241_css = arima(BC.eggs, order = c(2,4,1), method = 'CSS')
coeftest(model_241_css)
model_241_ml = arima(BC.eggs, order = c(2,4,1), method = 'ML')
coeftest(model_241_ml)


#Parameter Estimation of the ARIMA(0,4,2)
model_042_css = arima(BC.eggs, order = c(0,4,2), method = 'CSS')
coeftest(model_042_css)
model_042_ml = arima(BC.eggs, order = c(0,4,2), method = 'ML')
coeftest(model_042_ml)

#Function to sort AIC and BIC score of probable models
sort.score <- function(x, score = c("bic", "aic")){
  if (score == "aic"){
    x[with(x, order(AIC)),]
  } else if (score == "bic") {
    x[with(x, order(BIC)),]
  } else {
    warning('score = "x" only accepts valid arguments ("aic","bic")')
  }
}


#Calling sort.score function with probable models
sort.score(AIC(model_042_ml,model_141_ml, model_041_ml, model_140_ml), score = "aic")
sort.score(BIC(model_042_ml,model_141_ml, model_041_ml, model_140_ml), score = "bic")
```

```
#Parameter Estimation of ARIMA(0,4,3)
model_043_css = arima(BC.eggs, order = c(0,4,3), method = 'CSS')
coeftest(model_043_css)
model_043_ml = arima(BC.eggs, order = c(0,4,3), method = 'ML')
coeftest(model_043_ml)

#Function to test residual analysis of best model
residual.analysis <- function(model, std = TRUE,start = 2, class = c("ARIMA","GARCH","ARMA-GA
RCH")[1]){
  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$residuals
  }else {
    stop("The argument 'class' must be either 'ARIMA' or 'GARCH' ")
  }
  par(mfrow=c(3,2))
  plot(res.model,type='o',ylab='Standardised residuals', main="Time series plot of standardis
ed residuals")
  abline(h=0)
  hist(res.model,main="Histogram of standardised residuals")
  acf(res.model,main="ACF of standardised residuals")
  pacf(res.model,main="PACF of standardised residuals")
  qqnorm(res.model,main="QQ plot of standardised residuals")
  qqline(res.model, col = 2)
  print(shapiro.test(res.model))
  print(Box.test(res.model, lag=10, type="Ljung-Box"))
  k=0
  LBQPlot(res.model, lag.max = length(model$residuals)-1, StartLag = k + 1, k = 0, SquaredQ =
FALSE)
}

#calling residual analysis function for model ARIMA(0,4,2)
residual.analysis(model_042_ml)
par(mfrow = c(1,1))

# fitting time series data to model ARIMA(0,4,2)
fit = Arima(eggs, c(0,4,2),lambda = .45)

#Plotting Time series graph with forecast for next 5 years
plot(forecast(fit, h = 5), ylim=c(-2,3))

#Forecasting for next 5 years
forecast(fit, h = 5)
```