Analysis & Prediction of Egg Depositions of age-3 Lake Huron Bloaters (Coregonus hoyi)

Time Series Analysis [redacted]

Vishesh Jain [redacted]

06 May, 2018

Contents

1 INTRODUCTION

The objective of this report is to analyze the egg depositions of age-3 Lake Huron Bloasters from 1981 to 1996 and provide forecast next 5 years, using the best fit model.

To find the best fit model, both trend models and ARIMA models are explored and diagnosed using various methods.

After selecting the best model, depositions from 1997 to 2001 are predicted and visualised.

2 PACKAGES

Following packages will be used for analysis and prediction.

```
library(TSA)
library(dplyr)
library(knitr)
library(FSAdata)
library(fUnitRoots)
library(lmtest)
library(AID)
library(forecast)
library(ggplot2)
```

3 DATA

Following information is given for the data:

- The dataset gives Egg depositions (in millions) of age-3 Lake Huron Bloaters (Coregonus hoyi) between years 1981 and 1996
- Column 'eggs' will be used as it represent the depositions
- Currently the dataset is a dataframe and not a time series

```
data("BloaterLH")
class(BloaterLH)
```

```
## [1] "data.frame"
```

Table 1: Sample Data

year	eggs
1981	0.0402
1982	0.0602
1983	0.1205
1984	0.1807
1985	0.7229

3.1 Data Transformation

Function ts() from TSA() package will be used to convert the given data set into a time series.

```
eggs <- BloaterLH[,2] #Using only eggs column
eggs.ts <- ts(as.vector(eggs), start = 1981)
```

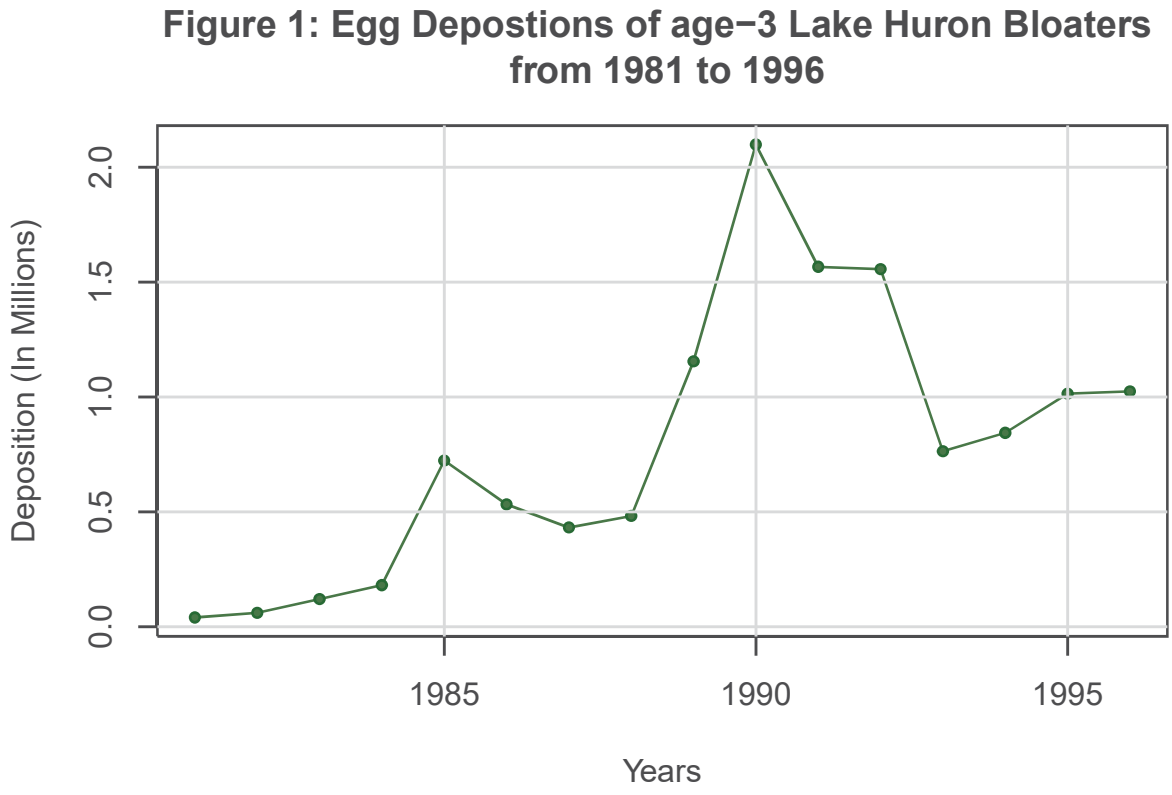
Checking the class of our data set, it shows that it has been successfully converted into a time-series data.

3.2 Data Exploration

Here we will use visualisation and statistical methods to explore our time-series

3.2.1 Series Visualisation

```
plot((eggs.ts), type = "o", ylab = 'Deposition (In Millions)', xlab = 'Years',
     main = 'Figure 1: Egg Depostions of age-3 Lake Huron Bloaters \n from 1981 to 1996',
     col = 'darkgreen', pch = 20, lwd=2)
grid()
```



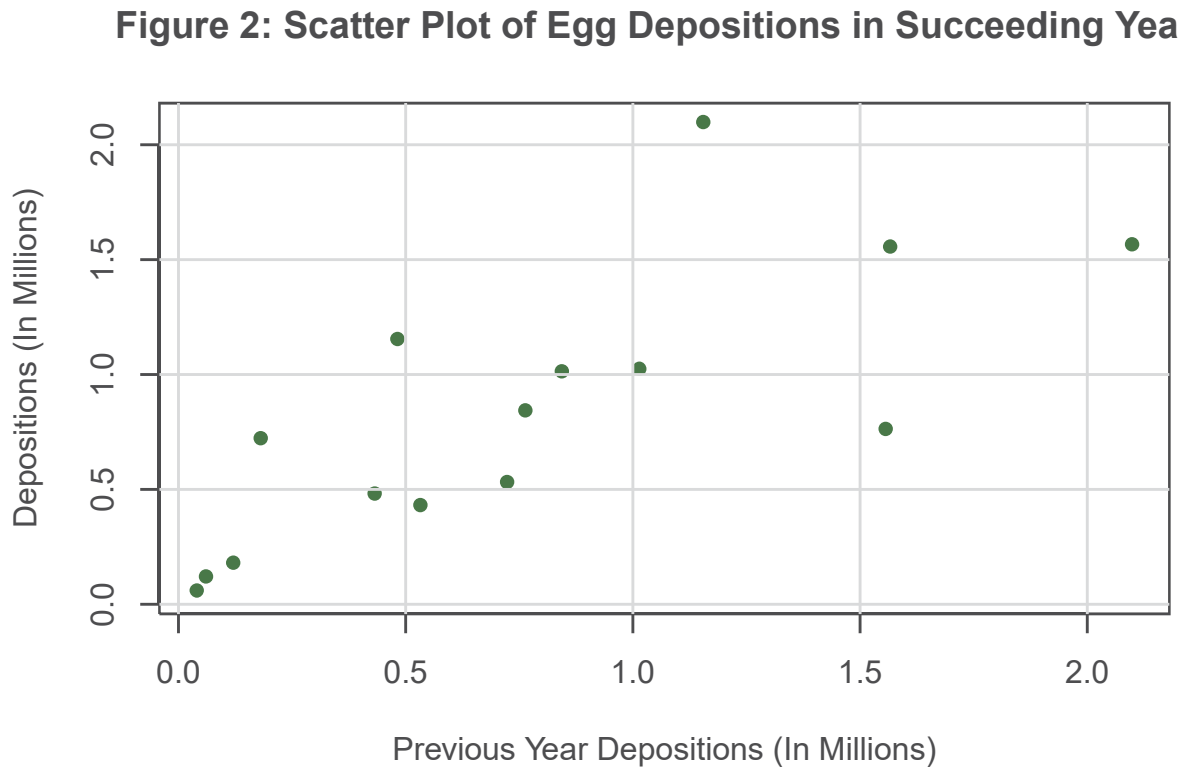
As per Fig. 1, There was a peak in deposition in 1990 but overall there is an upward trend.

- **Trend** : There is clearly an updward trend
- **Stationarity** : As there is a trend, the series is non-stationary
- **Seasonality** : Seasonality cannot be seen in the series
- **Behavior** : Auto - Regressive behavior

- **Variance** : Change in Variance
- **Pattern** : No repeating pattern can be seen

3.2.2 Relation Between Succeeding Points

```
plot(y = eggs.ts, x = zlag(eggs.ts),
     main = 'Figure 2: Scatter Plot of Egg Depositions in Succeeding Years',
     ylab = 'Depositions (In Millions)', xlab = 'Previous Year Depositions (In Millions)',
     col = 'darkgreen', pch = 16)
grid()
```



Checking the relation between depositions in succeeding years in Fig. 2:

- An upward trend is observed
- Lower changes are followed by lower changes in next year and higher changes are followed by higher changes
- *Positive correlation* can be seen

The degree of correlation can be calculated using:

```
y = eggs.ts
x = zlag(eggs.ts) #Previous Year change with Lag = 1
index = 2:length(x)
cor(y[index],x[index])
```

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

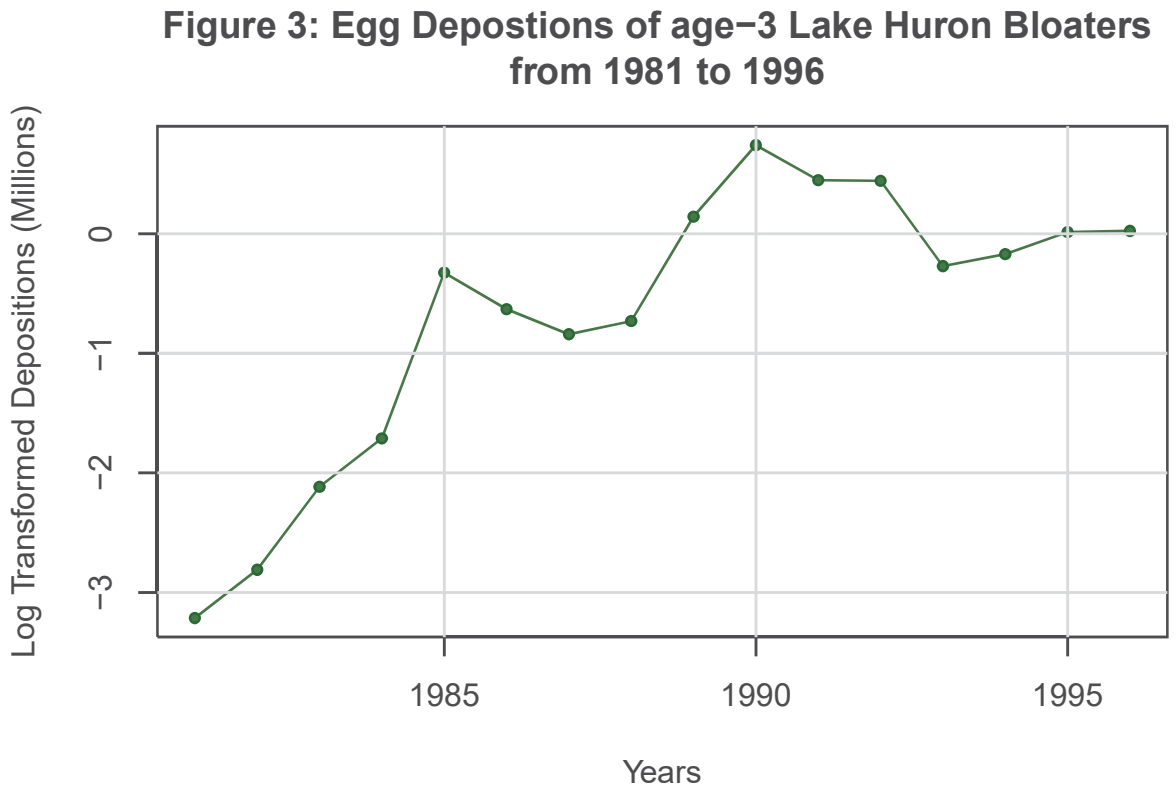
As expected, there is a **fairly high positive correlation (0.74)** between ozone layer change in succeeding years. We can say that there is a auto-regressive behavior in the time series.

3.3 Handling Change in Variance

We will use log-transformation to decrease the change in variance.

```
t.eggs = log(eggs.ts)

plot((t.eggs), type = "o", ylab = 'Log Transformed Depositions (Millions)', xlab = 'Years',
     main = 'Figure 3: Egg Depostions of age-3 Lake Huron Bloaters \n from 1981 to 1996',
     col = 'darkgreen', pch = 20, lwd=2)
grid()
```



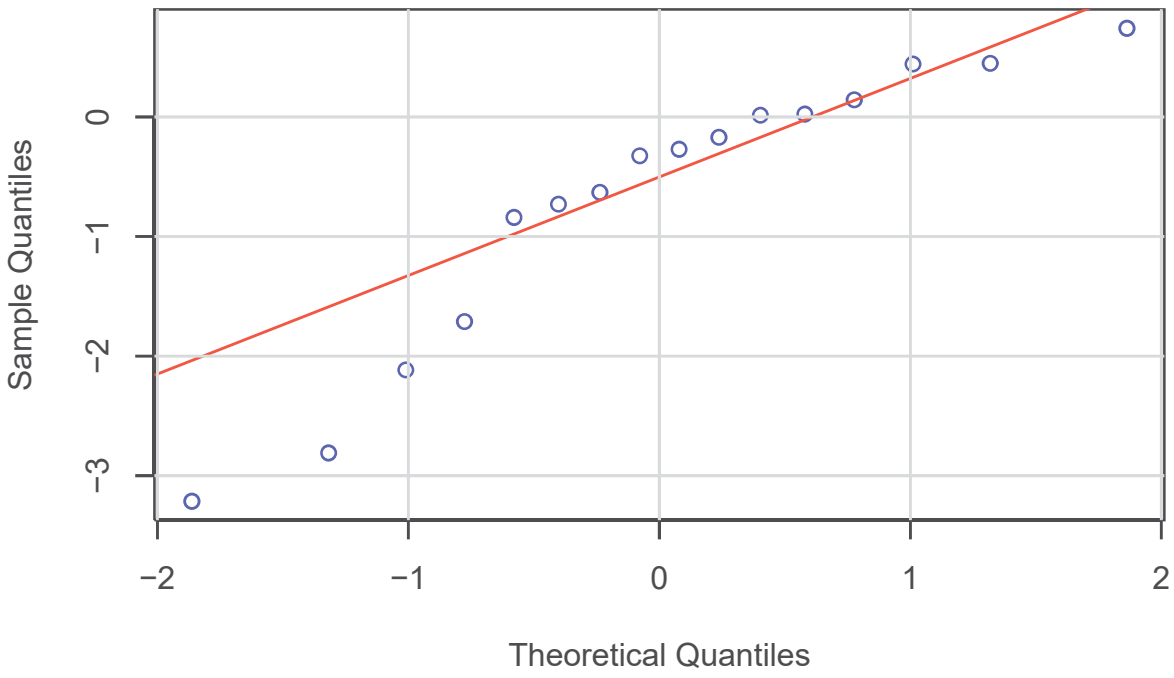
As per Fig. 3, the variance is more stabilised as compared to original series. We will use this tranformed data while using trend models.

3.4 Normality Of the Series

We will explore the distribution of the series through Q-Q plot and hypthesis test of normality.

```
qqnorm(t.eggs, col="blue", main = "Figure 4: Normal Q-Q Plot of the Transformed Data")
qqline(t.eggs, col=2)
grid()
```

Figure 4: Normal Q−Q Plot of the Transformed Data




```
# Hypothesis Test of Normality

shapiro.test(t.eggs)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  t.eggs
## W = 0.88762, p-value = 0.05108
```

- As per Fig. 4, we can observe departure of points from the normality line (red), this shows that the series is not normally distributed
- The p-value in shapiro-wilk test is greater than 0.5 (Significance Level). Therefore, the series is normal by the hypothesis test
- Because of a small sample size, normality tests have little power to reject the null hypothesis that the data come from a normal distribution. Therefore, small samples always pass normality tests.

4 TREND MODELING

4.1 Linear Trend Model

The series is treated as a linear time trend and the slope & intercept are calculated using least squares regression approach.

```
t <- time(t.eggs)
model1 = lm(t.eggs ~ t) #Linear Model with one time coefficient
summary(model1)
```

```
##
## Call:
## lm(formula = t.eggs ~ t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.02584 -0.58678 -0.03264  0.57675  1.12882
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -398.47541   77.70875  -5.128 0.000154 ***
## t              0.20004    0.03908   5.119 0.000156 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7206 on 14 degrees of freedom
## Multiple R-squared:  0.6518, Adjusted R-squared:  0.6269
## F-statistic:  26.2 on 1 and 14 DF,  p-value: 0.0001561
```

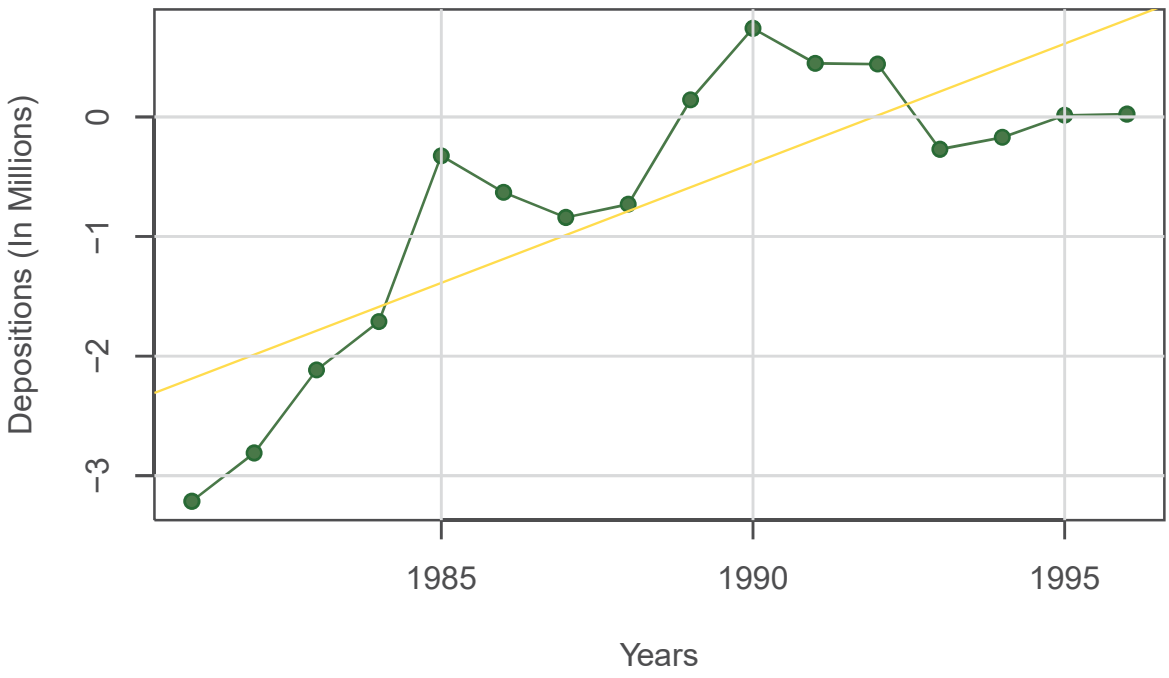
Regression Output:

1. **Coefficients**:
   - *Slope*: 0.084, statistically significant
   - *Intercept*: -165.98, statistically significant
2. **Adjusted R-square** = 0.40, the linear trend regression model is able to only explain **63%** of the variance.
3. **Residual Standard Error** : The actual change in the depositions can deviate from the regression line by 0.4598
4. **F_statistic**: It shows that there is a relationship between time and depositions

4.1.1 Plotting the Linear Trend Line

```
plot(t.eggs, type = 'o', ylab = 'Depositions (In Millions)', xlab = 'Years',
     main = 'Figure 5: Fitted Linear Trend to the Egg Depositon Data',
     pch = 19, col = 'darkgreen', lw = 1.8)
abline(model1, col = 'gold') #Adding the trend line
grid()
```

Figure 5: Fitted Linear Trend to the Egg Depositon Data



As per fig. 5, - Distance between line and points is at optimal level. It captures the trend but it doesnt capture auto correlation of succeeding points. - The regression line is also not able to capture change in variance.

4.2 Quadratic Trend Model

Now, fitting a quadratic trend model to the data.

```
t <- time(t.eggs)   #Firt variable
t2 <- t^2             #t-square as the second variable
model2 = lm(t.eggs ~  t + t2)
summary(model2)
```

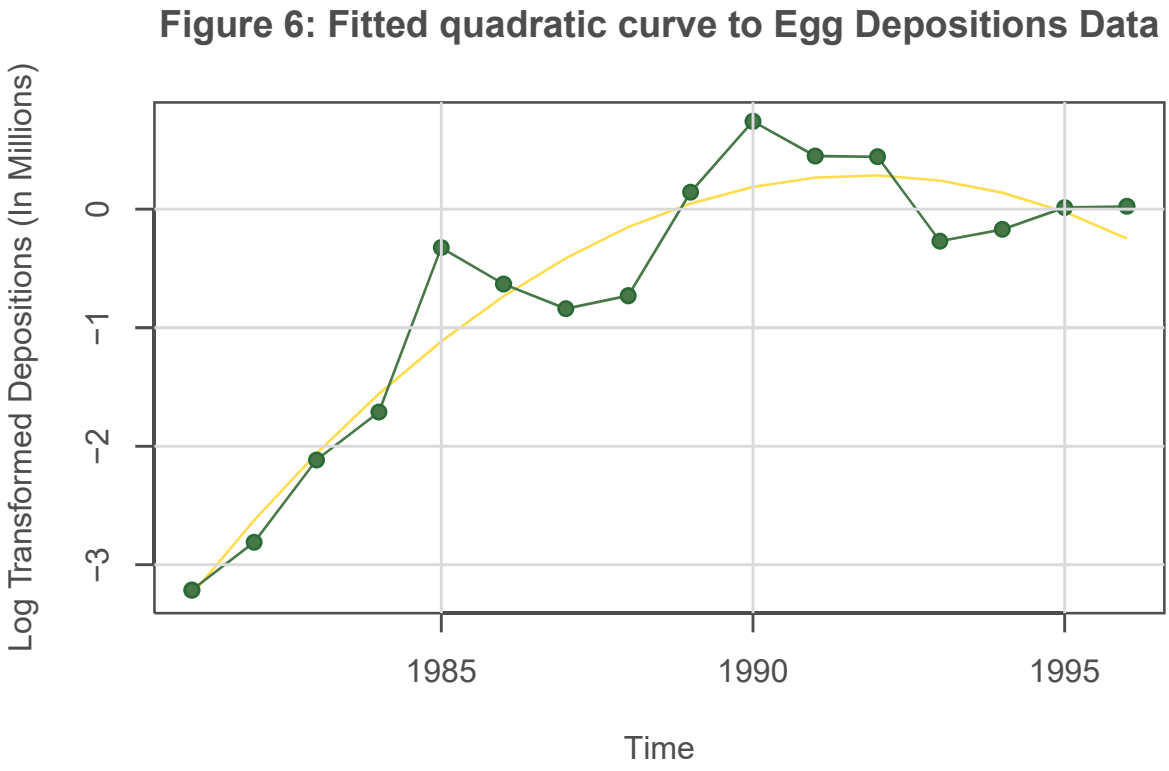
```
##
## Call:
## lm(formula = t.eggs ~ t + t2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57869 -0.21678  0.03634  0.16323  0.79064
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.202e+05  2.064e+04  -5.825 5.93e-05 ***
## t            1.207e+02  2.076e+01   5.815 6.02e-05 ***
## t2          -3.030e-02  5.220e-03  -5.806 6.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3945 on 13 degrees of freedom
## Multiple R-squared:  0.9031, Adjusted R-squared:  0.8882
## F-statistic: 60.56 on 2 and 13 DF,  p-value: 2.581e-07
```

Regression Output:

1. **Coefficients**:
   - *Intercept* : -4.647e+04, statistically significant
   - t: 4.665e+01, statistically significant
   - t^2: -1.171e-02, statistically significant
2. **Adjusted R-square** = 0.73, the linear trend regression model is able to explain **89%** of the variance.
3. **Residual Standard Error** : The actual change in the depositions can deviate from the regression line by 0.4092
4. **F_statistic**: It shows that there is a relationship between time and change in ozone layer thickness

4.2.1 Plotting the Quadratic Trend Line

```
plot(ts(fitted(model2), frequency = 1, start = c(1981,1)), ylim =
       c(min(c(fitted(model2),as.vector(t.eggs))),
         max(c(fitted(model2),as.vector(t.eggs)))),
     ylab='Log Transformed Depositions (In Millions)' ,
     main = "Figure 6: Fitted quadratic curve to Egg Depositions Data", col = 'gold')
lines(t.eggs,type="o", col = 'darkgreen', pch = 19, lw = 1.8)
grid()
```



As per fig. 6, - Distance between line and points is at optimal level. It captures the trend better than the linear trend model.

4.3 Harmonic Trend

As there was *no cyclic trend* or *seasonality* in the series, harmonic model will not be used.

4.4 Diagnostic Checking

4.4.1 Time-Series Plot of Residuals

```
par(mfrow = c(1,2))

plot(y=rstudent(model1),x=as.vector(time(t.eggs)), xlab="Time",
     ylab="Standardized Residuals",type='o',
     main = "Figure 7: Linear Model",
     col = 'darkblue', pch = 20, lwd = 2)
abline(h=0, col = 'red')
grid()

plot(y=rstudent(model2),x=as.vector(time(t.eggs)), xlab="Time",
     ylab="Standardized Residuals",type='o',
     main = "Figure 8: Quadratic Model",
     col = 'darkblue', pch = 20, lwd = 2)
abline(h=0, col = 'red')
grid()
```

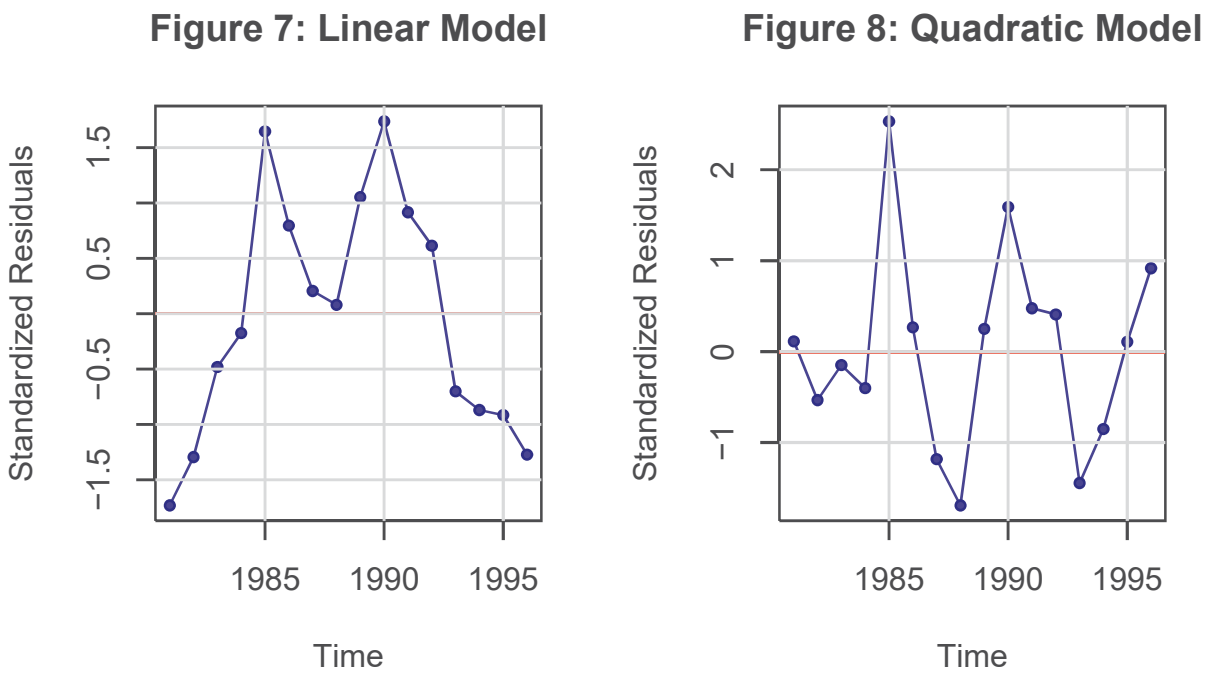


Figure 7: Linear Model

- A *trend* is observed in the standardised residual plot. This suggests that the residuals are not a true stochastic component and the linear trend model shall not be used.

Figure 8: Quadratic Model

- Comparing to the linear trend model, there is no trend in the plot. However, there are departures from the randomness and a pattern is observed.

4.4.2 Plotting Residuals Vs Fitted values

```
par(mfrow = c(1,2))

plot(y=rstudent(model1),x=as.vector(fitted(model1)),
     xlab='Fitted Trend Values', ylab='Standardized Residuals',
     type='p', main = "Figure 9: Linear Model",
     pch = 20, col = 'darkgreen')
abline(h=0, col = "gold", lty = 2, lwd = 2)
grid(lwd = 1.5)

plot(y=rstudent(model2),x=as.vector(fitted(model2)),
     xlab='Fitted Trend Values', ylab='Standardized Residuals',
     type='p', main = "Figure 10: Quadratic Model",
     col = 'darkgreen', pch = 20)
abline(h=0, col = 'gold', lty = 2, lwd = 2)
grid(lwd = 1.5)
```

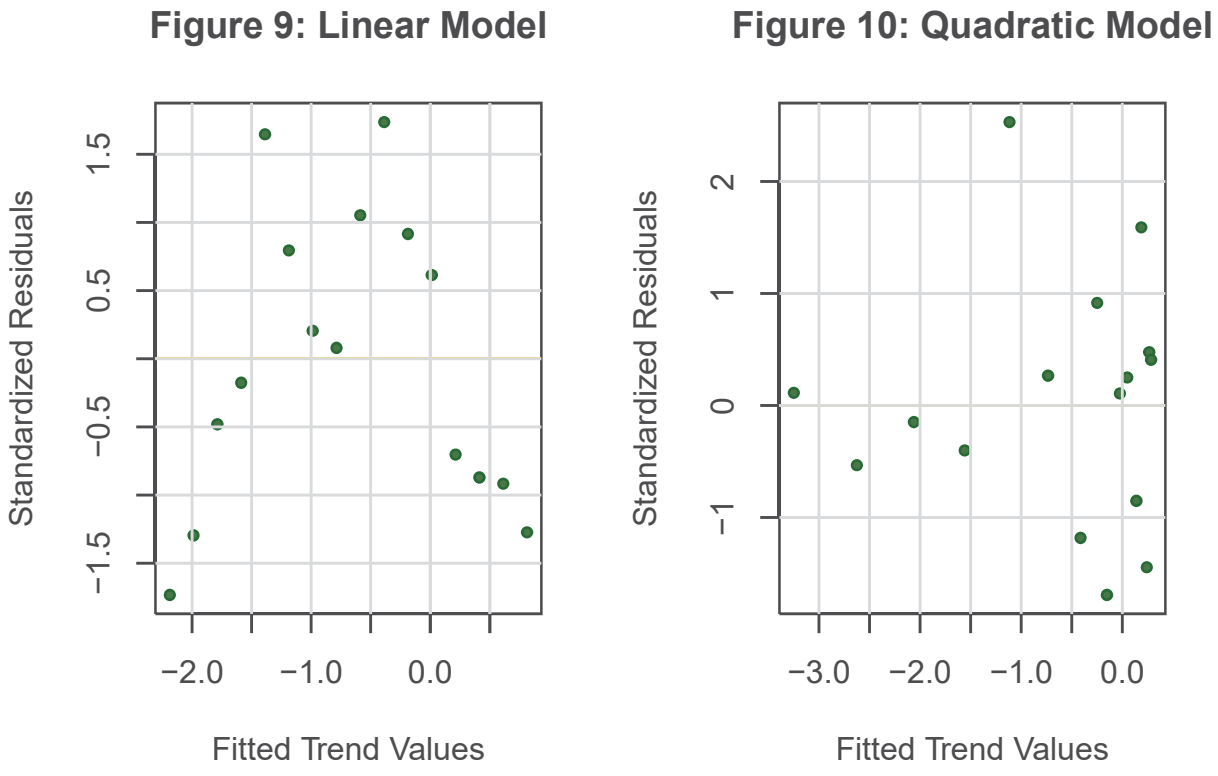


Figure 9: Linear Model

- The distribution is not random, there is a trend in distribution.

Figure 10: Quadratic Model

- Unlike white noise more residuals are observed with fitted values close to 0.

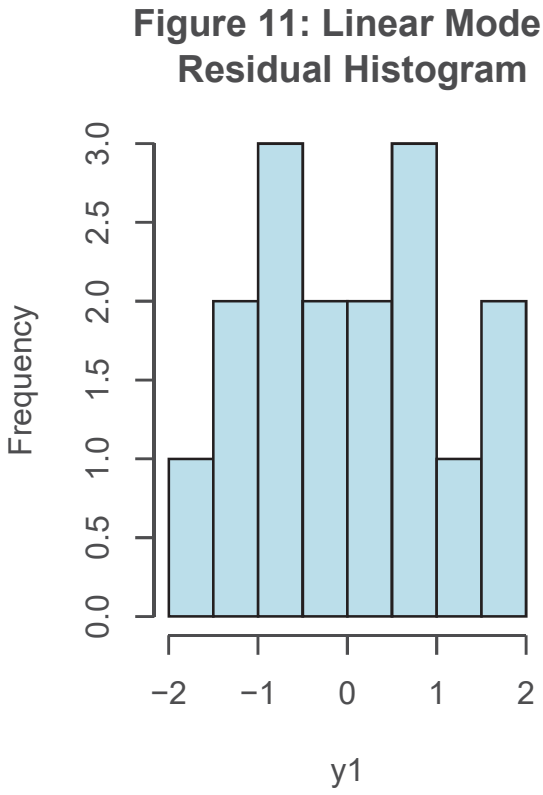
4.4.3 Normality of Residuals

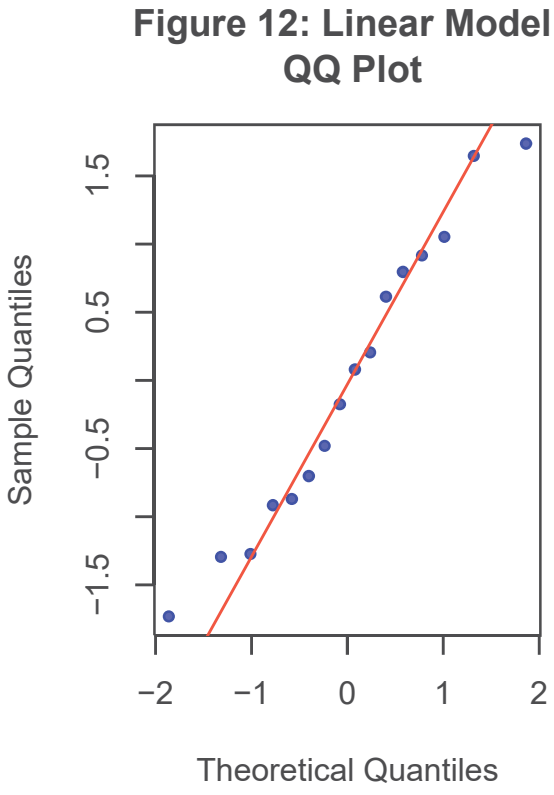
- Histograms can be used to visualise the distribution of the standardised residuals
- Q-Q plot is used to present the normality assumption of the residuals
- Hypothesis test, Shapiro-Wilk will be used to check the same.

Visualisation

```
par(mfrow=c(1, 2))
y1 = rstudent(model1) #Linear Model Standardised Residuals
y2 = rstudent(model2) #Quadratic Model Standardised Residuals
hist(y1, main = "Figure 11: Linear Model \n Residual Histogram", col = 'lightblue')

qqnorm(y1, main = 'Figure 12: Linear Model \n QQ Plot', col = 'blue', pch = 20)
qqline(y1, col = 2, lwd = 1, lty = 2)
```





- As per fig. 11, a symmetric ditribution around 0 is expected in the histogram for normal distribution, but, the same cannot be seen here.
- It can be seen from the fig. 12 that there are departures from the reference line. Therefore, it seems that the residuals are not distributed normally.

```
par(mfrow=c(1, 2))
hist(y2, main = "Figure 13: Quadratic Model \n Residual Histogram", col = 'lightblue')

qqnorm(y2, main = 'Figure 14: Quadratic Model \n QQ Plot', col = 'blue', pch = 20)
qqline(y2, col = 2, lwd = 1, lty = 2)
```

Figure 13: Quadratic Model Residual Histogram

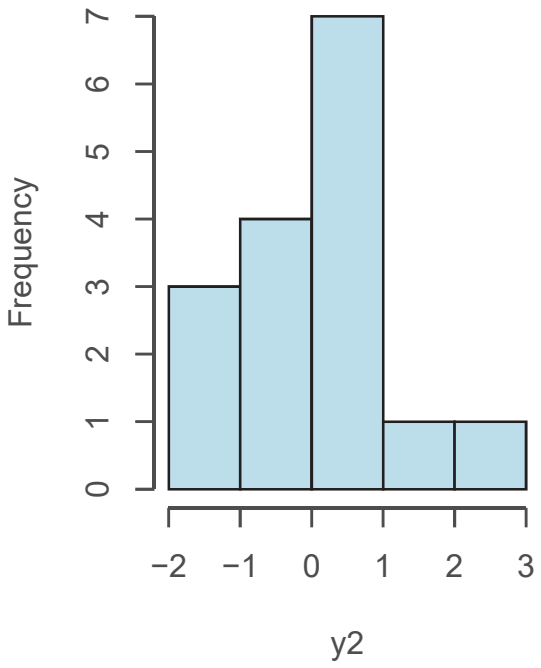
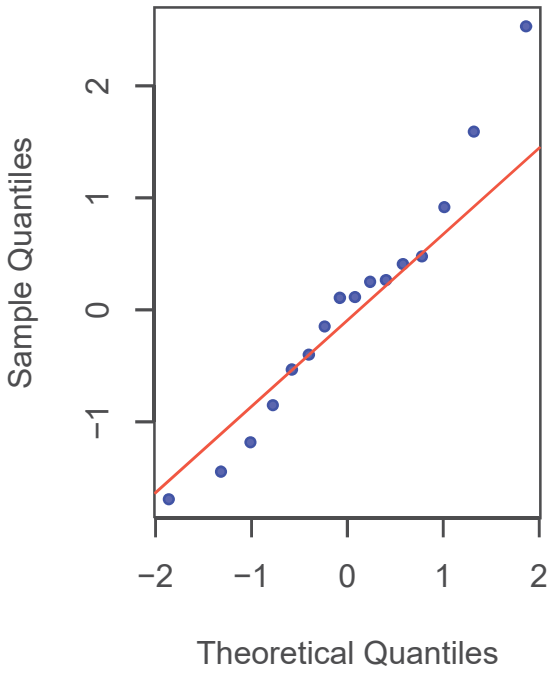


Figure 14: Quadratic Model QQ Plot



- As per fig. 13, Same as the residuals of the linear trend model, normal distribution cannot be observed in Histogram.
- As per fig. 14, there are departures from the normality reference line.

Hypothesis Testing

Hypothesis testing is used to check the normality of the series.

```
shapiro.test(rstudent(model1)) # Linear Model Hypothesis Testing
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstudent(model1)
## W = 0.95936, p-value = 0.6503
```

As the *p-value* is greater than the significance level of 0.05, we *cannot reject* the null hypothesis that the residuals are from the population which is normally distributed.

```
shapiro.test(rstudent(model2)) # Quadratic Model Hypothesis Testing
```

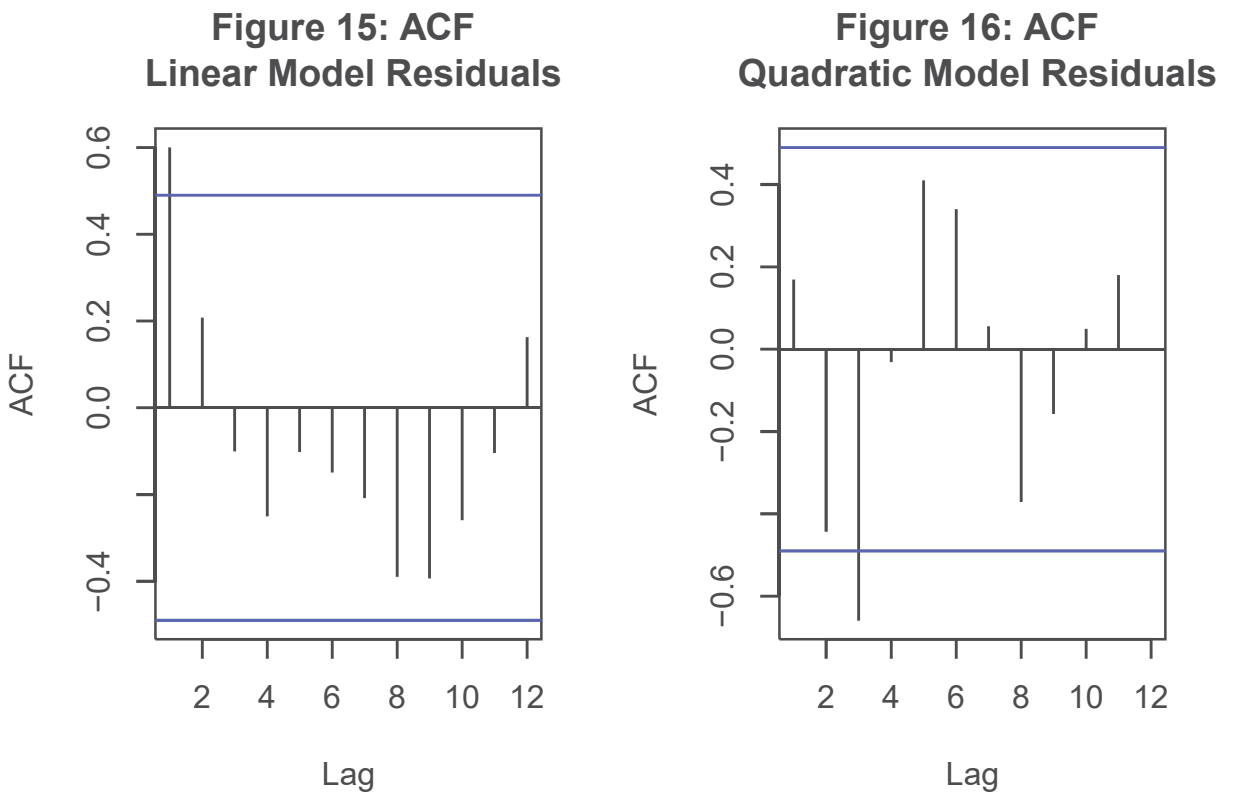
```
##
##  Shapiro-Wilk normality test
##
## data:  rstudent(model2)
## W = 0.96331, p-value = 0.7221
```

As the p-value is greater than 0.05, we cannot reject the hypthesis that the residuals are from the population which has a normal distribution.

4.4.4 Auto - Correlation Function

```
par(mfrow=c(1,2))

acf(rstudent(model1), main = "Figure 15: ACF \n Linear Model Residuals")
acf(rstudent(model2), main = "Figure 16: ACF \n Quadratic Model Residuals")
```



As per fig. 15 & fig. 16, that there are still correlation values higher than the confidence bound which is not expected in white noise.

4.5 Conclusion

Through residual analysis, it is observed that there is still correlation in the residual values. Residuals are not normally distributed in both linear & quadratic trend model. As, the trend in the time-series was not accounted for before modeling, stochastic component is not truely a white noise.

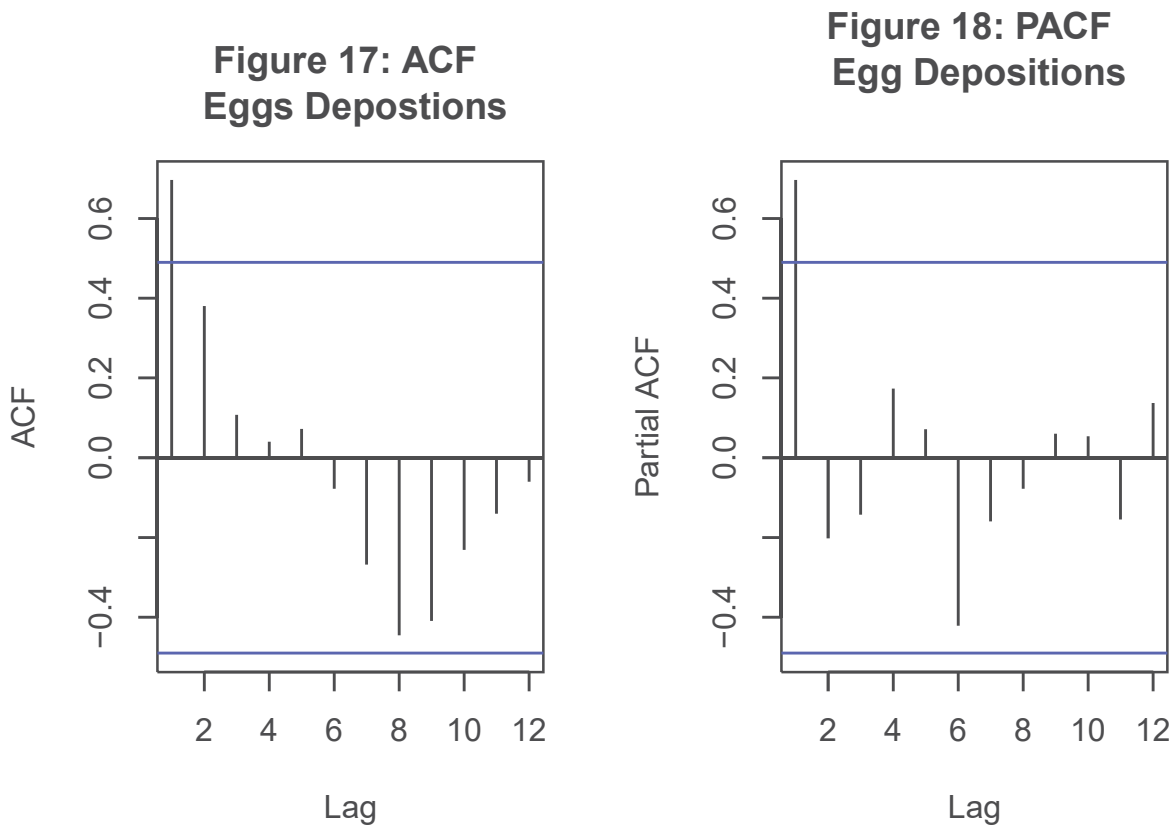
Therefore, from here we will depart from the regression approach and will explore *time-series models* below.

5 ARIMA MODELING

To begin with ARIMA models, we first have to deal with trend and non-stationarity in the series.

5.1 Trend in the Series

```
par(mfrow = c(1,2))
acf(eggs.ts, main = "Figure 17: ACF \n Eggs Depostions")
pacf(eggs.ts, main = "Figure 18: PACF \n Egg Depositions")
```



As per fig. 17 & fig. 18, we can confirm the following:

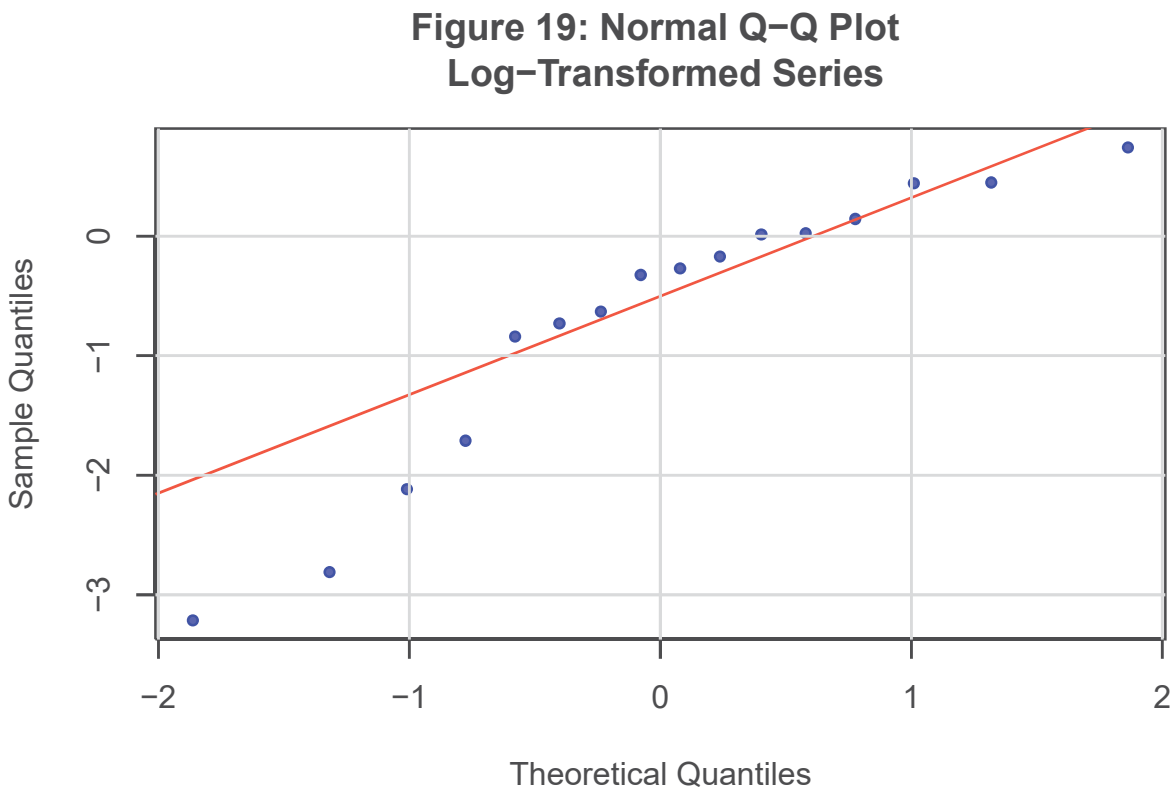
- Highly signifcant lag in the ACF plot
- There is a wave pattern in the ACF plot, we can say that the series is auto-regressive
- Highly significant lag in PACF plot suggests AR(1)

But the series also has an upward trend, so we will remove non-stationarity from the series and then explore the possisbility of AR(1) model.

5.1.1 De-Trending the Series

Following from the observation in `section 3.4`, we will try to achieve normality in the series.

```
qqnorm(t.eggs, col='blue', pch=20, main="Figure 19: Normal Q-Q Plot \n Log-Transformed Series")
qqline(t.eggs, col=2, lwd=1.8)
grid()
```



As per fig. 19, the log transformation used in trend modeling does not help in achieving normal distribution in the series.

We will use $Box - CoxTransformationSearch$ to achieve normal distribution.

```
transformation = BoxCoxSearch((eggs.ts))
```

```
## [1] "------------- Shapiro-Francia Test -------------"
##
##   Box-Cox power transformation
## ---------------------------------------------------------------
##   data : y
##
##   lambda.hat : 0.6
##
##
##   Shapiro-Francia normality test for transformed data (alpha = 0.05)
## ---------------------------------------------------------------
##
##   statistic  : 0.978985
##   p.value    : 0.9066446
##
##   Result     : Transformed data are normal.
## ---------------------------------------------------------------
##
## [1] "No results for this test!"
```

```
## [1] "------------- Shapiro-Wilk  Test -------------"
##
##   Box-Cox power transformation
## ---------------------------------------------------------------
##   data : y
##
##   lambda.hat : 0.61
##
##
##   Shapiro-Wilk normality test for transformed data (alpha = 0.05)
## ---------------------------------------------------------------
##
##   statistic  : 0.9682362
##   p.value    : 0.809291
##
##   Result     : Transformed data are normal.
## ---------------------------------------------------------------
##
## [1] "No results for this test!"
## [1] "------------- Anderson-Darling Test -------------"
##
##   Box-Cox power transformation
## ---------------------------------------------------------------
##   data : y
##
##   lambda.hat : 0.66
##
##
##   Anderson-Darling normality test for transformed data (alpha = 0.05)
## ---------------------------------------------------------------
##
##   statistic  : 0.1851461
##   p.value    : 0.8908249
##
##   Result     : Transformed data are normal.
## ---------------------------------------------------------------
##
## [1] "No results for this test!"
## [1] "------------- Cramer-von Mises Test -------------"
##
##   Box-Cox power transformation
## ---------------------------------------------------------------
##   data : y
##
##   lambda.hat : 0.7
##
##
##   Cramer-von Mises normality test for transformed data (alpha = 0.05)
## ---------------------------------------------------------------
##
##   statistic  : 0.02202402
##   p.value    : 0.9397773
##
##   Result     : Transformed data are normal.
```

```
## ------------------------------------------------------------------
##
## [1] "No results for this test!"
## [1] "------------- Pearson Chi-square Test -------------"
##
##   Box-Cox power transformation
## ------------------------------------------------------------------
##   data : y
##
##   lambda.hat : 0.33
##
##
##   Pearson Chi-square normality test for transformed data (alpha = 0.05)
## ------------------------------------------------------------------
##
##   statistic  : 1.5
##   p.value    : 0.8266415
##
##   Result     : Transformed data are normal.
## ------------------------------------------------------------------
##
## [1] "No results for this test!"
## [1] "------------- Lilliefors Test -------------"
##
##   Box-Cox power transformation
## ------------------------------------------------------------------
##   data : y
##
##   lambda.hat : 0.93
##
##
##   Lilliefors normality test for transformed data (alpha = 0.05)
## ------------------------------------------------------------------
##
##   statistic  : 0.09786507
##   p.value    : 0.9498772
##
##   Result     : Transformed data are normal.
## ------------------------------------------------------------------
##
## [1] "No results for this test!"
## [1] "------------- Jarque-Bera Test -------------"
##
##   Box-Cox power transformation
## ------------------------------------------------------------------
##   data : y
##
##   lambda.hat : 0.65
##
##
##   Jarque-Bera normality test for transformed data (alpha = 0.05)
## ------------------------------------------------------------------
##
##   statistic  : 0.4975422
```

```
##   p.value    : 0.7797584
##
##   Result     : Transformed data are normal.
## ------------------------------------------------------------------------
##
## [1] "No results for this test!"
```

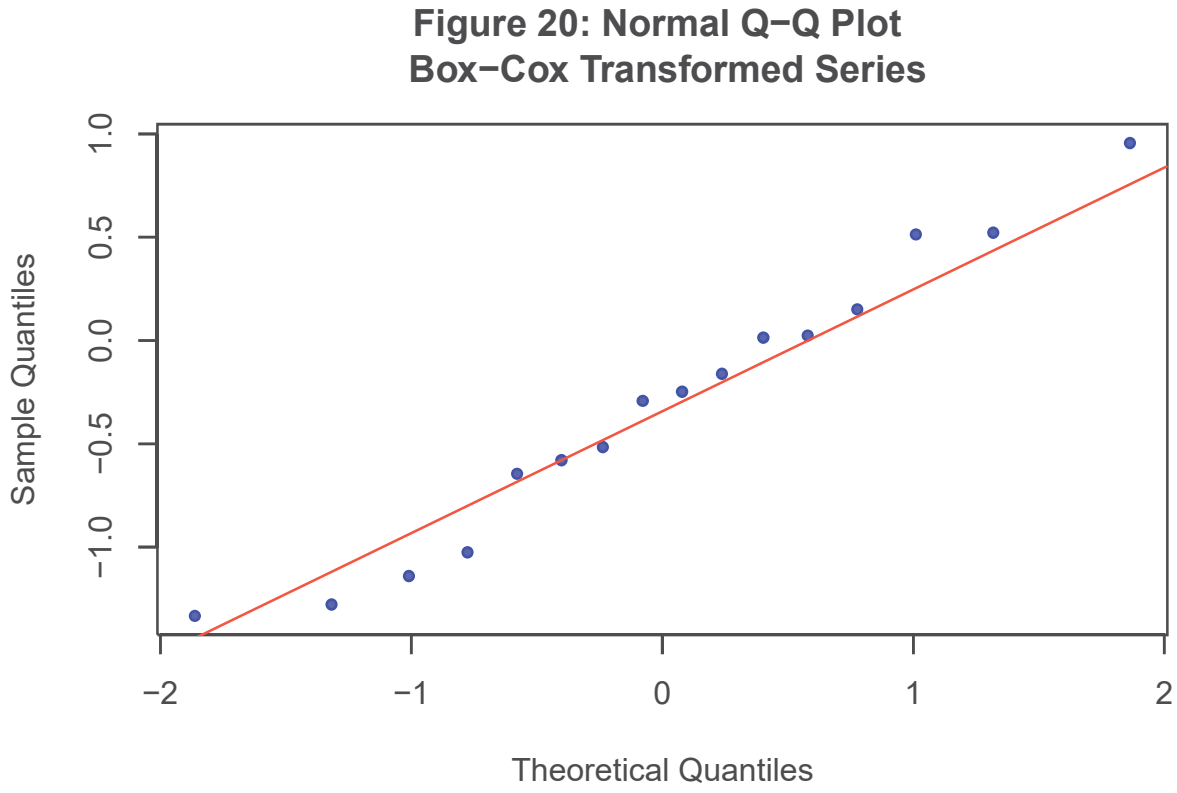
Series Transformation

From above search for transformation, we found $lambda = 0.66$ and this will be used to transform the dataset.

```
lambda = 0.66
BC.eggs = ((eggs.ts^lambda)-1)/lambda
```

Visualisation of Transformed Data

```
qqnorm(BC.eggs, col='blue', pch=20, main = "Figure 20: Normal Q-Q Plot \n Box-Cox Transformed Series")
qqline(BC.eggs, col = 2, lwd = 1, lty = 2)
```



As per fig. 20, The normality of distribution is considerably improved by Box-Cox Transformation.

5.2 Stationarity

To see if the series is still non-stationary after the transformation, we will apply $Augmentd Dickey - Fuller Test$.

```
ar(diff(BC.eggs))
```

```
##
## Call:
## ar(x = diff(BC.eggs))
##
##
## Order selected 0  sigma^2 estimated as  0.1792
```

Order obtained from the test is 0.

```
adfTest(BC.eggs, lags = 0, title = NULL, description = NULL, type="ct")
```

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 0
##   STATISTIC:
##     Dickey-Fuller: -1.6185
##   P VALUE:
##     0.7177
##
## Description:
##  Sun May 06 22:31:44 2018 by user: vishe
```

We conclude that the series is still non-stationary at 5% significance level.

We will apply $first - difference$ and test again.

```
eggs.diff = diff(BC.eggs)
ar(diff(eggs.diff))
```

```
##
## Call:
## ar(x = diff(eggs.diff))
##
## Coefficients:
##       1        2        3        4
## -0.7559  -0.5074  -0.7051  -0.5589
##
## Order selected 4  sigma^2 estimated as  0.2492
```

Order of the Test obtained is 4.

Also, we will check the $ACF\ and\ PACF$ plot to check the significant lags.

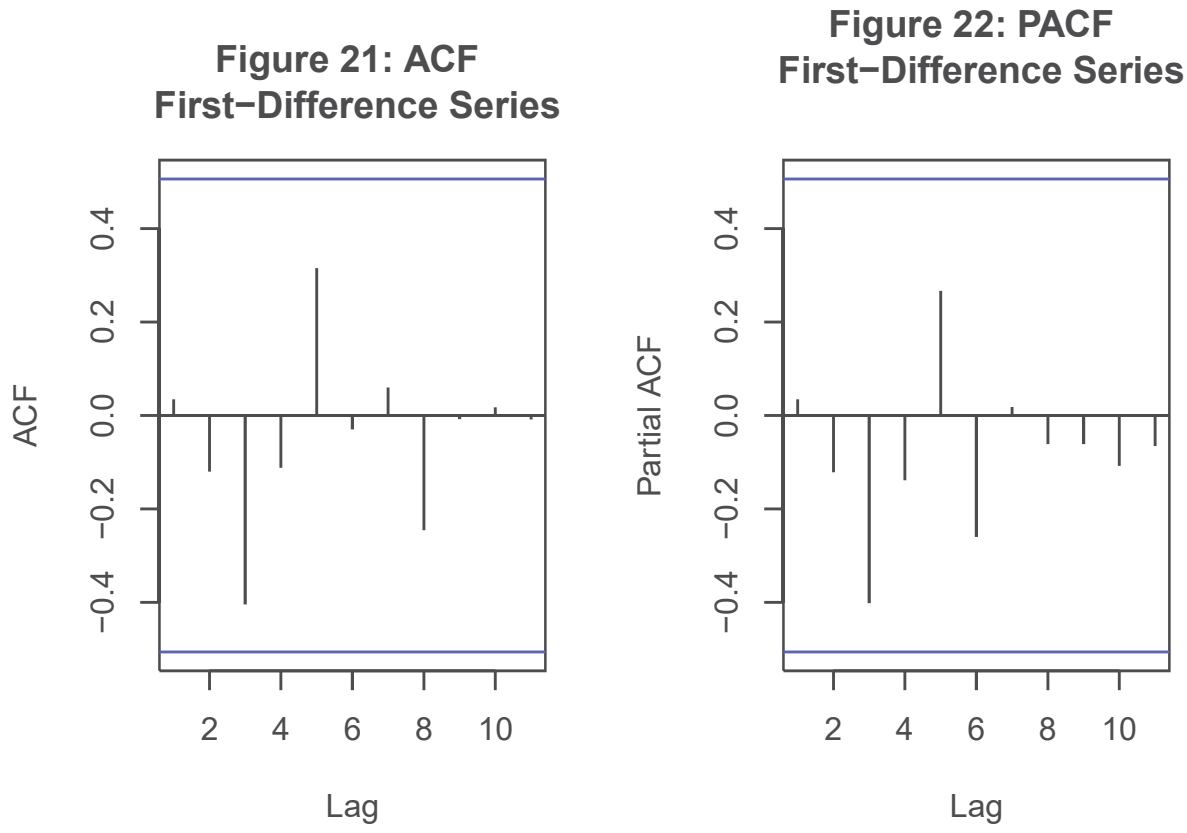
```
adfTest(eggs.diff, lags = 4, title = NULL, description = NULL, type="c")
```

```
##
## Title:
##  Augmented Dickey-Fuller Test
```

```
##
## Test Results:
##   PARAMETER:
##     Lag Order: 4
##   STATISTIC:
##     Dickey-Fuller: -0.4216
##   P VALUE:
##     0.8817
##
## Description:
##  Sun May 06 22:31:44 2018 by user: vishe
```

We can see above that $Augmented Dickey - Fuller Test$ says, the series is non-stationary. We will also observe the ACF and PACF plots after differencing.

```
par(mfrow = c(1,2))
acf(eggs.diff, main="Figure 21: ACF \n First-Difference Series")
pacf(eggs.diff, main = "Figure 22: PACF \n First-Difference Series")
```



As per fig. 21 & fig. 22, we can see that there are no significant lags.

We will proceed with the $second - difference$ and test again.

```
eggs.diff2 = diff(eggs.diff)
ar(diff(eggs.diff2))
```

```
##
## Call:
## ar(x = diff(eggs.diff2))
```

```
##
## Coefficients:
##       1        2        3        4
## -1.0685  -0.7678  -0.8342  -0.6317
##
## Order selected 4  sigma^2 estimated as  0.4952
```

```
adfTest(eggs.diff2, lags = 4, title = NULL,description = NULL, type="c")
```

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 4
##   STATISTIC:
##     Dickey-Fuller: -1.2216
##   P VALUE:
##     0.5986
##
## Description:
##  Sun May 06 22:31:44 2018 by user: vishe
```

As per the unit root test, the series is still non stationary, we will also check ACF and PACF to decide if we will proceed with third-difference.

```
par(mfrow = c(1,2))

acf(eggs.diff2, main = "Figure 23: ACF \n Second-Difference Series")
pacf(eggs.diff2, main = "Figure 24: PACF \n Second-Difference Series")
```

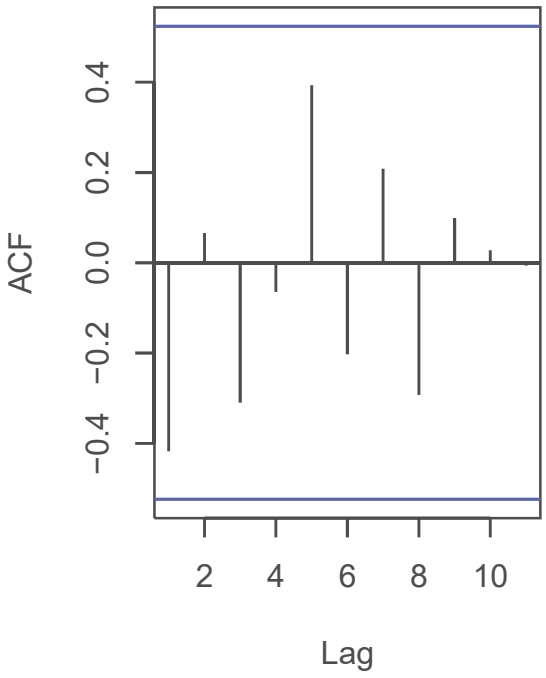


Figure 23: ACF
Second−Difference Series

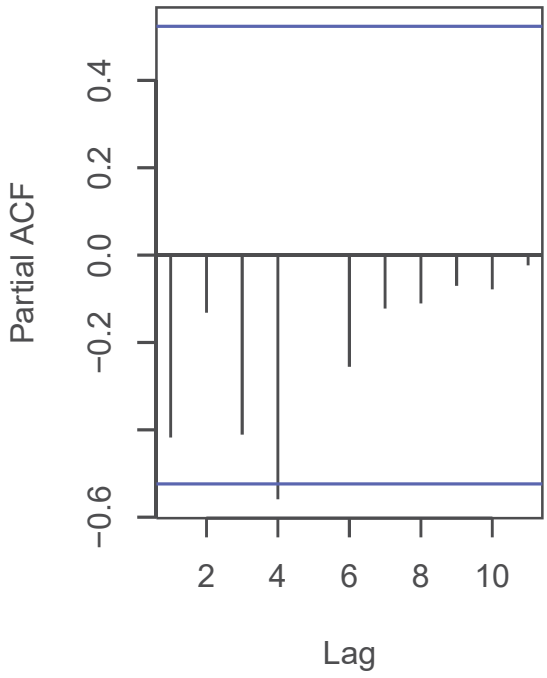


Figure 24: PACF
Second−Difference Series

As per fig. 23, there are no significant lags in the series. But in fig. 24, we can see that there is one signigicant lag.

For now, we will proceed with the $third - difference$.

```
eggs.diff3 = diff(eggs.ts, differences=3)
ar(diff(eggs.diff3))
```

```
##
## Call:
## ar(x = diff(eggs.diff3))
##
## Coefficients:
##       1        2        3        4
## -1.3417  -0.9946  -0.8473  -0.5699
##
## Order selected 4  sigma^2 estimated as  1.226
```

```
adfTest(eggs.diff3, lags = 4, title = NULL, description = NULL)
```

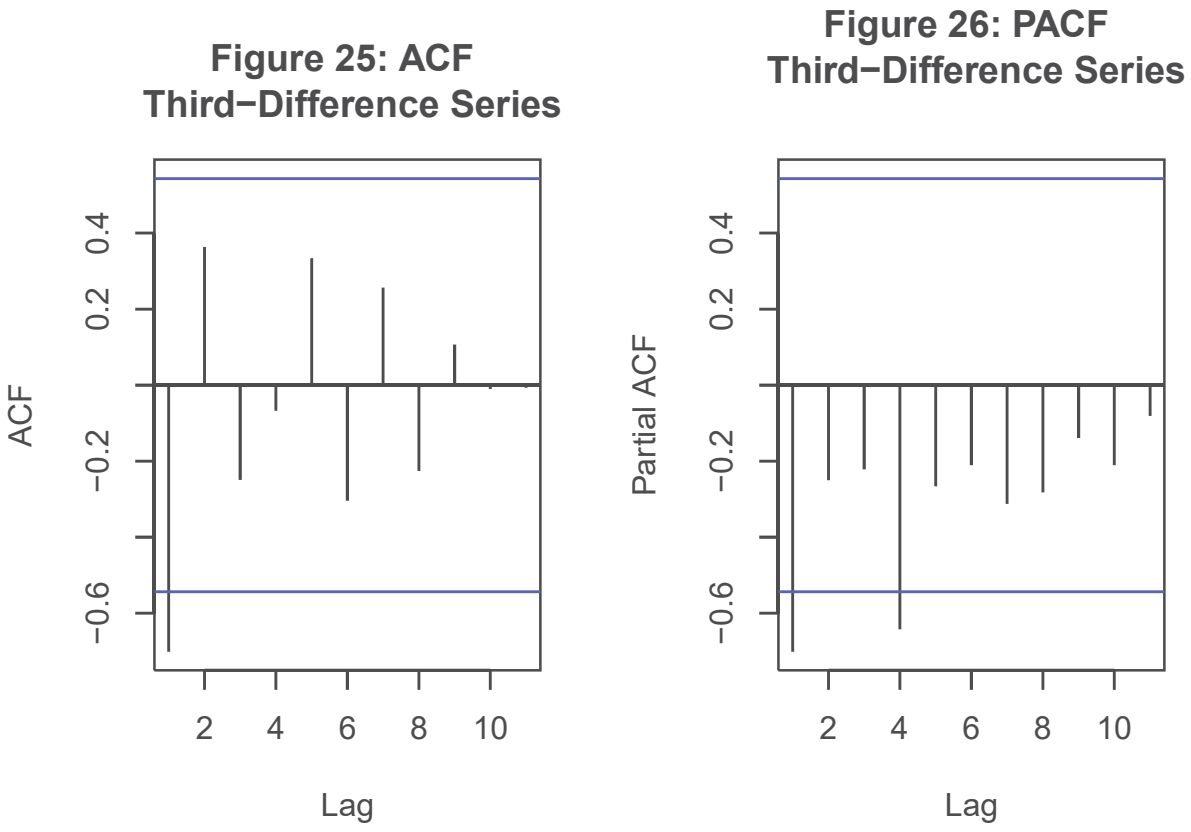
```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 4
##   STATISTIC:
```

```
##     Dickey-Fuller: -0.7284
##   P VALUE:
##     0.3767
##
## Description:
##  Sun May 06 22:31:44 2018 by user: vishe
```

As per the $ADF - Test$, the series is still non-stationary at 5% significance level.

But we will look at the ACF and PACF plots also.

```
par(mfrow=c(1,2))
acf(eggs.diff3, main="Figure 25: ACF \n Third-Difference Series")
pacf(eggs.diff3, main="Figure 26: PACF \n Third-Difference Series")
```



As per fig. 25 & fig. 26, we can see significant lags. This introduces unneccessary correlations in the series and is a sign of *overdifferencing*.

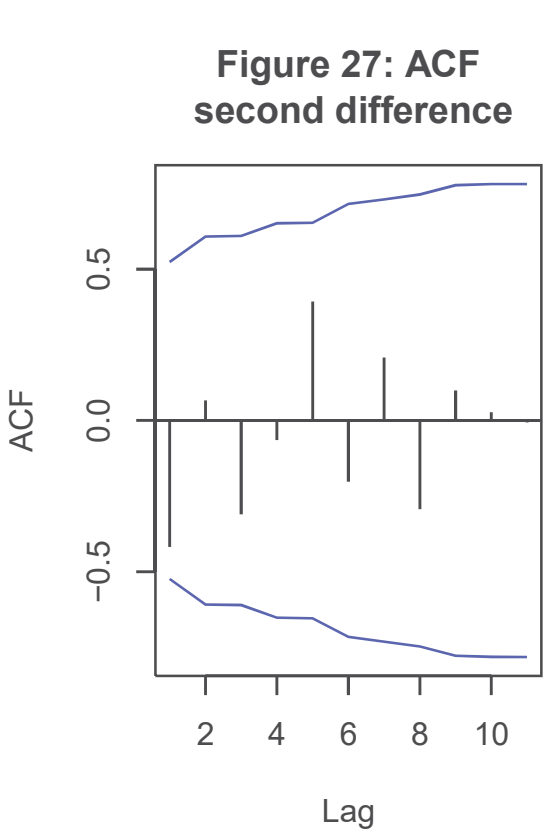
Thereofore, although unit root test suggests the series to be non-stationary, we will not perform *fourth-difference*.

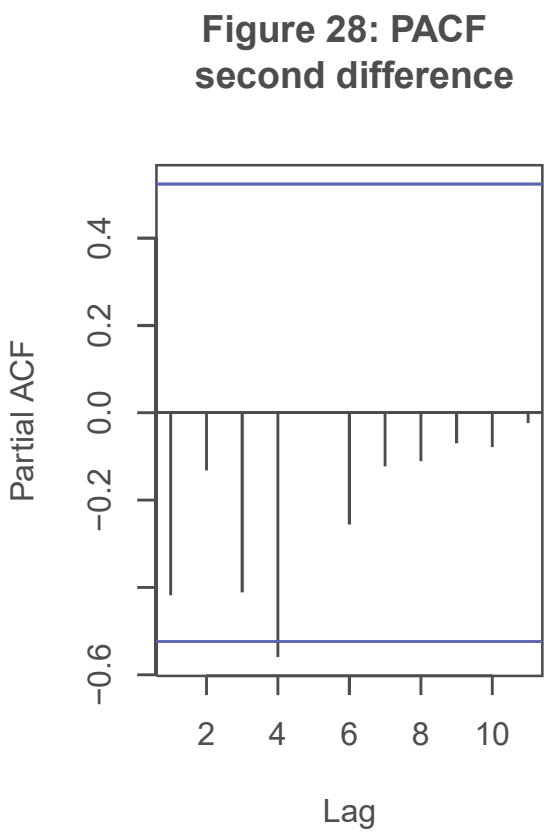
we will obtain candidate models from both second & third difference as taking MA or AR terms may compensate underdifferencing or overdifferencing. Robert Nau - Duke University

6 MODEL SELECTION

6.1 ACF and PACF

```
par(mfrow=c(1,2))
acf(eggs.diff2, ci.type='ma', main=" Figure 27: ACF \n second difference")
pacf(eggs.diff2, main="Figure 28: PACF \n second difference")
```

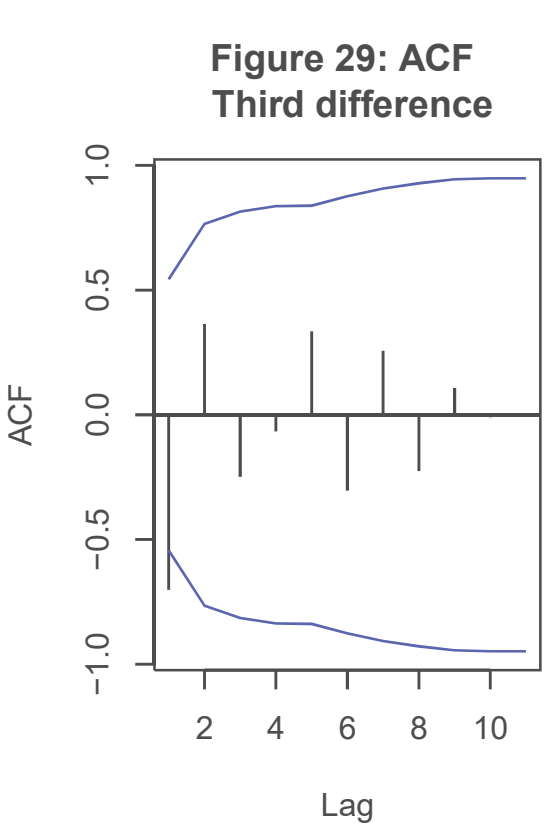


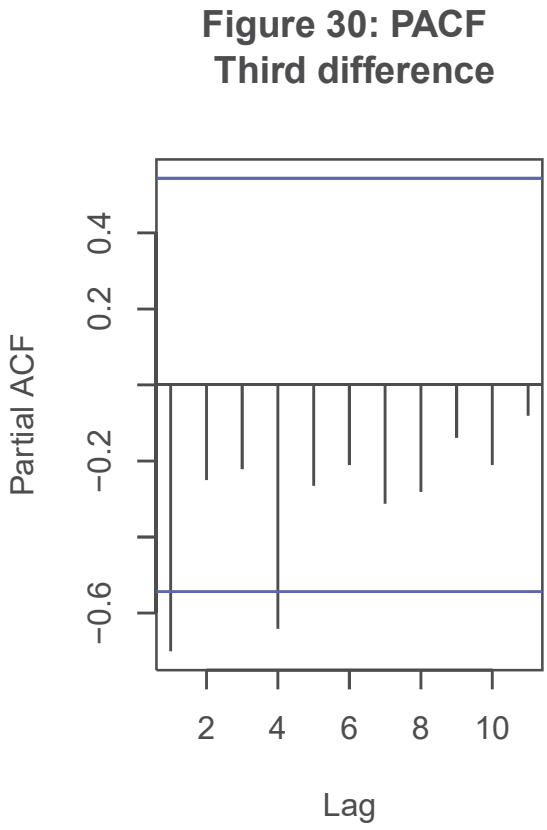


As per fig. 27 & 28, we see alternate decaying pattern in ACF and 1 significant lag in PACF.

So, a possible model from here is *{ARIMA(1,2,0)}*.

```
par(mfrow=c(1,2))
acf(eggs.diff3, ci.type='ma', main="Figure 29: ACF \n Third difference")
pacf(eggs.diff3, main="Figure 30: PACF \n Third difference")
```





As per fig. 29 & 30, we see alternate descaying pattern in ACF and 2 signigicant lags in PACF. So, a possible model from here is *{ARIMA(1,3,0), ARIMA(2,3,0)}*.

6.2 Extended Auto-Correlation Function (EACF)

```
eacf(eggs.diff2, ar.max = 3, ma.max = 2)
```

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

From the above EACF, the top-left 'o' symbol is located at AR=0 and MA=0.

Possible candidate models from here are: *{ARIMA(1,2,0), ARIMA(0,2,1)}*

```
eacf(eggs.diff3, ar.max = 3, ma.max = 2)
```

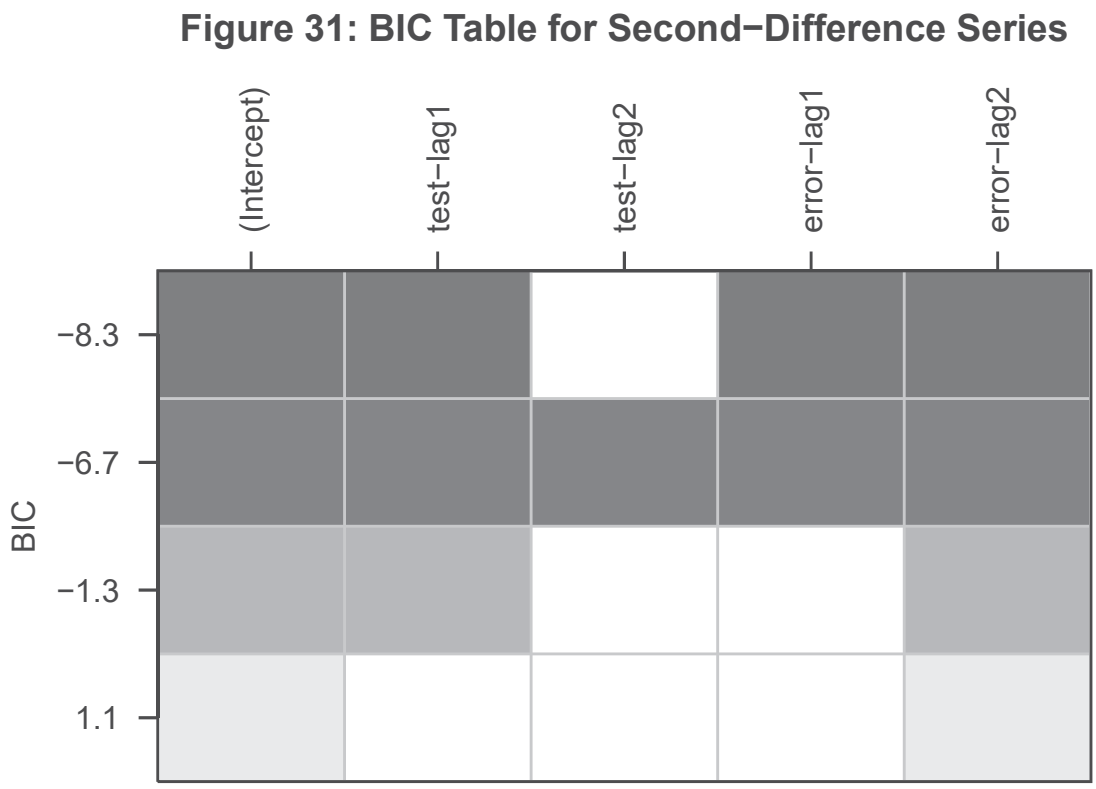
```
## AR/MA
##   0 1 2
## 0 x o o
## 1 o o o
## 2 o o o
## 3 o o o
```

From the above EACF, the top-left 'o' symbol is located at AR=0 and MA=1.

Possible candidate models from here are: *{ARIMA(0,3,1), ARIMA(1,3,1), ARIMA(1,3,0)}*

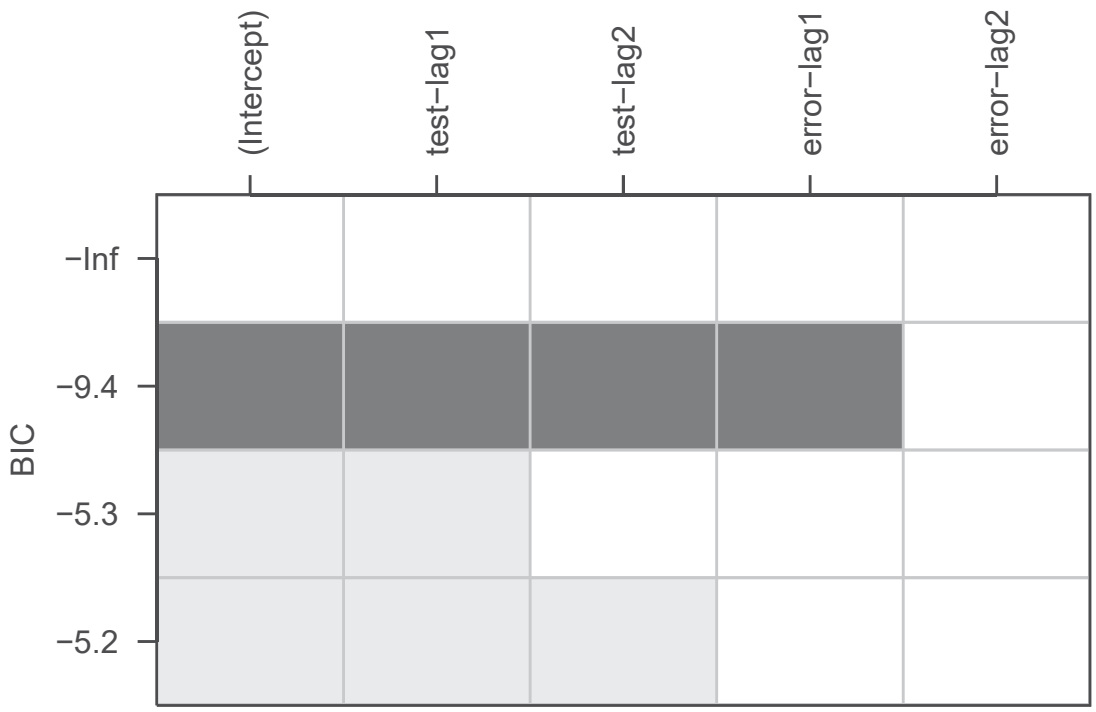
6.3 BIC Table

```
res1 = armasubsets(y=eggs.diff2,nar=2,nma=2,y.name='test',ar.method='ols')
plot(res1)
title("Figure 31: BIC Table for Second-Difference Series", line = 6)
```



As per fig. 31, we read the models *{ARIMA(0,2,2), ARIMA(1,2,2)}*.

```
res2 = armasubsets(y=eggs.diff3,nar=2,nma=2,y.name='test',ar.method='ols')
plot(res2)
title("Figure 32: BIC Table for Third-Difference Series", line = 6)
```



As per fig. 32, we read the models *{ARIMA(1,3,0), ARIMA(2,3,0)}*.

6.4 Possible Candidate Models

From `Section 6.1 to 6.3`, we get the following set of possible candidate models:

- ARIMA(0,2,1)
- ARIMA(0,2,2)
- ARIMA(1,2,0)
- ARIMA(1,2,2)
- ARIMA(0,3,1)
- ARIMA(1,3,0)
- ARIMA(2,3,0)

Now, we will proceed with the model fitting and find their parameter estimates.

7 MODEL FITTING

7.1 ARIMA(0,2,1)

```
model.021 = arima(BC.eggs,order=c(0,2,1),method='ML')
coeftest(model.021)
```

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value Pr(>|z|)
## ma1 -1.00000    0.32788 -3.0498  0.00229 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The MA coefficient of $ARIMA(0,2,1)$ is *significant* at 5% significance level.

7.2 ARIMA(0,2,2)

```
model.022 = arima(BC.eggs,order=c(0,2,2),method='ML')
coeftest(model.022)
```

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value Pr(>|z|)
## ma1 -0.89941    0.36875 -2.4391  0.01472 *
## ma2 -0.10059    0.25286 -0.3978  0.69076
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

MA2 coefficient of $ARIMA(0,2,2)$ is not *significant* at 5% significance level.

7.3 ARIMA(1,2,0)

```
model.120 = arima(BC.eggs,order=c(1,2,0),method='ML')
coeftest(model.120)
```

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value Pr(>|z|)
## ar1 -0.39278    0.23305 -1.6854   0.0919 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The AR coefficient of $ARIMA(1,2,0)$ is *significant* at 5% significance level.

7.4 ARIMA(1,2,2)

```
model.122 = arima(BC.eggs,order=c(1,2,2),method='ML')
coeftest(model.122)
```

```
##
## z test of coefficients:
##
##      Estimate Std. Error z value Pr(>|z|)
## ar1  0.029572   1.013681  0.0292   0.9767
## ma1 -0.926413   1.005368 -0.9215   0.3568
## ma2 -0.073498   0.969075 -0.0758   0.9395
```

All coefficients of $ARIMA(1,2,2)$ are not *significant* at 5% significance level.

7.5 ARIMA(0,3,1)

```
model.031 = arima(BC.eggs,order=c(0,3,1),method='ML')
coeftest(model.031)
```

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value  Pr(>|z|)
## ma1 -1.00000    0.19722 -5.0704 3.969e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

MA1 coefficient of $ARIMA(0,3,1)$ is *significant* at 5% significance level.

7.6 ARIMA(1,3,0)

```
model.130 = arima(BC.eggs,order=c(1,3,0),method='ML')
coeftest(model.130)
```

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value Pr(>|z|)
## ar1 -0.62781    0.19248 -3.2618 0.001107 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

AR1 coefficient of $ARIMA(1,3,0)$ is *significant* at 5% significance level.

7.7 ARIMA(1,3,1)

```
model.131 = arima(BC.eggs,order=c(1,3,1),method='ML')
coeftest(model.131)
```

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value  Pr(>|z|)
```

```
## ar1 -0.34974    0.24746 -1.4133    0.1576
## ma1 -0.99999    0.21801 -4.5870 4.497e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

AR1 coefficient of $ARIMA(1,3,1)$ is not *significant* at 5% significance level.

7.8 ARIMA(2,3,0)

```
model.230 = arima(BC.eggs,order=c(2,3,0),method='ML')
coeftest(model.230)
```

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value Pr(>|z|)
## ar1 -0.81995    0.26617 -3.0806 0.002066 **
## ar2 -0.26744    0.26550 -1.0073 0.313778
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

AR2 coefficient of $ARIMA(2,3,0)$ is not *significant* at 5% significance level.

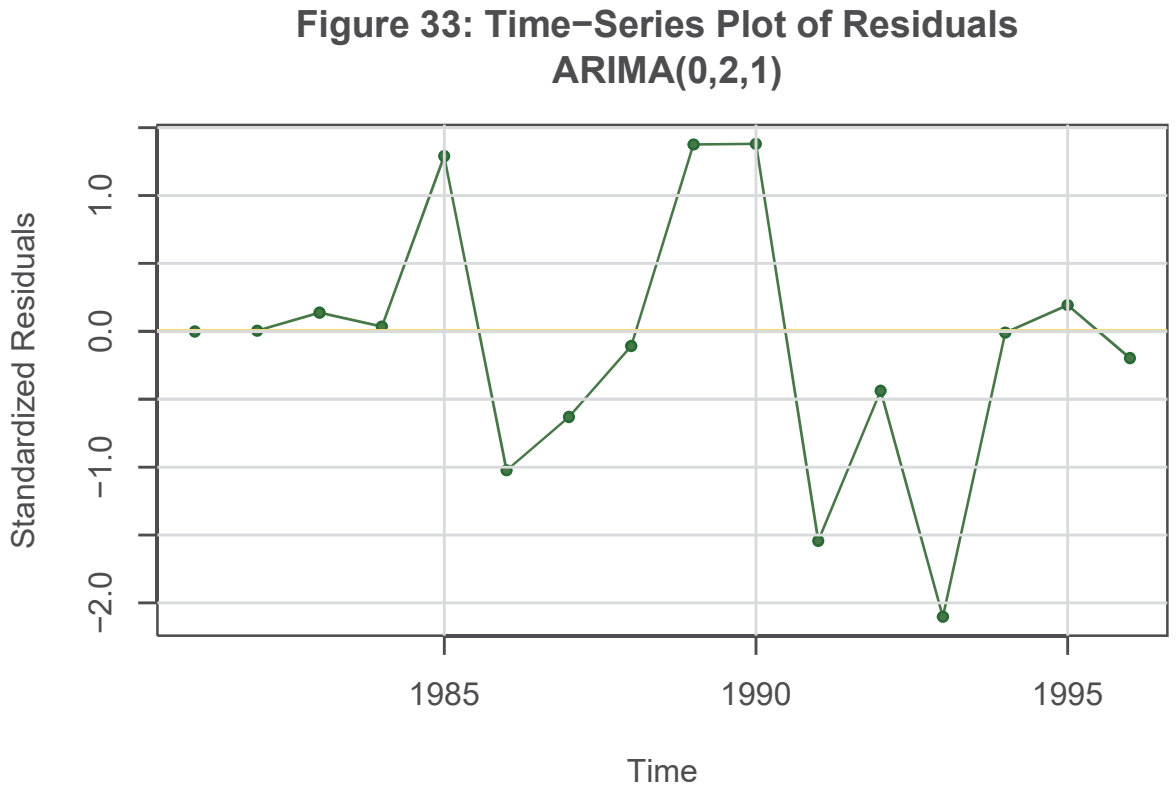
8 DIAGNOSTIC CHECKING

To perform the given activity, we will analyse *Residuals* of each model.

8.1 Residual Analysis - ARIMA(0,2,1)

8.1.1 Time-Series Plot of Residuals

```
plot(rstandard(model.021),ylab ='Standardized Residuals',type='o',main="Figure 33: Time-Series Plot of
abline(h=0, col="gold", lty=2, lw=2)
grid()
```

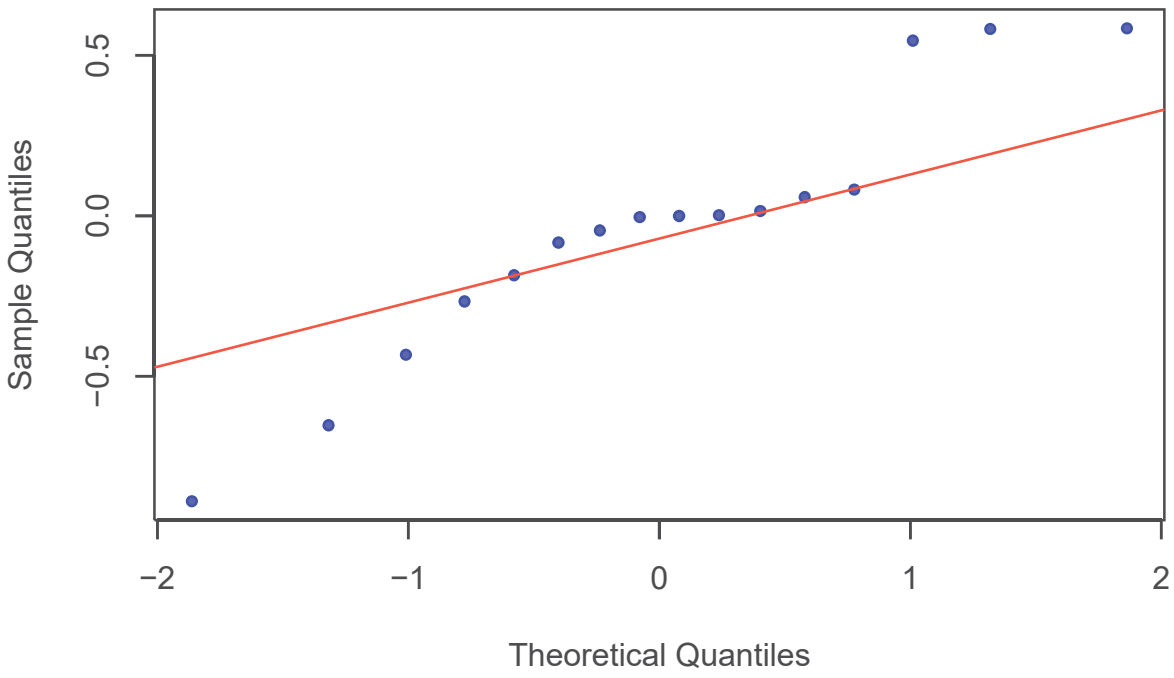


From fig. 33, we can say that there is trend and there is an obvious change in variance in the residuals

8.1.2 Normality Check of Residuals

```
qqnorm(residuals(model.021), pch = 20, col="blue", main = "Figure 34: Normal Q-Q Plot")
qqline(residuals(model.021), lty=2, col="red", lw=1.5)
```

Figure 34: Normal Q−Q Plot




```
shapiro.test(residuals(model.021))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model.021)
## W = 0.92467, p-value = 0.2005
```

Although we can see departures from the normality line in fig. 34, but, Shapiro Test gives the residuals to be normally distributed at 5% signigicane level.

8.1.3 ACF and PACF of Residuals

```
par(mfrow=c(1,2))

acf(residuals(model.021), main = "Figure 35: ACF \n Residuals of ARIMA(0,2,1)")
pacf(residuals(model.021), main = "Figure 36: PACF \n Residuals of ARIMA(0,2,1)")
```

Figure 35: ACF
Residuals of ARIMA(0,2,1)

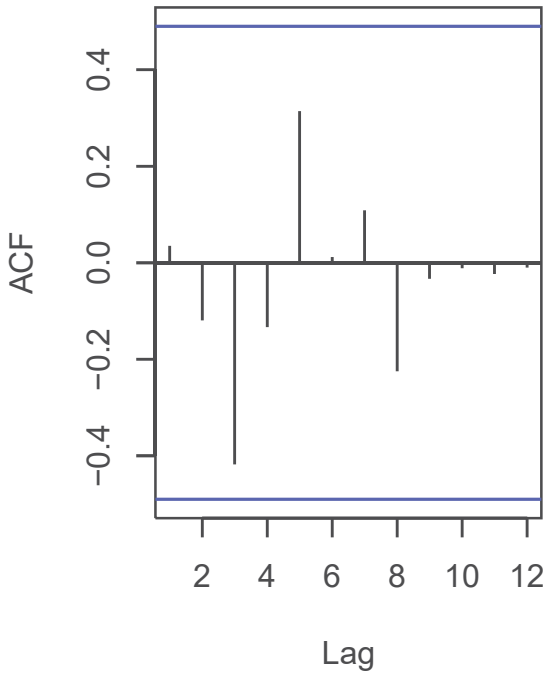
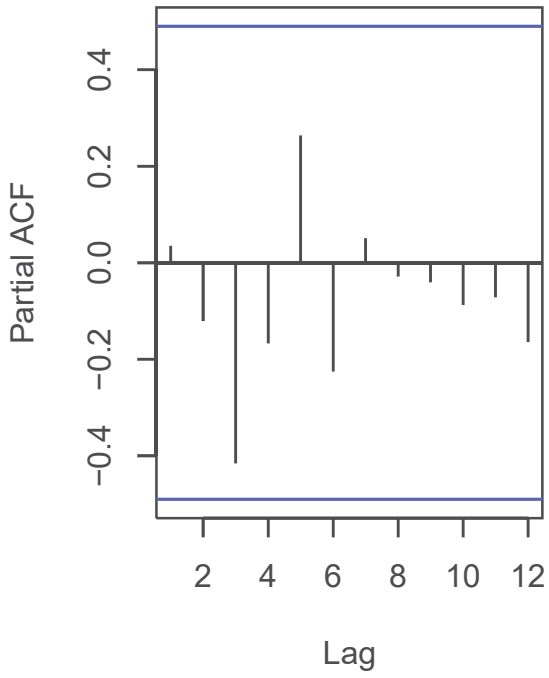


Figure 36: PACF
Residuals of ARIMA(0,2,1)



From ACF and PACF of the residual, we can conclude that the residuals constitute a white noise series as there are no highly significant correlation.

8.1.4 Box-Ljung Test

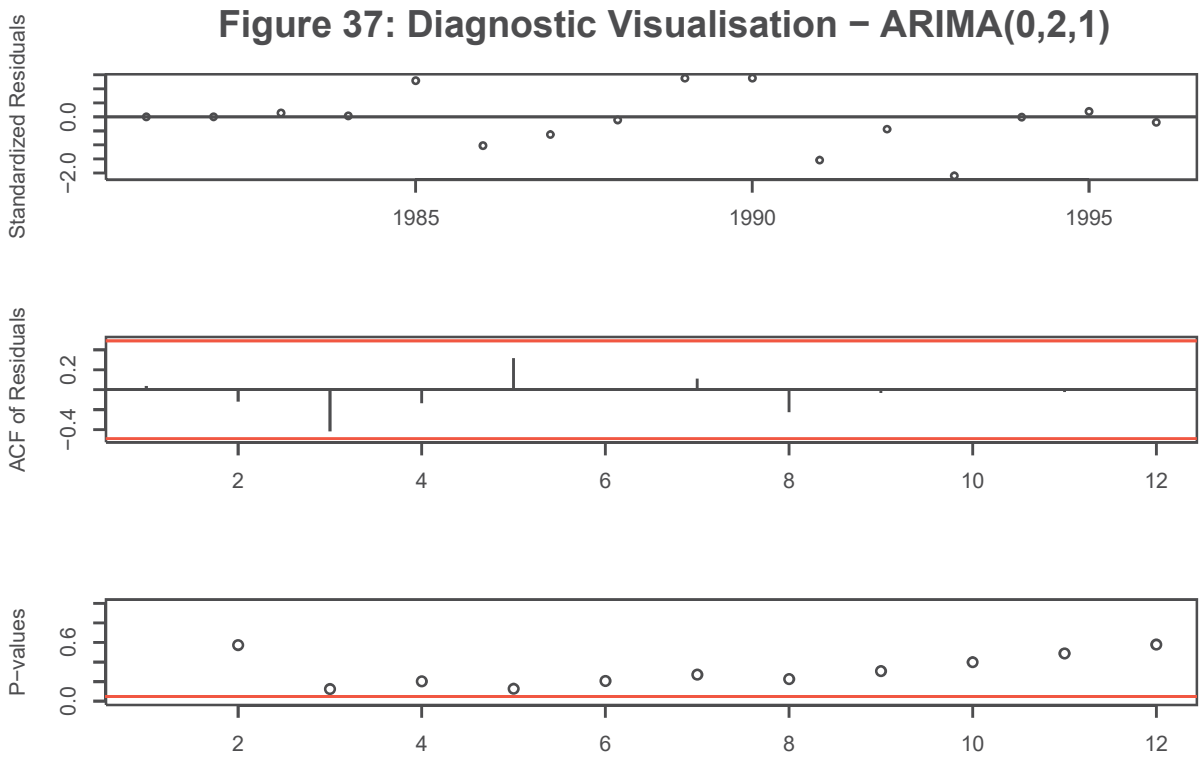
```
Box.test(residuals(model.021), lag = 12, type = "Ljung-Box", fitdf = 0)
```

```
##
##  Box-Ljung test
##
## data:  residuals(model.021)
## X-squared = 9.4716, df = 12, p-value = 0.6622
```

As seen in ACF & PACF plots, Ljung-Box Test also supports the non-existence of correlation in residuals at 5% significance level.

8.1.5 Diagnostics Visualisation

```
tsdiag(model.021,gof=12,omit.initial=F)
title("Figure 37: Diagnostic Visualisation - ARIMA(0,2,1)")
```

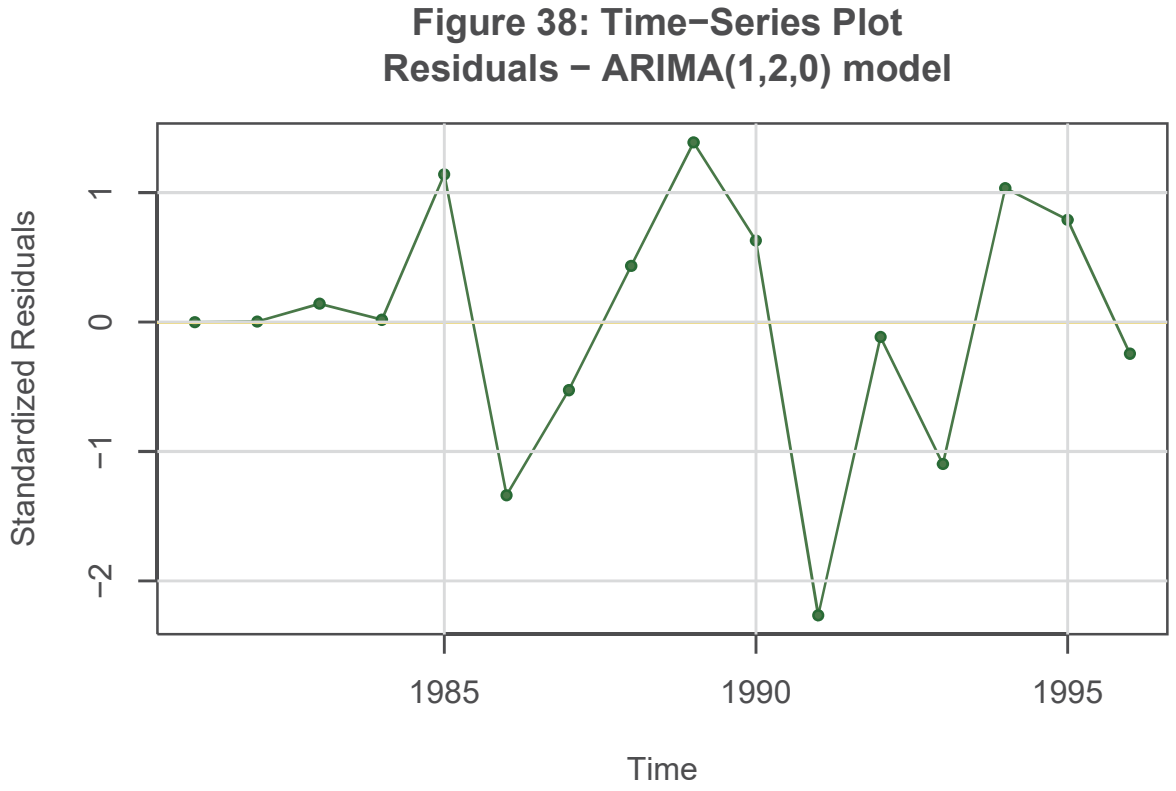


From the above diagnostic visualisation, it can be clearly seen that $p-value$ of the Ljung-Box Test statistic for the whole range are not significant. Therefore, there is no existence of correlation in residuals of ARIMA(0,2,1) model.

8.2 Residual Analysis - ARIMA(1,2,0)

8.2.1 Time-Series Plot of Residuals

```
plot(rstandard(model.120),ylab ='Standardized Residuals',type='o',main="Figure 38: Time-Series Plot \n
abline(h=0, col="gold", lty=2, lw=2)
grid()
```

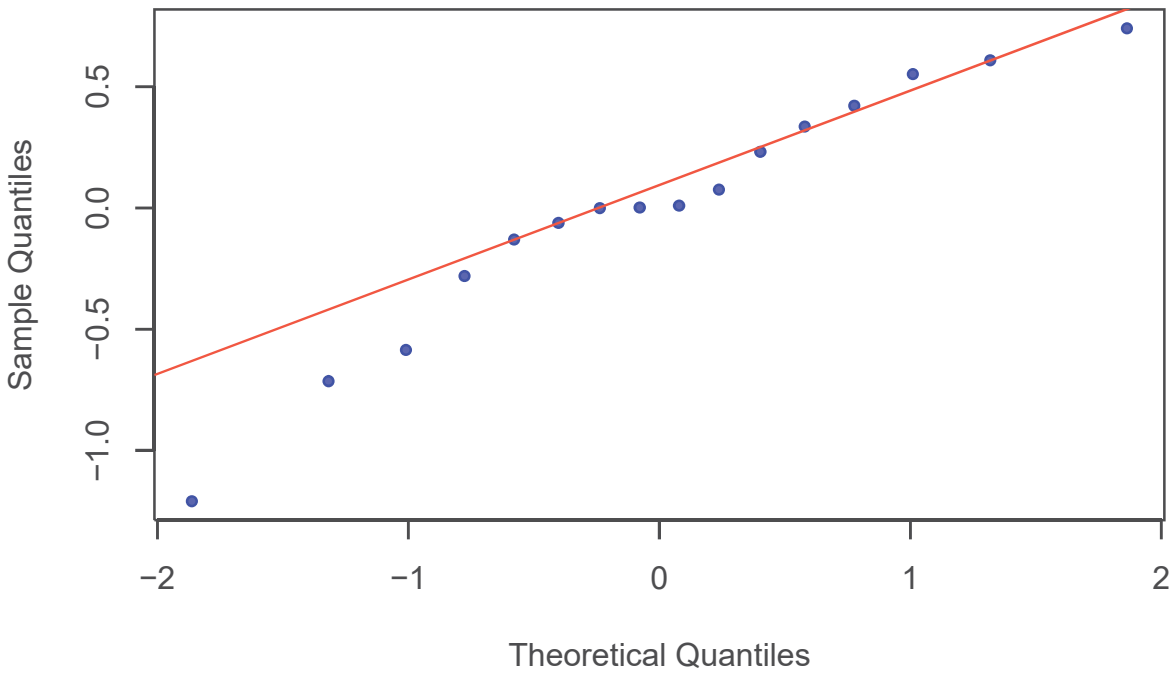


From above plot, we can say that there is no trend, but, there is an obvious change in variance in the residuals

8.2.2 Normality Check of Residuals

```
qqnorm(residuals(model.120), pch = 20, col="blue", main="Figure 39: Normal Q-Q Plot")
qqline(residuals(model.120), lty=2, col="red", lw=1.5)
```

Figure 39: Normal Q−Q Plot




```
shapiro.test(residuals(model.120))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model.120)
## W = 0.94907, p-value = 0.475
```

Although we can see departures from the normality line in fig. 39, but, Shapiro Test gives the residuals to be normally distributed at 5% signigicane level.

8.2.3 ACF and PACF of Residuals

```
par(mfrow=c(1,2))

acf(residuals(model.120), main = "Figure 40: ACF \n Residuals of ARIMA(1,2,0)")
pacf(residuals(model.120), main = "Figure 41: PACF \n Residuals of ARIMA(1,2,0)")
```

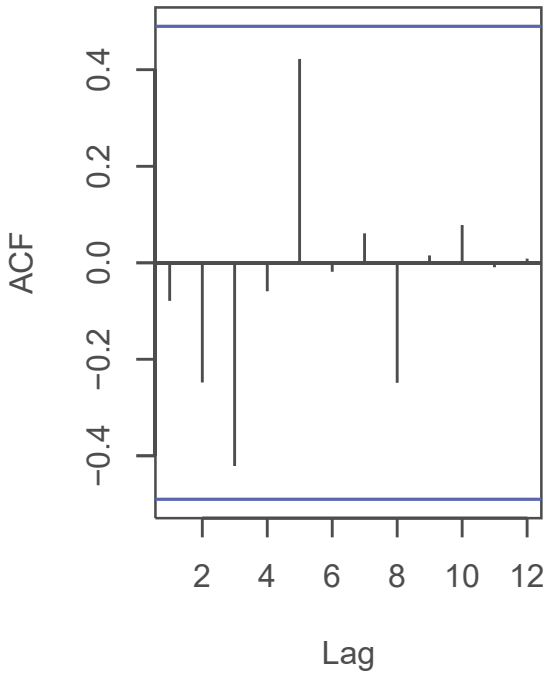


Figure 40: ACF
Residuals of ARIMA(1,2,0)

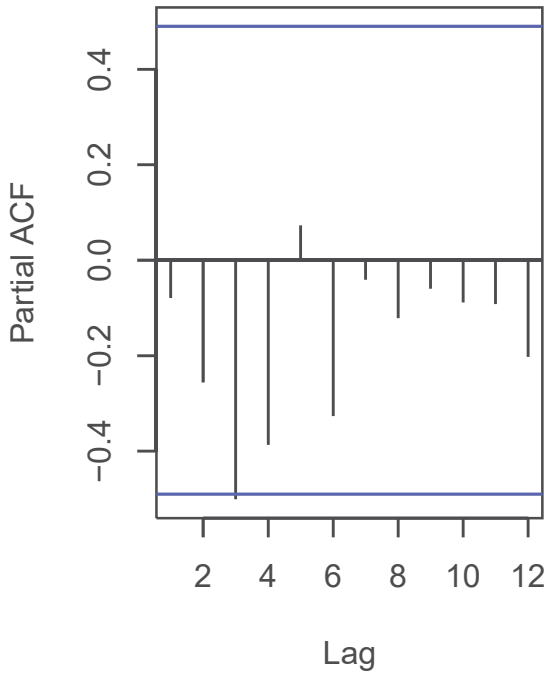


Figure 41: PACF
Residuals of ARIMA(1,2,0)

From ACF and PACF of the residual, we can conclude that the residuals may constitute a white noise series as there are no highly significant correlation in ACF.

We will explore this further using Ljung-Box Test and visualisation tool.

8.2.4 Box-Ljung Test of Residuals

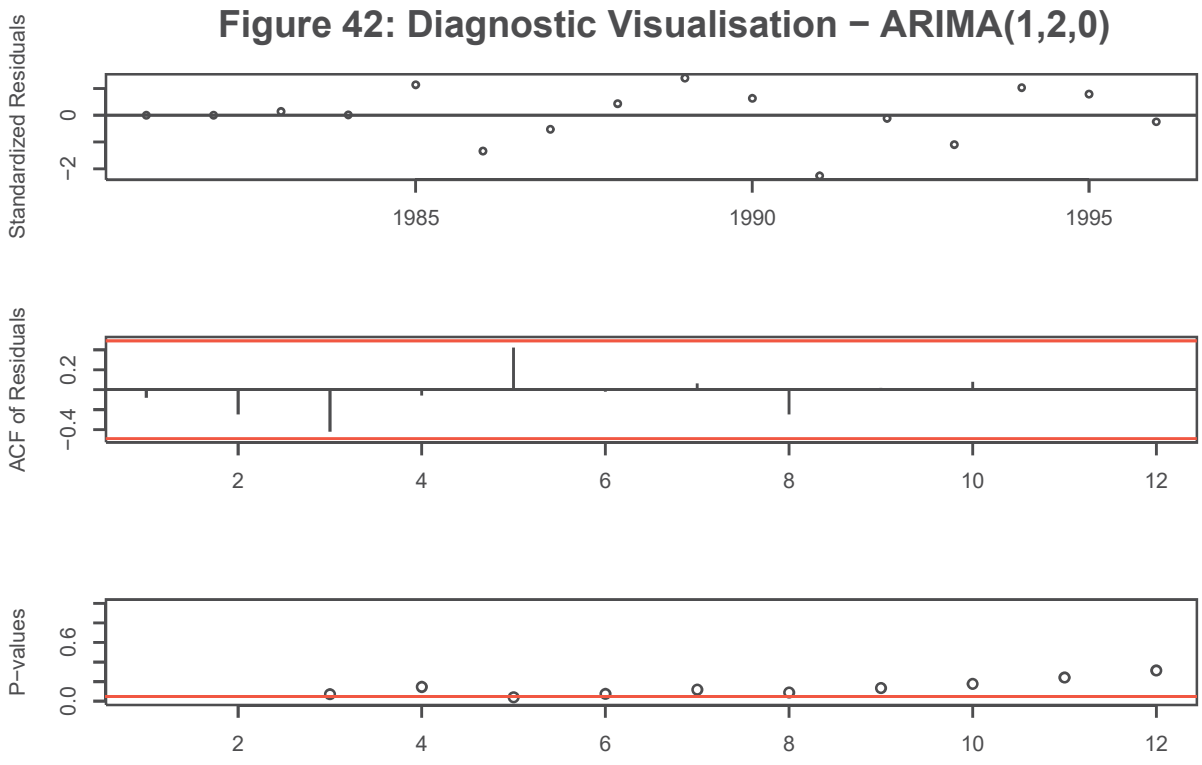
```
Box.test(residuals(model.120), lag = 12, type = "Ljung-Box", fitdf = 0)
```

```
##
##  Box-Ljung test
##
## data:  residuals(model.120)
## X-squared = 12.7, df = 12, p-value = 0.3912
```

Ljung-Box Test also supports the non-existence of correlation in residuals at 5% significance level.

8.2.5 Diagnostics Visualisation

```
tsdiag(model.120,gof=12,omit.initial=F)
title("Figure 42: Diagnostic Visualisation - ARIMA(1,2,0)")
```

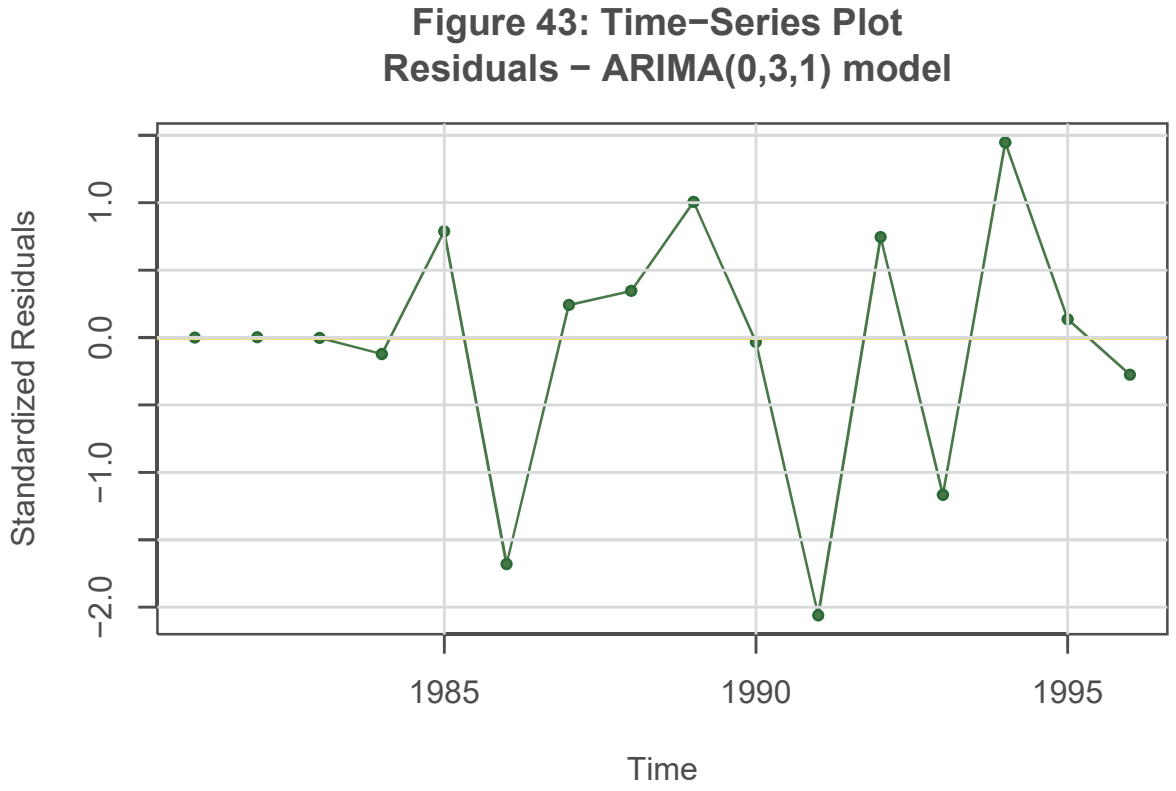


From the above diagnostic visualisation, it can be clearly seen that $p - value$ of the Ljung-Box Test statistic for some lags are *significant*. Therefore, there is existence of correlation in residuals of ARIMA(1,2,0) model.

8.3 Residual Analysis - ARIMA(0,3,1)

8.3.1 Time-Series Plot of Residuals

```
plot(rstandard(model.031),ylab ='Standardized Residuals',type='o',main="Figure 43: Time-Series Plot \n
abline(h=0, col="gold", lty=2, lw=2)
grid()
```

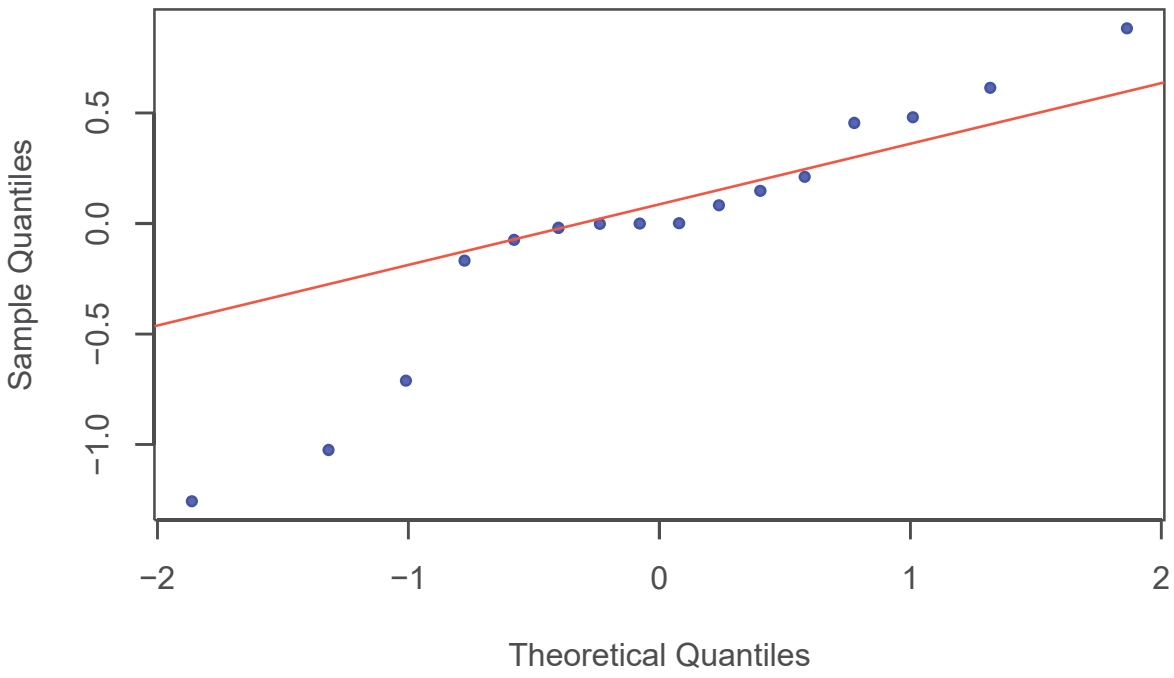


From fig. 43, we can say that there is no trend, but, there is an obvious change in variance in the residuals

8.3.2 Normality Check of Residuals

```
qqnorm(residuals(model.031), pch = 20, col="blue", main="Figure 44: Normal Q-Q Plot")
qqline(residuals(model.031), lty=2, col="red", lw=1.5)
```

Figure 44: Normal Q−Q Plot




```
shapiro.test(residuals(model.031))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model.031)
## W = 0.91807, p-value = 0.157
```

Although we can see departures from normality line in fig. 44, Shapiro-Wilk Test gives residuals to be normally distributed at 5% signigicane level.

8.3.3 ACF and PACF of Residuals

```
par(mfrow=c(1,2))
acf(residuals(model.031), main = "Figure 45: ACF \n Residuals of ARIMA(0,3,1)")
pacf(residuals(model.031), main = "Figure 46: PACF \n Residuals of ARIMA(0,3,1)")
```

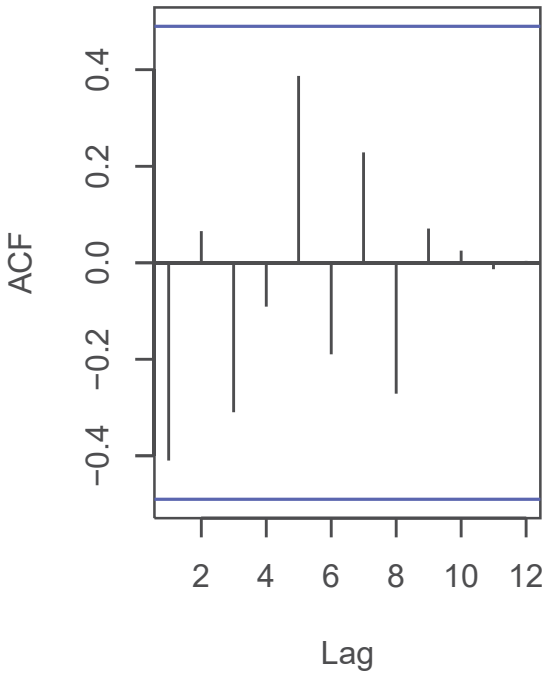


Figure 45: ACF
Residuals of ARIMA(0,3,1)

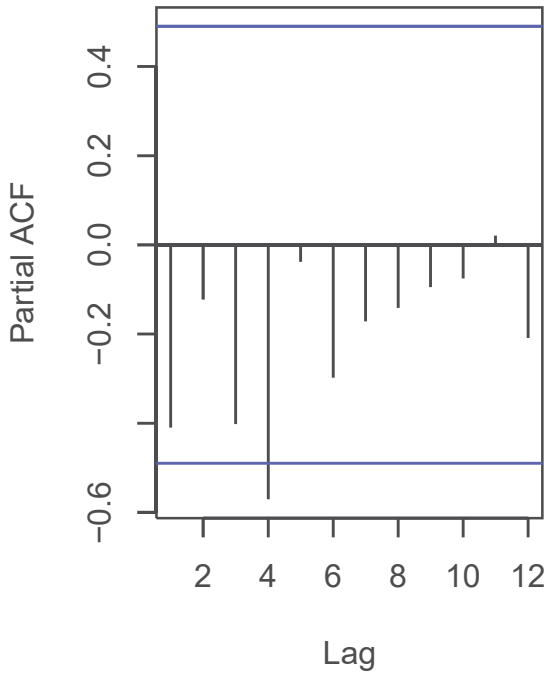


Figure 46: PACF
Residuals of ARIMA(0,3,1)

From fig. 46, we can conclude that the residuals does not constitute a white noise series as there are highly significant correlation.

8.3.4 Box-Ljung Test of Residuals

```
Box.test(residuals(model.031), lag = 12, type = "Ljung-Box", fitdf = 0)
```

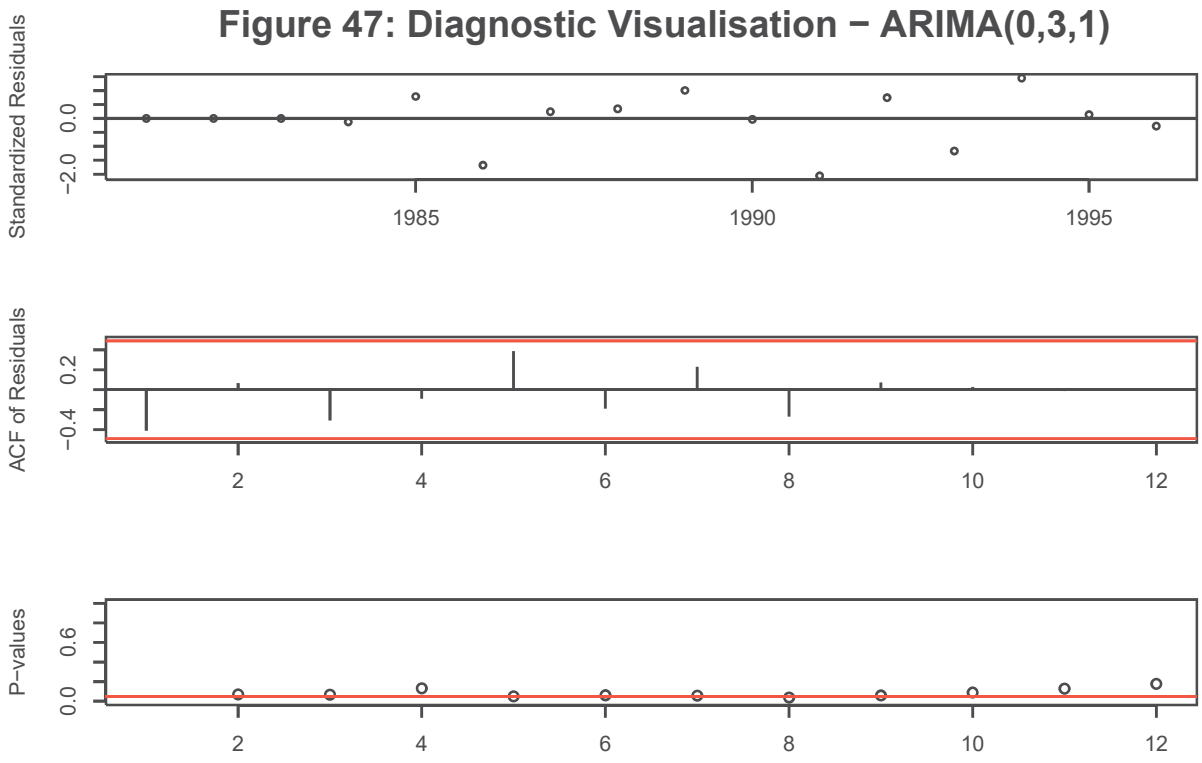
```
##
##  Box-Ljung test
##
## data:  residuals(model.031)
## X-squared = 15.118, df = 12, p-value = 0.2351
```

As seen in ACF plot, Ljung-Box Test shows the non-existence of correlation in residuals at 5% significance level.

We will explore this further using a visualisation tool.

8.3.5 Diagnostics Visualisation

```
tsdiag(model.031,gof=12,omit.initial=F)
title("Figure 47: Diagnostic Visualisation - ARIMA(0,3,1)")
```

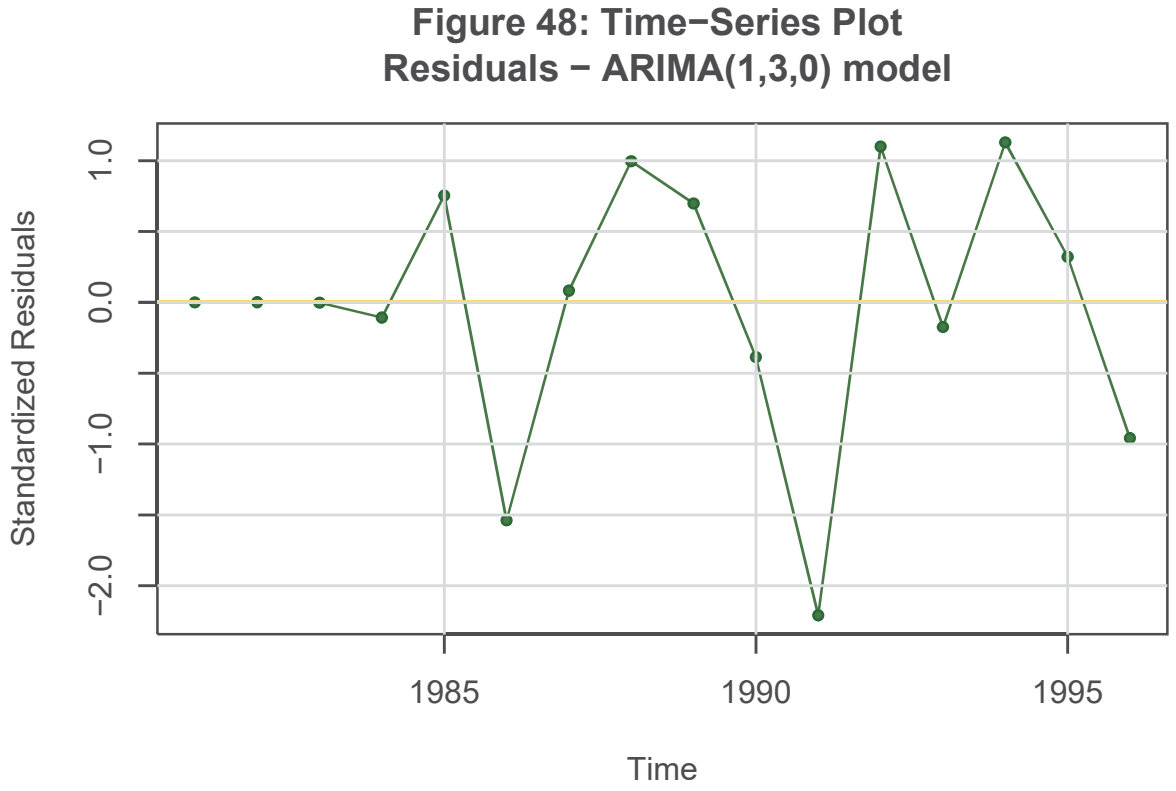


From fig. 47, it can be clearly seen that $p-value$ of the Ljung-Box Test statistic shows existence of correlation at most lags. Therefore, there is existence of correlation in residuals of ARIMA(0,3,1) model.

8.4 Residual Analysis - ARIMA(1,3,0)

8.4.1 Time-Series Plot of Residuals

```
plot(rstandard(model.130),ylab ='Standardized Residuals',type='o',main="Figure 48: Time-Series Plot \n
abline(h=0, col="gold", lty=2, lw=2)
grid()
```

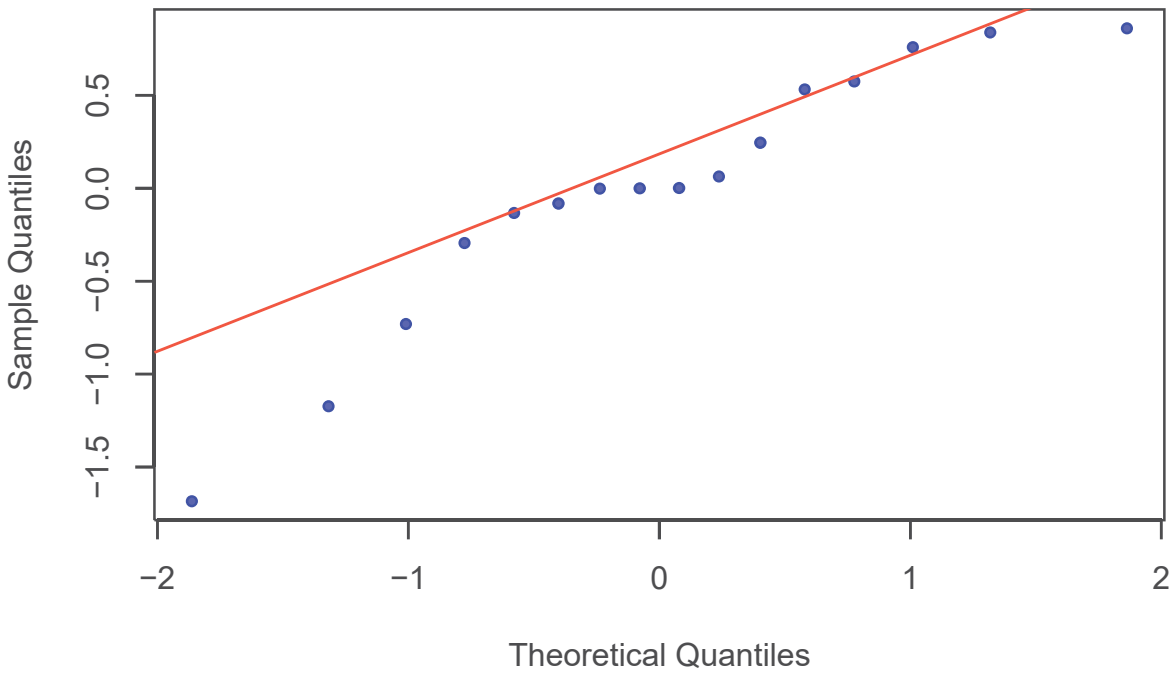


From fig. 48, we can say that there is no trend, but, there is an obvious change in variance in the residuals

8.4.2 Normality Check of Residuals

```
qqnorm(residuals(model.130), pch = 20, col="blue", main = "Figure 49: Normal Q-Q Plot")
qqline(residuals(model.130), lty=2, col="red", lw=1.5)
```

Figure 49: Normal Q−Q Plot




```
shapiro.test(residuals(model.130))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model.130)
## W = 0.91285, p-value = 0.1293
```

Although we can see departures from the normality line in Q-Q Plot, but, Shapiro Test gives the residuals to be normally distributed at 5% signigicane level.

8.4.3 ACF and PACF of Residuals

```
par(mfrow=c(1,2))
acf(residuals(model.130), main = "Figure 50: ACF \n Residuals of ARIMA(1,3,0)")
pacf(residuals(model.130), main = "Figure 51: PACF \n Residuals of ARIMA(1,3,0)")
```

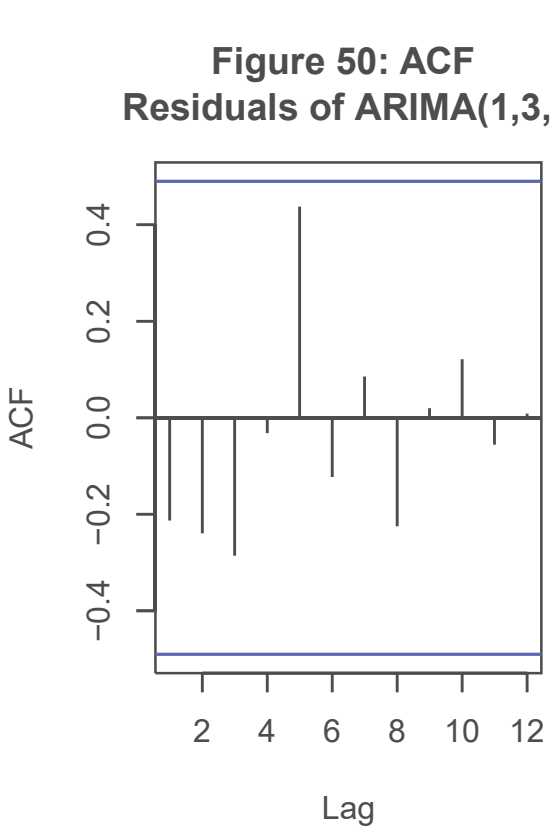


Figure 50: ACF
Residuals of ARIMA(1,3,0)

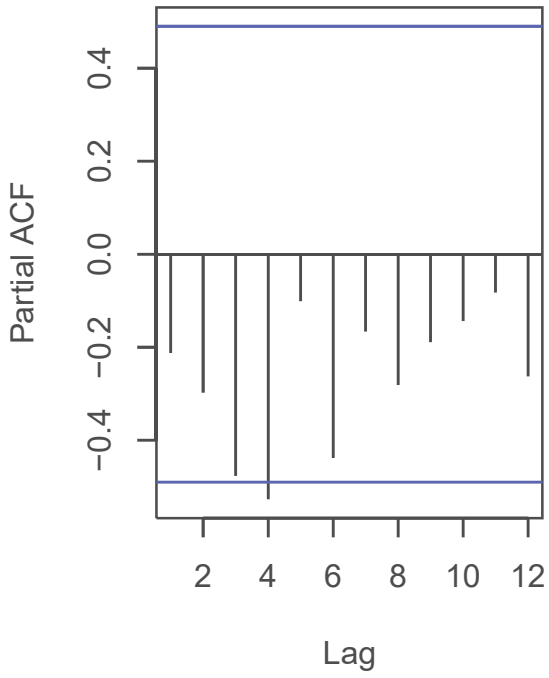


Figure 51: PACF
Residuals of ARIMA(1,3,0)

From fig. 50 & fig. 51, we can conclude that the residuals may constitute a white noise series as there are significant correlation.

We will check this hypthesis using the Ljung-Box Test.

8.4.4 Box-Ljung Test of Residuals

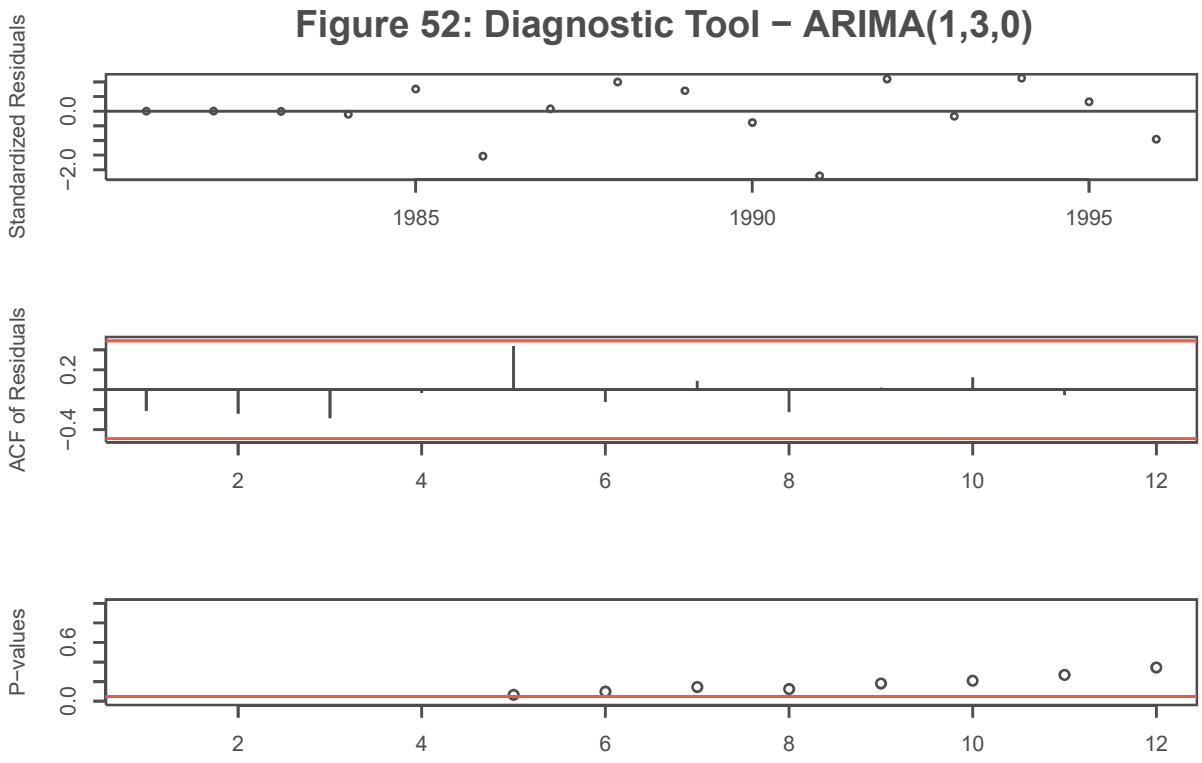
```
Box.test(residuals(model.130), lag = 12, type = "Ljung-Box", fitdf = 0)
```

```
##
##  Box-Ljung test
##
## data:  residuals(model.130)
## X-squared = 12.255, df = 12, p-value = 0.4254
```

As seen in ACF & PACF plots, Ljung-Box Test also supports the non-existence of correlation in residuals at 5% significance level.

8.4.5 Diagnostics Visualisation

```
tsdiag(model.130,gof=12,omit.initial=F)
title("Figure 52: Diagnostic Tool - ARIMA(1,3,0)")
```

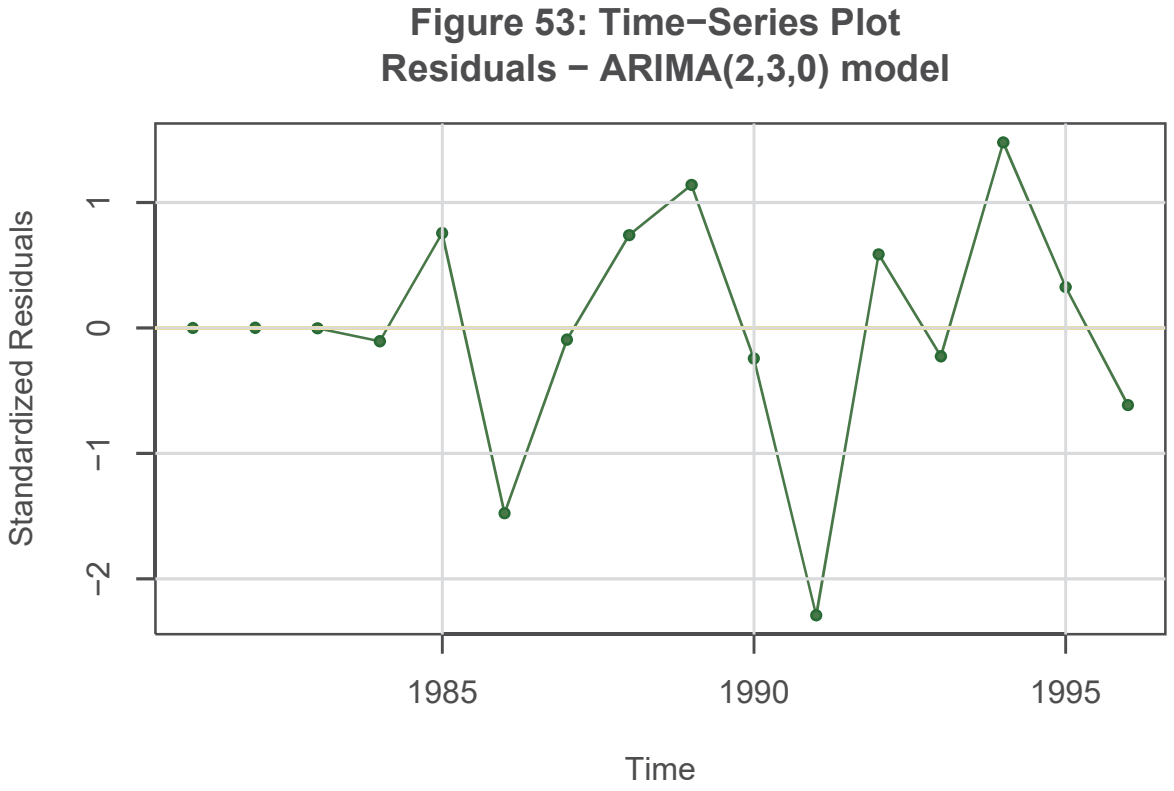


From the fig. 52, it can be clearly seen that $p - value$ of the Ljung-Box Test statistic for the whole range supports non-existence of correlation. Therefore, there is no existence of correlation in residuals of ARIMA(1,3,0) model.

8.5 Residual Analysis - ARIMA(2,3,0)

8.5.1 Time-Series Plot of Residuals

```
plot(rstandard(model.230),ylab ='Standardized Residuals',type='o',main="Figure 53: Time-Series Plot \n
abline(h=0, col="gold", lty=2, lw=2)
grid()
```

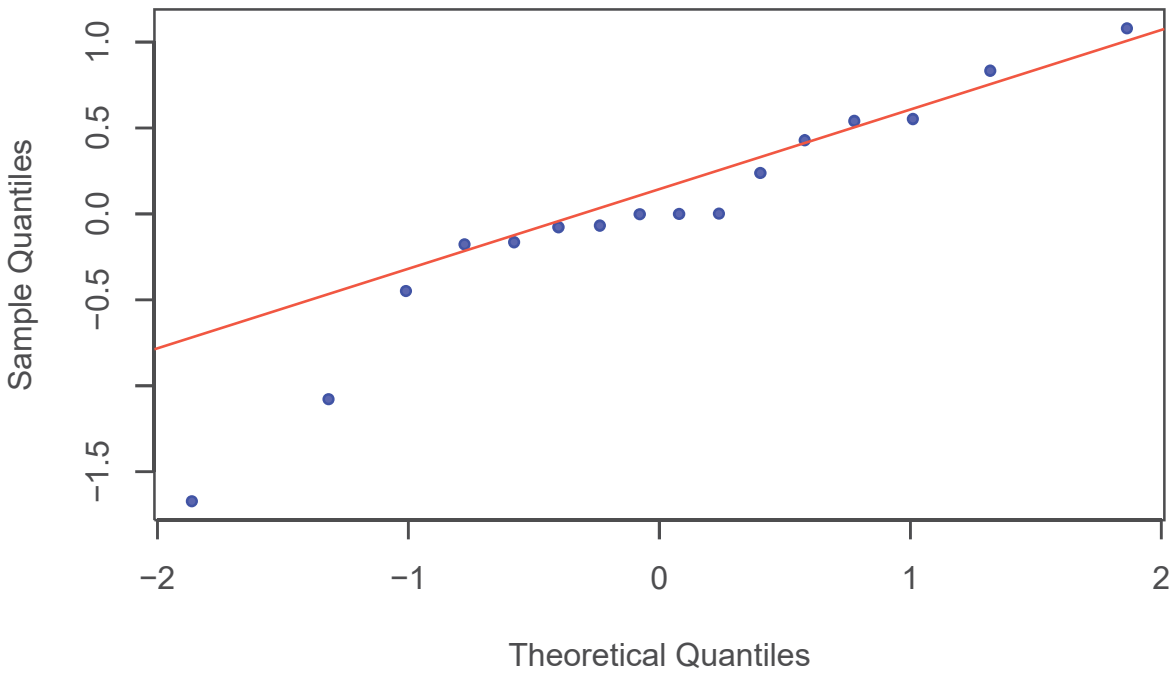


From fig. 53, we can say that there is no trend, but, there is an obvious change in variance in the residuals

8.5.2 Normality Check of Residuals

```
qqnorm(residuals(model.230), pch = 20, col="blue", main = "Figure 54: Normal Q-Q Plot")
qqline(residuals(model.230), lty=2, col="red", lw=1.5)
```

Figure 54: Normal Q−Q Plot




```
shapiro.test(residuals(model.230))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model.230)
## W = 0.92681, p-value = 0.2169
```

Although we can see departures from the normality line in fig. 54, the Shapiro-Wilk Test gives the residuals to be normally distributed at 5% signigicane level.

8.5.3 ACF and PACF of Residuals

```
par(mfrow=c(1,2))

acf(residuals(model.230), main = "Figure 55: ACF \n Residuals of ARIMA(2,3,0)")
pacf(residuals(model.230), main = "Figure 56: ACF \n Residuals of ARIMA(2,3,0)")
```

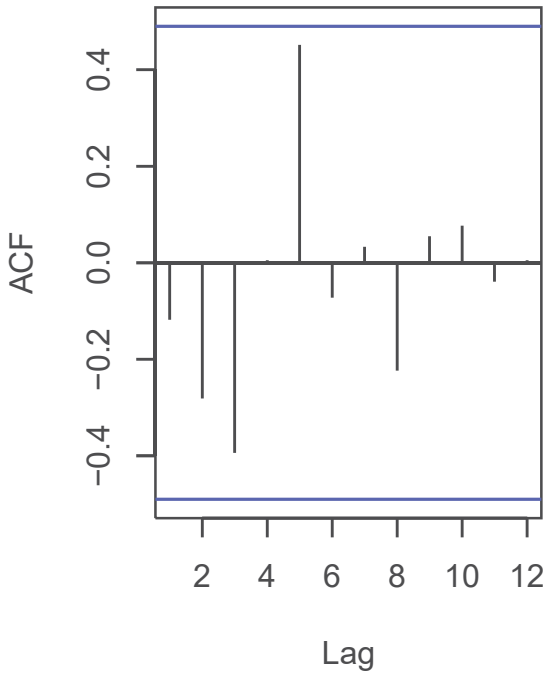


Figure 55: ACF Residuals of ARIMA(2,3,0)

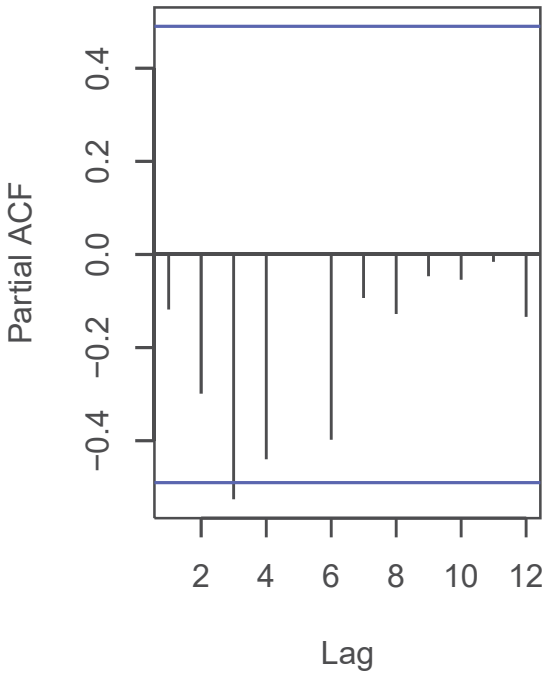


Figure 56: ACF Residuals of ARIMA(2,3,0)

From fig. 55 & fig. 56, we can conclude that the residuals may constitute a white noise series as there is a significant correlation.

8.5.4 Box-Ljung Test of Residuals

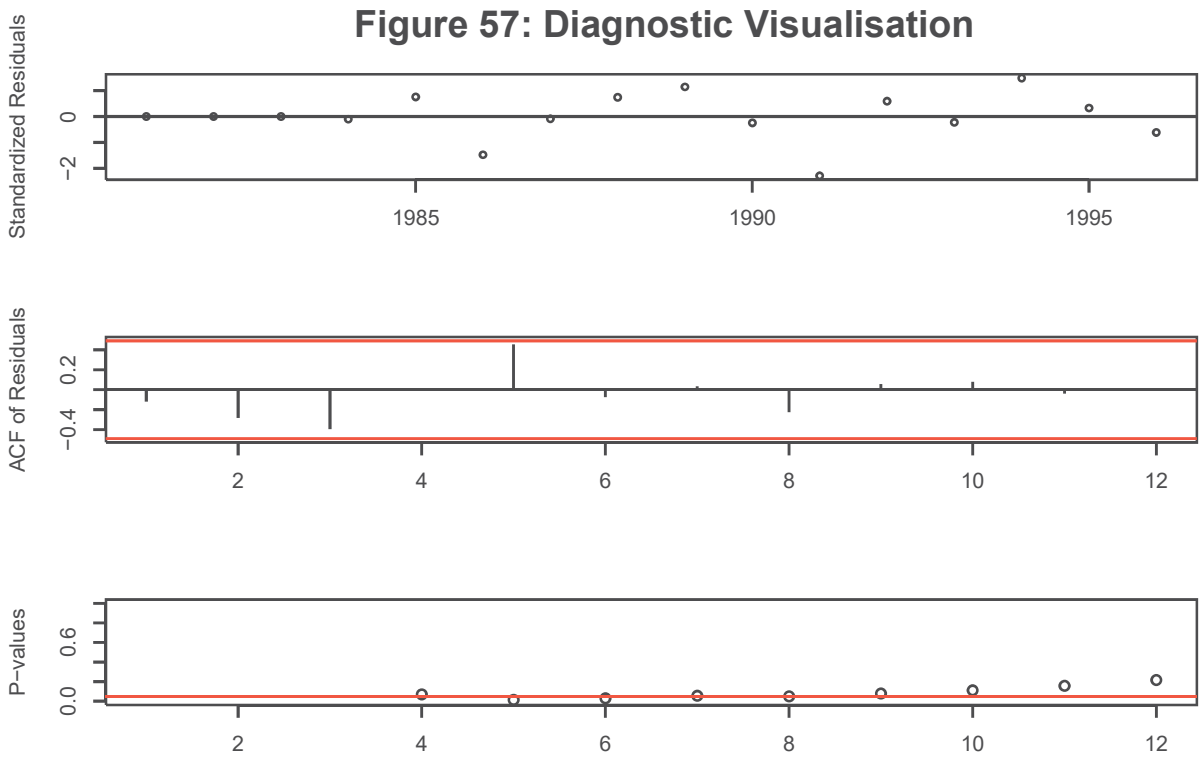
```
Box.test(residuals(model.230), lag = 12, type = "Ljung-Box", fitdf = 0)
```

```
##
##  Box-Ljung test
##
## data:  residuals(model.230)
## X-squared = 13.13, df = 12, p-value = 0.3597
```

As seen in ACF & PACF plots, Ljung-Box Test also supports the non-existence of correlation in residuals at 5% significance level.

8.5.5 Diagnostics Visualisation

```
tsdiag(model.230,gof=12,omit.initial=F)
title("Figure 57: Diagnostic Visualisation")
```



From the fig. 57, it can be clearly seen that $p - value$ of the Ljung-Box Test statistic at most lags support existence of correlation. Therefore, there is existence of correlation in residuals of ARIMA(2,3,0) model.

8.6 Residual Analysis - Conclusion

As seen from various diagnostic checks, following candidate models will be taken in `Model Selection`: + ARIMA(0,2,1) + ARIMA(1,3,0)

9 MODEL SELECTION

We will use AIC and BIC values for model selection.

9.1 AIC

```
AIC(model.021, model.130)
```

```
## Warning in AIC.default(model.021, model.130): models are not all fitted to
## the same number of observations
```

```
##           df      AIC
## model.021  2 22.36645
## model.130  2 34.34278
```

9.2 BIC

```
BIC(model.021, model.130)
```

```
## Warning in BIC.default(model.021, model.130): models are not all fitted to
## the same number of observations
```

```
##           df      BIC
## model.021  2 23.64456
## model.130  2 35.47268
```

According to AIC and BIC, the *best model* is *ARIMA(0,2,1)*.

10 FORECASTING - ARIMA(0,2,1)

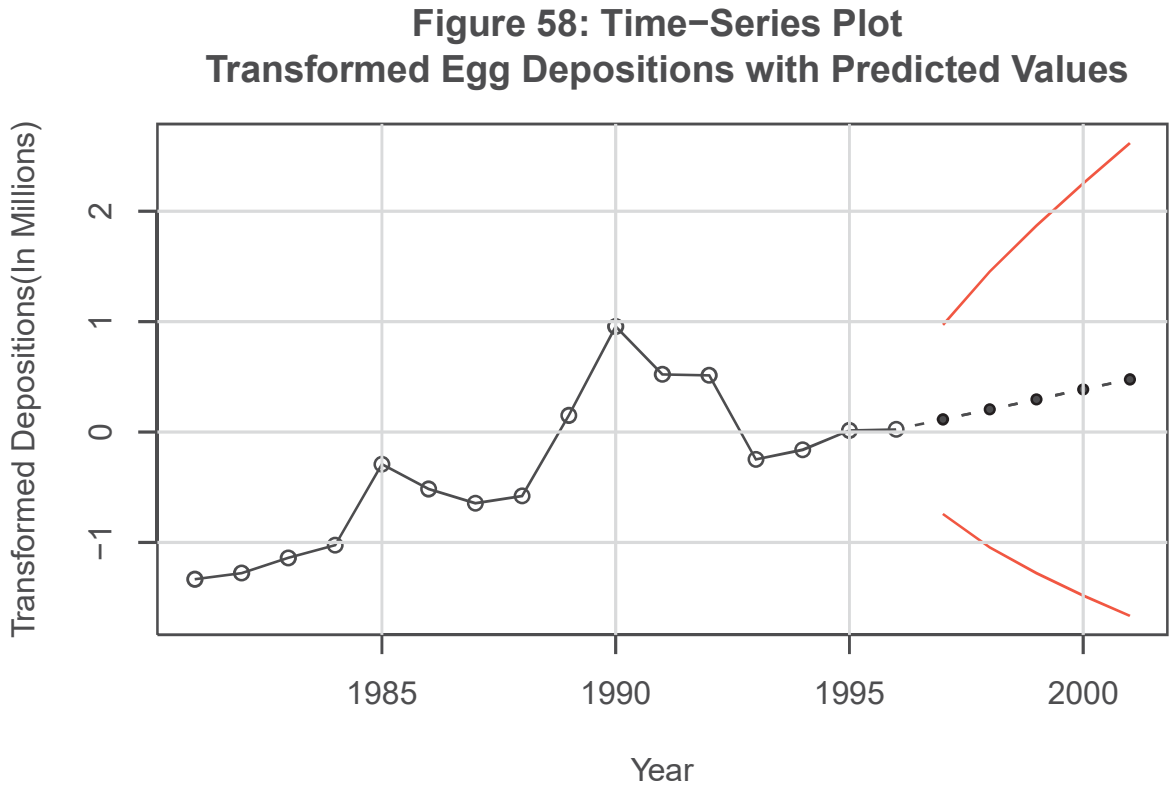
10.1 Prediction

```
eggs.predict.transform = predict(model.021, n.ahead = 5, newxreg = NULL, se.fit = TRUE)
eggs.predict.transform
```

```
## $pred
## Time Series:
## Start = 1997
## End = 2001
## Frequency = 1
## [1] 0.1145020 0.2050017 0.2955014 0.3860010 0.4765007
##
## $se
## Time Series:
## Start = 1997
## End = 2001
## Frequency = 1
## [1] 0.4371667 0.6372745 0.8031266 0.9527828 1.0929168
```

10.2 Plot of Transformed Series

```
bestfitmodel = arima(BC.eggs, order=c(0,2,1), xreg=data.frame(constant=seq(eggs.ts))) # Create matrix o
n=length(eggs.ts)
n.ahead=5 #Forecast 5 years ahead
newxreg=data.frame(constant=(n+1):(n+n.ahead))
dataTransform =
plot.Arima(bestfitmodel, n.ahead=n.ahead, newxreg=newxreg,ylab='Transformed Depositions(In Millions)',x
grid()
```

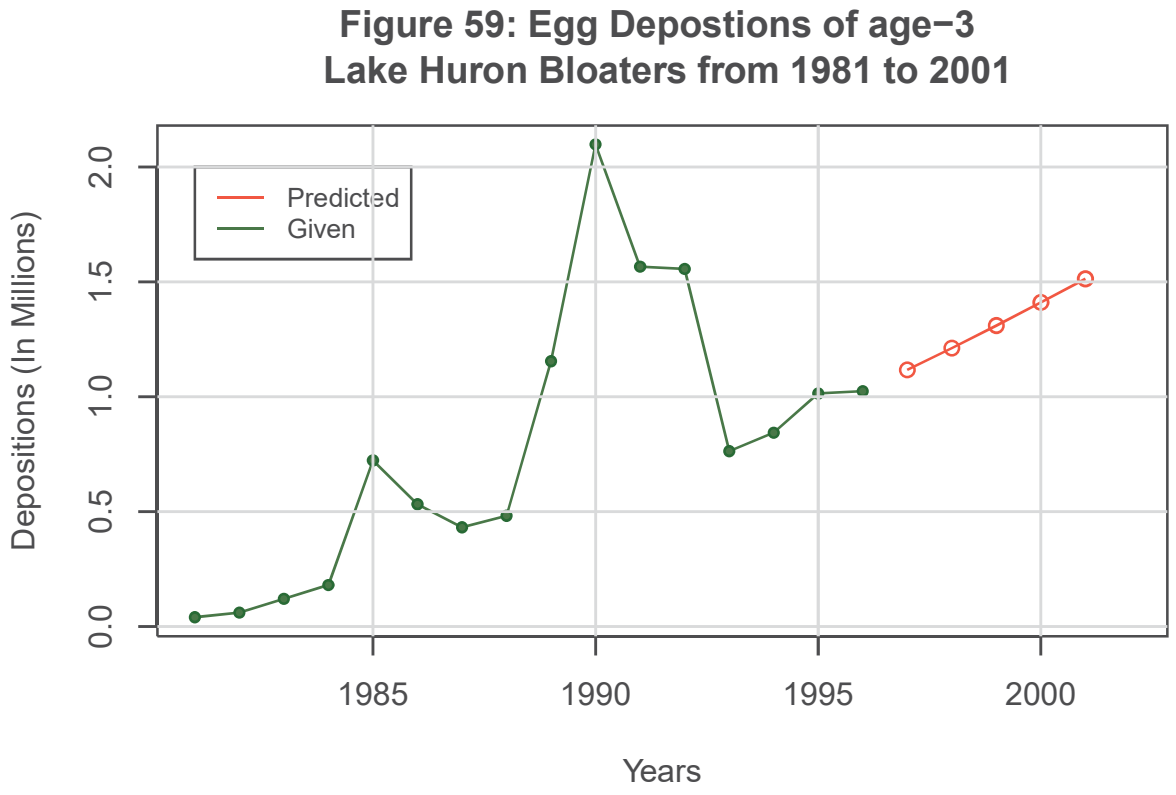


10.3 Plot of Time-Series with Predications

```
eggs.predict = data.frame(eggs.predict.transform)
eggs.predict[] <- lapply(eggs.predict[], function(x) (0.66*(x)+1)^(1/0.66))

eggs.predict.ts = ts(eggs.predict$pred, start=c(1997,1))

plot((eggs.ts), type = "o", ylab = 'Depositions (In Millions)', xlab = 'Years',
     main = 'Figure 59: Egg Depostions of age-3 \n Lake Huron Bloaters from 1981 to 2001',
     col = 'darkgreen', pch = 20, lwd=2, xlim=c(1981,2002))
lines(eggs.predict.ts, col = "red", type="o", lty=2)
legend(1981, 2, legend=c("Predicted", "Given"),
       col=c("red", "darkgreen"), lty=2:1, cex=0.8)
grid()
```



11 SUMMARY

In the given report, we observed that when the sample size is small, power of Shapiro-Wilk Test reduces. Therefore, visualisataions were relied on to check normality distribution of the series.

To model the egg depositions series, trend models were rejected based on diagnostic checking and its failure to capture auto-correlation in the series.

Various ARIMA models are taken to find the possible candidate models. We used order of differencing as 2 & 3, as beyond 3, correlations were introduced as per ACF & PACF. This gave us the signs of over-differencing. To accomodate mild under-differencing or over-differencing, candidate models from both orders were taken.

ARIMA(0,2,1) was found to be the best fit model. Also, as the sample size is very small, predicting depositions for next 5 years becomes unreliable with high Standard Error. The visualisation of series with predicted values in given in `Section 10.3`.

12 REFRENCE

Oztuna D, Elhan AH, Tuccar E. Investigation of four different normality tests in terms of type 1 error rate and power under different distributions. Turkish Journal of Medical Sciences. 2006;36(3):171-6.

Robert Nau - Duke University