Egg depositions forecasting

Code

Time Series Analysis

MUJEER M.

Introduction

Our eggs dataset from 1981 to 1996 shows the number of Coregonus hoyi of age-3 egg depositions made through the years (in millions). We will analyse this dataset to capture this trend and predict the outcome for the next five years.

Methodology

- To perform this analysis we have analysed the time series plot, ACF plot, PACF plot.
- Applied transformation and done differencing to the data due to non-stationarity.
- Parameter estimation after selecting the possible models from ACF, PACF, EACF, BIC.
- Filtering the models further based on AIC and BIC score.
- Diagnostic checks on 2 models selected ARIMA(0,4,2) and ARIMA(1,4,1)
- Forecasting on ARIMA model(1,4,1) for the next five years

Setup

Hide

```
library(TSA)
library(dplyr)
library(readr)
library(analytics)
library(ggplot2)
library(tidyr)
library(ggfortify)
library(tidyverse)
library(broom)
library(magrittr)
library(fUnitRoots)
library(lmtest)
library(FitAR)
library(tseries)
library(forecast)
```

Importing and converting dataset to “ts” format

Hide

```
eggs <- read.csv("C:/Users/Mohammed/Documents/eggs.csv", header=TRUE)
view(eggs)
class(eggs)
```

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

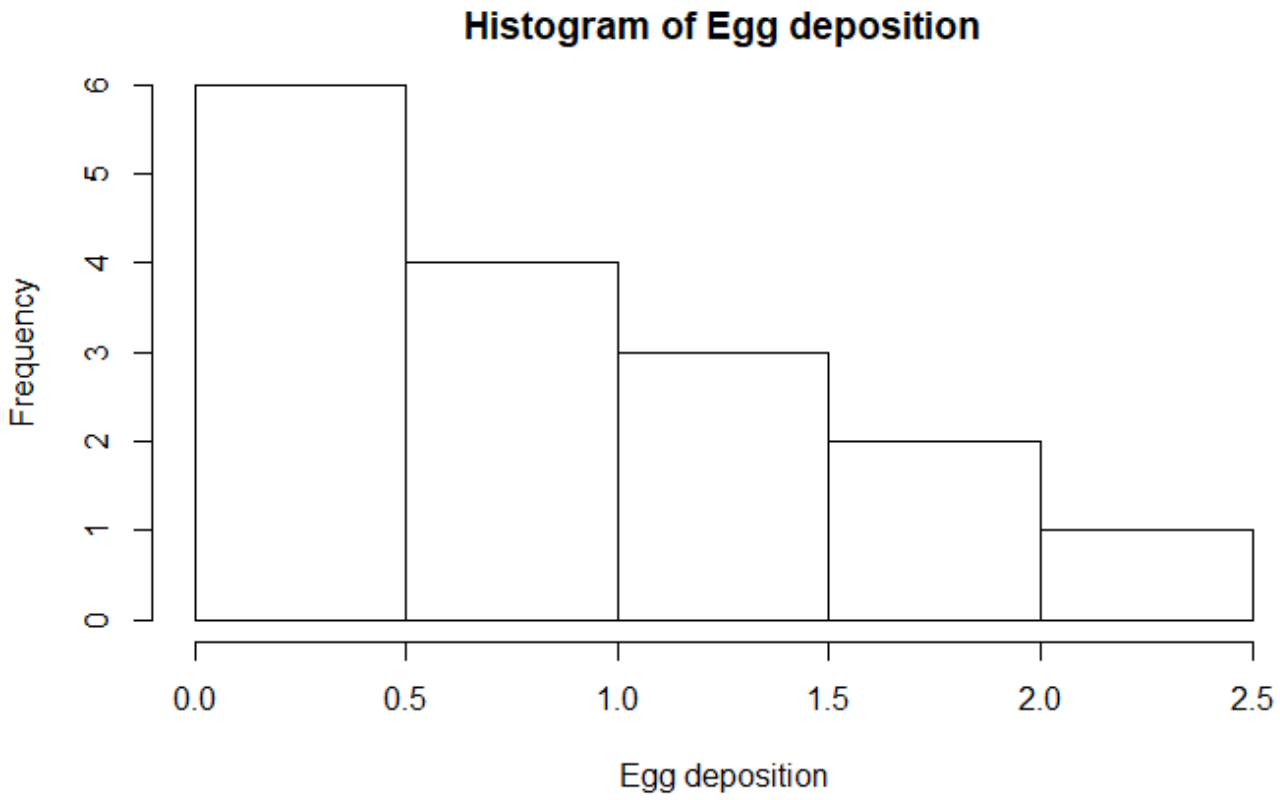
Hide

```
eggs_ts <- ts(eggs$eggs, start = 1981, end = 1996)
class(eggs_ts)
```

```
[1] "ts"
```

Hide

```
hist(eggs$eggs,xlab = "Egg deposition",main = "Histogram of Egg deposition")
```

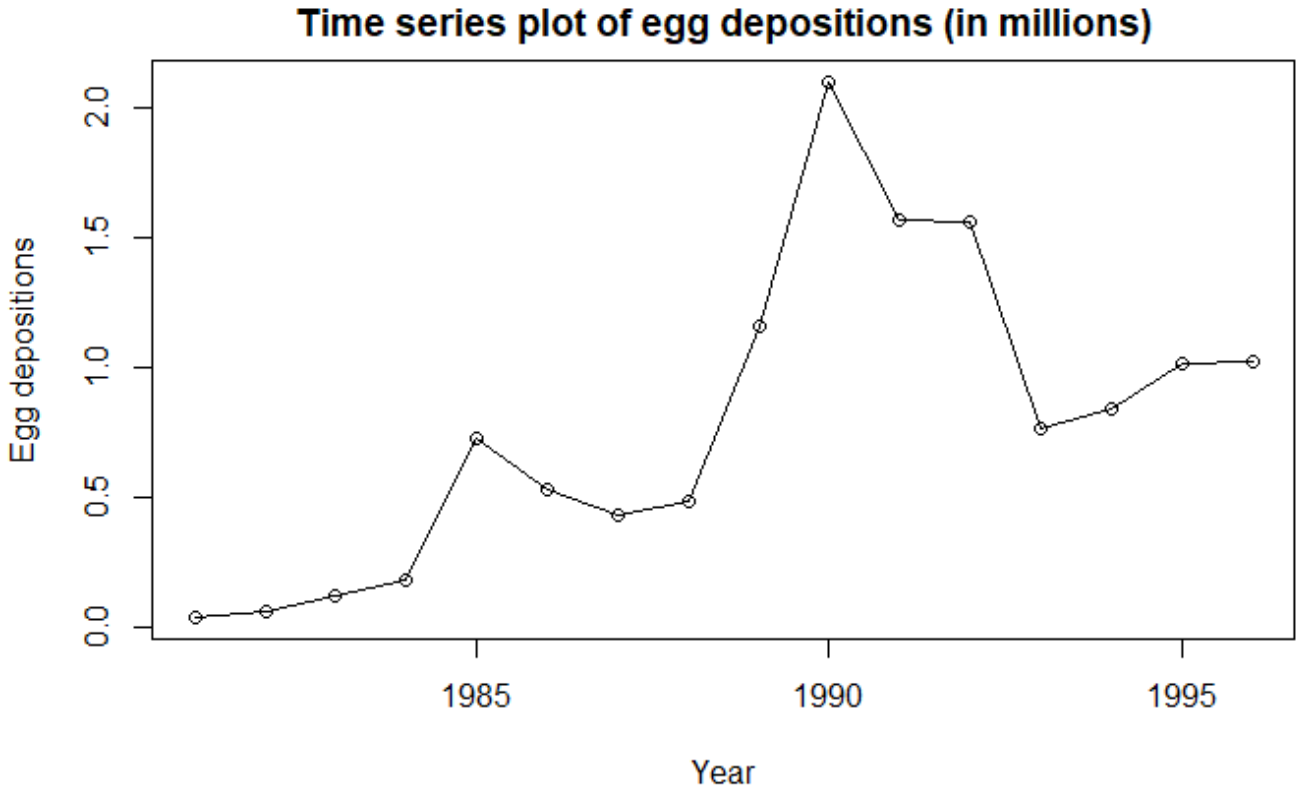


- The distribution is right skewed as egg deposition cannot be negative

Time series plot

Hide

```
plot(eggs_ts,ylab='Egg depositions',xlab='Year',type='o', main = "Time series plot of egg dep
ositions (in millions)")
```

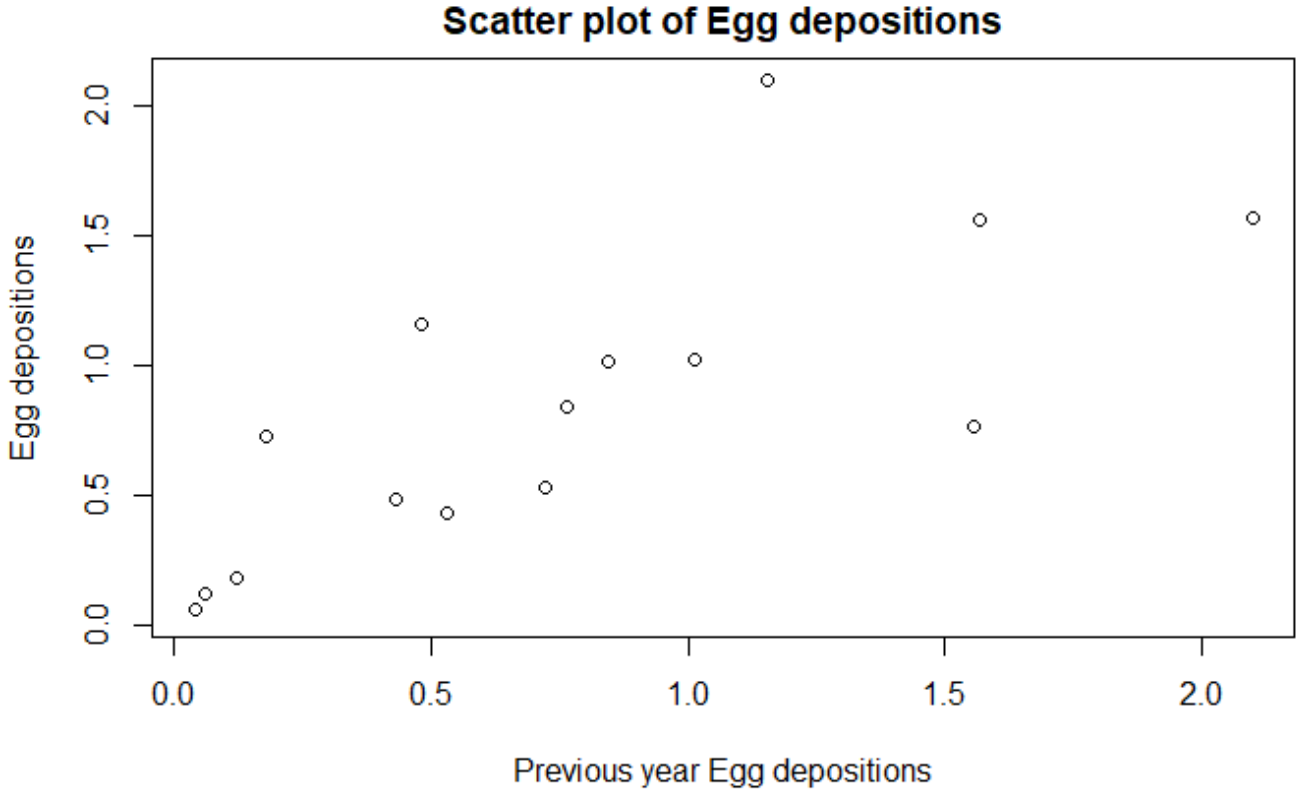


- Trend: there is an increasing change in the mean level
- Changing variance: There is not much of larger and smaller variation noted from the plot
- Seasonality: There are no obvious repeating patterns
- Autocorrelation structure:There are consecutive data points and some fluctuations, the behaviour of the series seems like autoregressive.
- Intervention:there is no change-point event or a sudden rise or drop in the data-points

Scatter plot of egg depositions

Hide

```
plot(y=eggs_ts,x=zlag(eggs_ts),ylab='Egg depositions', xlab='Previous year Egg depositions',m
ain = "Scatter plot of Egg depositions")
```

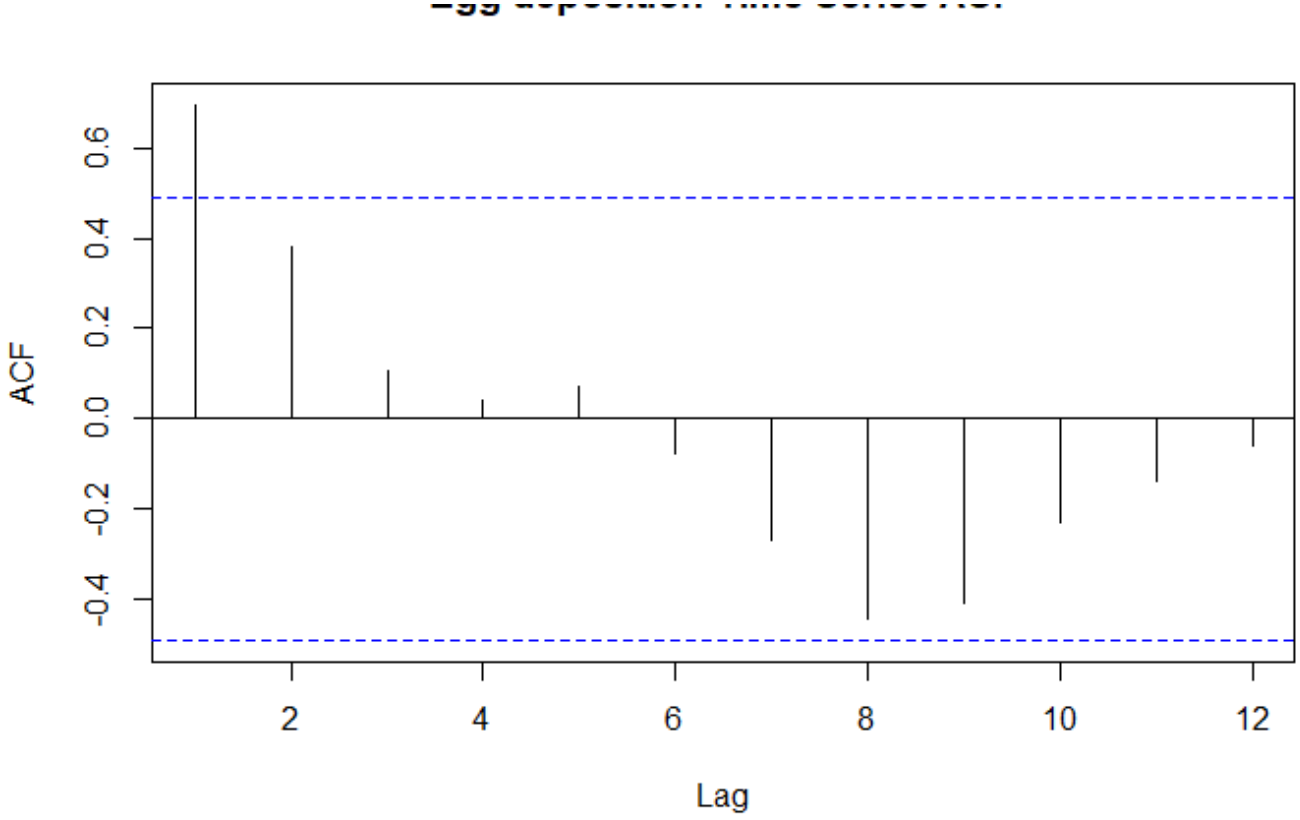


- In the above scatter plot we cannot see a trend. There is no autocorrelation in its first lag of the series

ACF plot

Hide

```
eggs_acf <- acf(eggs_ts,plot = FALSE)
plot(eggs_acf, main = "Egg deposition Time Series ACF")
```

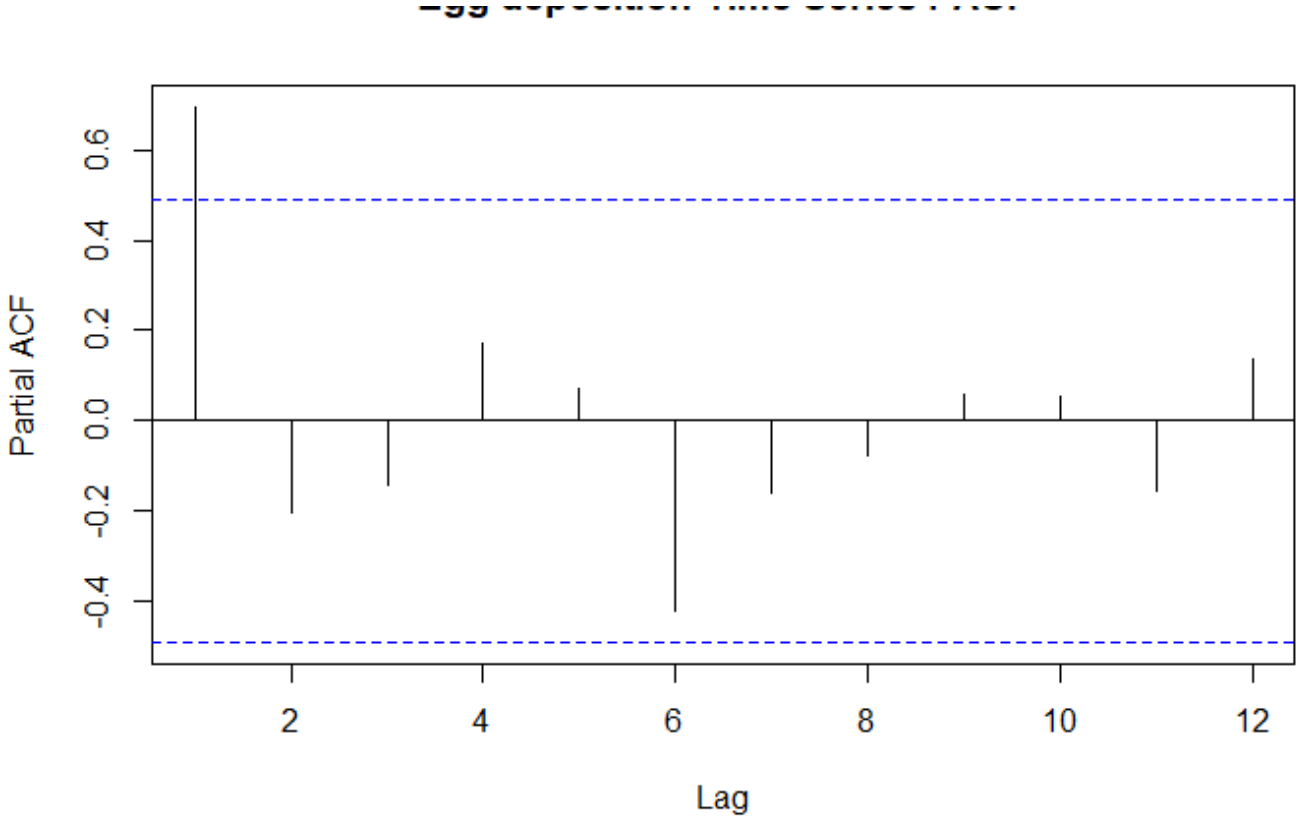


- One significant lag starting high and slowly decaying pattern, there is just one significant lag this could be because the series is small. This series is non-stationary.

PACF plot

Hide

```
eggs_pacf <- pacf(eggs_ts,plot = FALSE)
plot(eggs_pacf, main = "Egg deposition Time Series PACF")
```



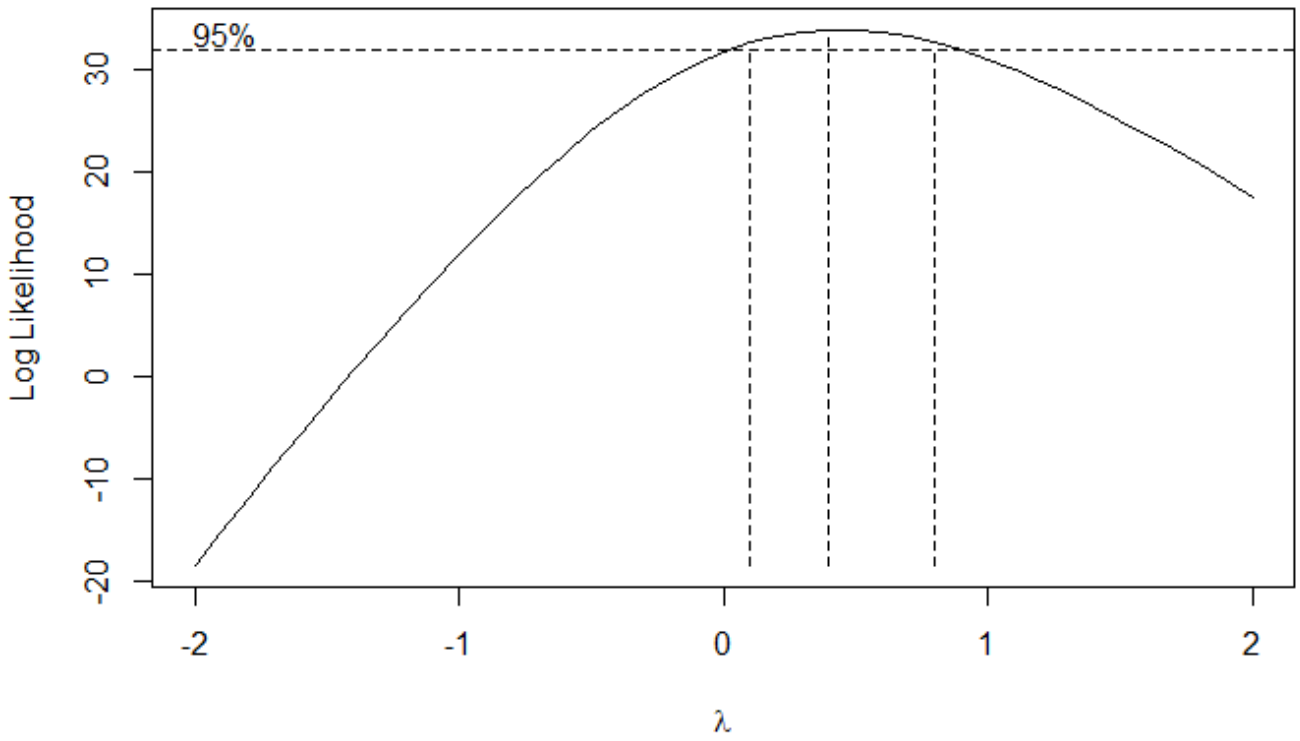
- Slowly decayin pattern in ACF and very high first correlation in PACF implies the existence of trend and nonstationarity. There is no trend or any pattern observed in the PACF plot. Hence, i am assuming non-stationarity.

Applying transformation

Hide

```
eggs_ts.transform = BoxCox.ar(eggs_ts, method = "yule-walker")
```

```
possible convergence problem: optim gave code = 1possible convergence problem: optim gave cod
e = 1
```



Hide

```
eggs_ts.transform$ci
```

```
[1] 0.1 0.8
```

Hide

```
lambda = 0.45
BC.eggs_ts = (eggs_ts^lambda-1)/lambda
```

- Applied box-cox transformation and got the lambda value from confidence interval (0.1 0.8) as 0.45

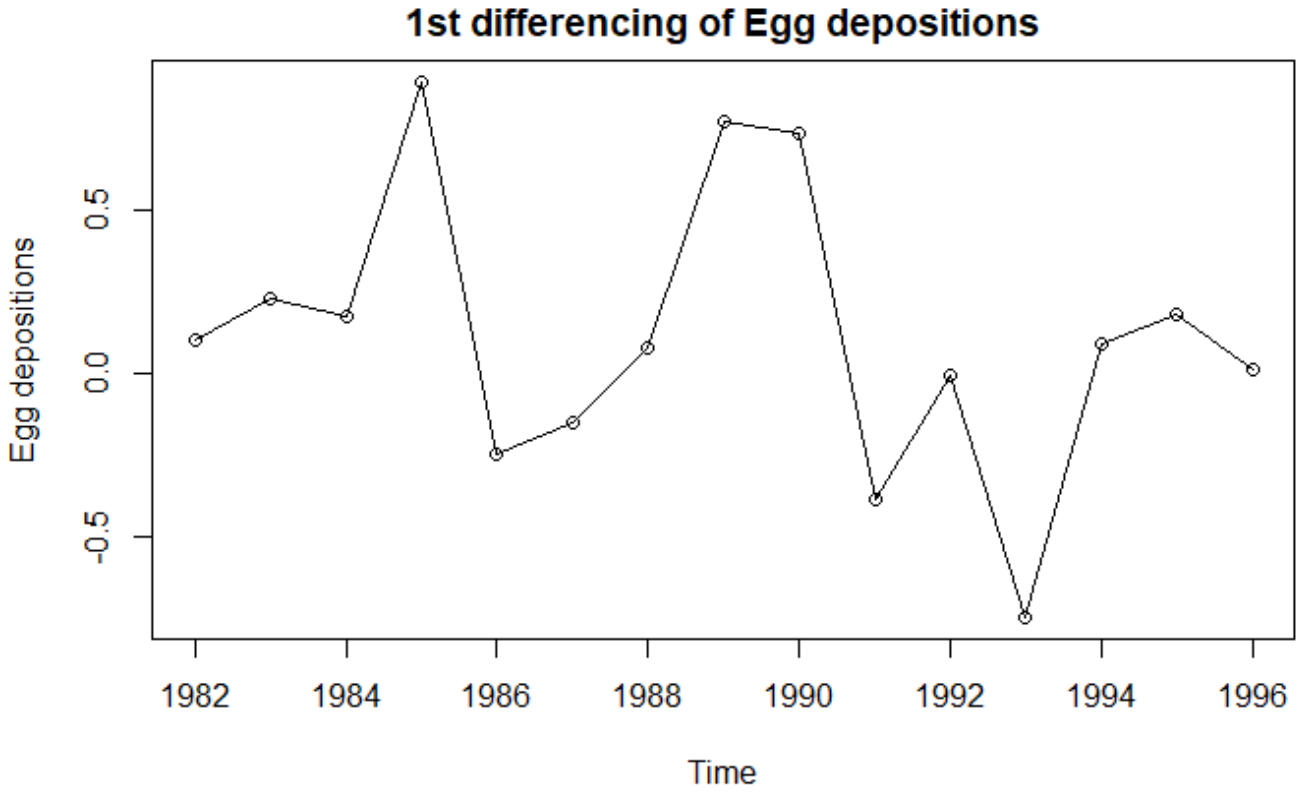
Differencing

- We are proceeding with differencing as the series is non-stationary.

1st differencing

Hide

```
diff.BC.eggs_ts = diff(BC.eggs_ts,differences=1)
plot(diff.BC.eggs_ts,type='o',ylab='Egg depositions', main='1st differencing of Egg depositio
ns')
```



ADF test on 1st differencing

Hide

```
order  = ar(diff(diff.BC.eggs_ts))$order
adfTest(diff.BC.eggs_ts, lags = order,  title = NULL,description = NULL)
```

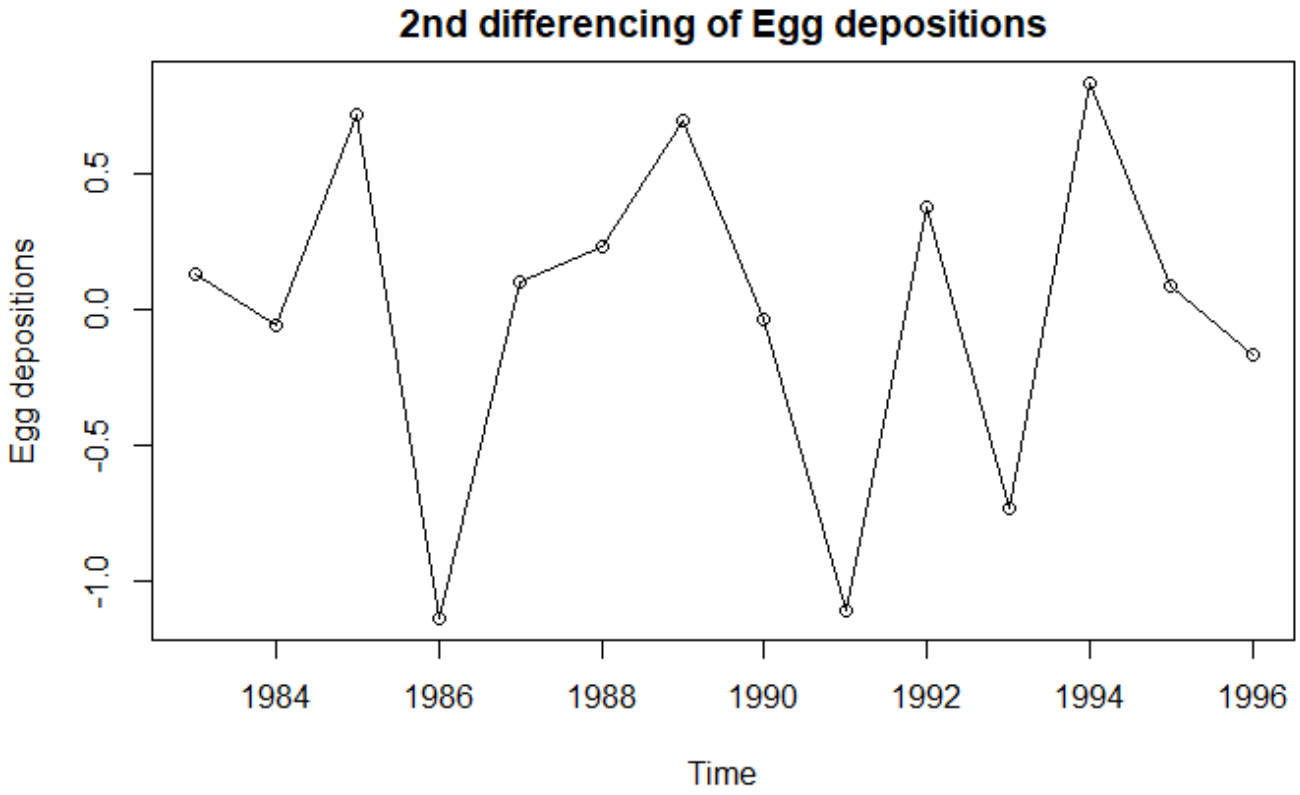
```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 4
  STATISTIC:
    Dickey-Fuller: -0.8222
  P VALUE:
    0.3469

Description:
 Mon May 13 00:46:42 2019 by user: Mohammed
```

- Since the p-value is 0.3469, there is 34.69% that my process is non stationary“. Therefore, we cannot reject H0 at, 1% and therefore we conclude that the process is non stationary.

2nd differencing

Hide

```
diff2.BC.eggs_ts = diff(BC.eggs_ts,differences=2)
plot(diff2.BC.eggs_ts,type='o',ylab='Egg depositions',main='2nd differencing of Egg depositio
ns')
```



ADF test on 2nd differencing

Hide

```
order  = ar(diff(diff2.BC.eggs_ts))$order
adfTest(diff2.BC.eggs_ts, lags = order,  title = NULL,description = NULL)
```

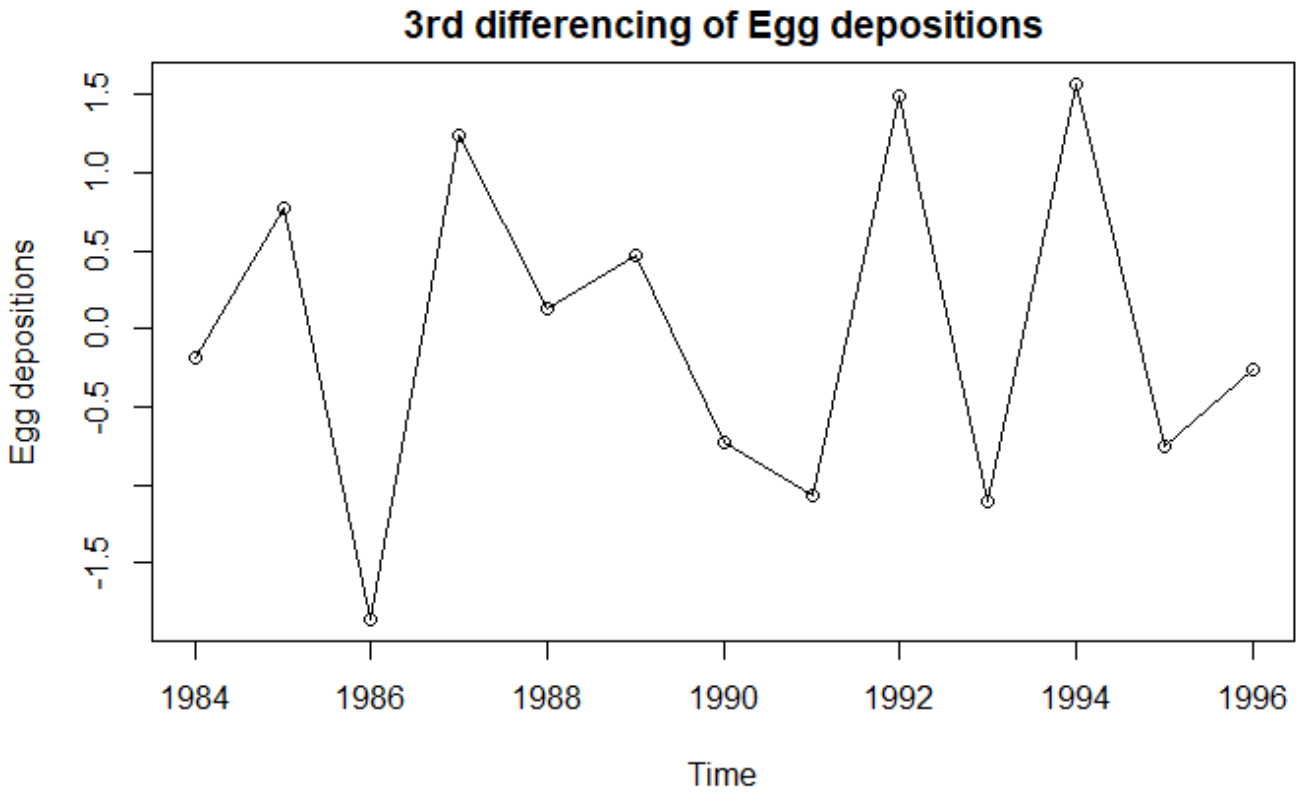
```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 4
  STATISTIC:
    Dickey-Fuller: -1.5692
  P VALUE:
    0.1098

Description:
 Mon May 13 00:46:43 2019 by user: Mohammed
```

- The p-value 0.1098 is greater than the chosen alpha level 1%. Hence we can say that the series is still non-stationary.

3rd differencing

Hide

```
diff3.BC.eggs_ts = diff(BC.eggs_ts,differences=3)
plot(diff3.BC.eggs_ts,type='o',ylab='Egg depositions',main='3rd differencing of Egg depositio
ns')
```



- The series still has seasonality and trend, we will check this with ADF test if series is stationary.

ADF test on 3rd differencing

Hide

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

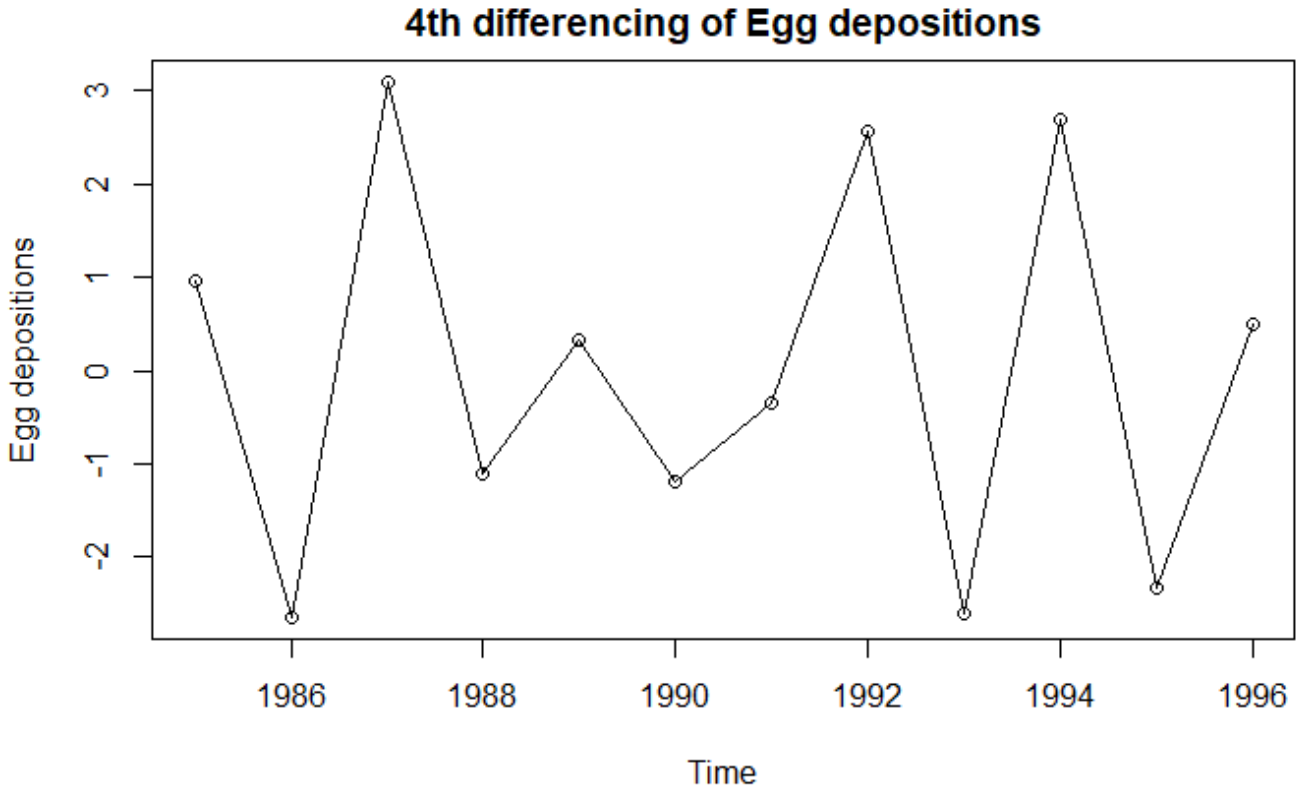
```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 4
  STATISTIC:
    Dickey-Fuller: -1.3368
  P VALUE:
    0.1836

Description:
 Mon May 13 00:46:43 2019 by user: Mohammed
```

- The p-value 0.1836 is greater than the chosen alpha level 1%. Hence we can say that the series is still non-stationary.

4th differencing

Hide

```
diff4.BC.eggs_ts = diff(BC.eggs_ts,differences=4)
plot(diff4.BC.eggs_ts,type='o',ylab='Egg depositions',main='4th differencing of Egg depositio
ns')
```



- There seems to be no trend or seasonality existing, we can further verify this by ADF test for stationarity.

ADF test on 4th differencing

Hide

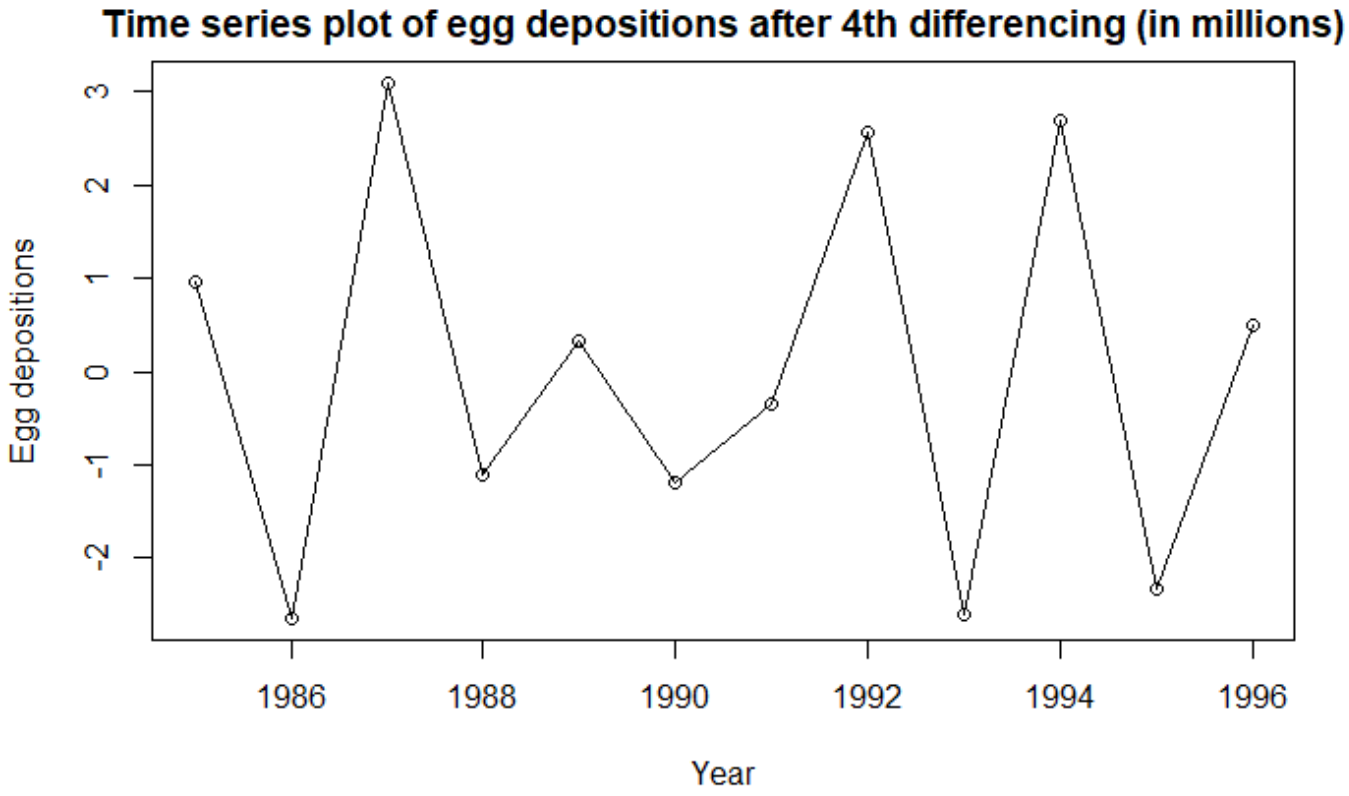
```
order  = ar(diff(diff4.BC.eggs_ts))$order
adfTest(diff4.BC.eggs_ts, lags = order,  title = NULL,description = NULL)
```

```

Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 2
  STATISTIC:
    Dickey-Fuller: -2.3228
  P VALUE:
    0.02265

Description:
 Mon May 13 00:46:43 2019 by user: Mohammed
```

- We are rejecting the null hypothesis since our p-value is less than 0.1. We can conclude that the series is stationary after 4th differencing.

Hide

```
plot(diff4.BC.eggs_ts,ylab='Egg depositions',xlab='Year',type='o', main = "Time series plot o
f egg depositions after 4th differencing (in millions)")
```

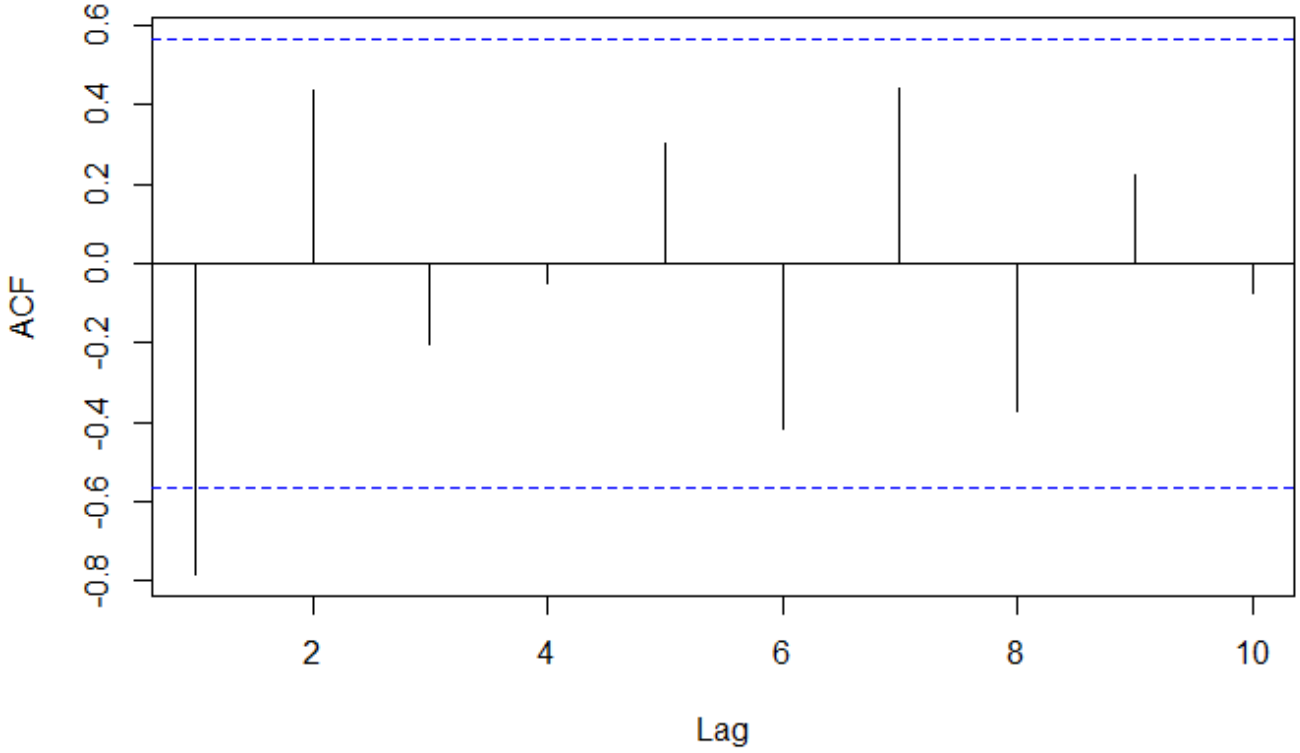


- Trend: Trend in this series is not so obvious.
- Changing variance:There could be variation between the year 1988 and 1990 against larger variations through the series.
- Seasonality:There are no obvious repeating patterns
- Autocorrelation structure:There are very few consecutive data points and more fluctuations in the series hence it could be moving average behaviour.
- Intervention:there is no change-point event or a sudden rise or drop in the data-points

ACF plot of 4th differencing

Hide

```
eggs4thdiff_acf <- acf(diff4.BC.eggs_ts,plot = FALSE)
plot(eggs4thdiff_acf, main = "4th differencing of Egg deposition Time Series ACF plot")
```

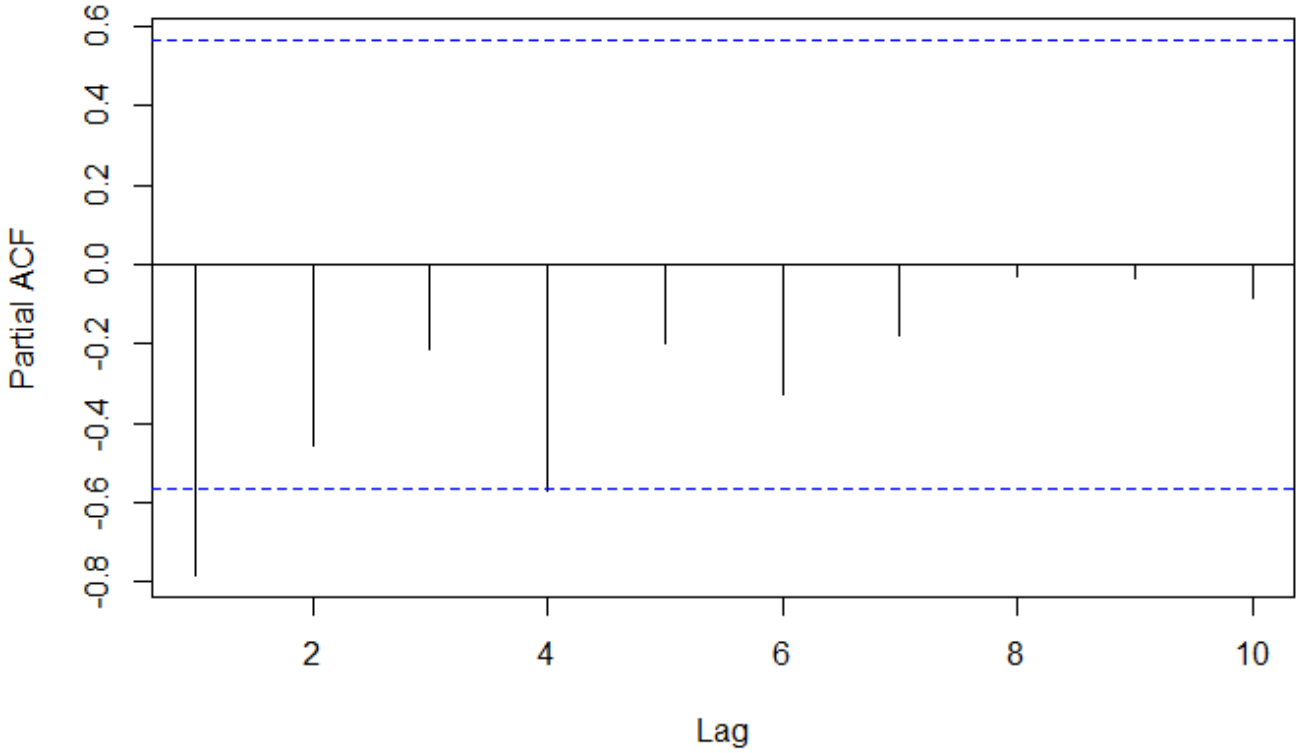


- We can observe some trend and seasonality from the plot and there is one significant lag in ACF.

PACF plot of 4th differencing

Hide

```
eggs4thdiff_pacf <- pacf(diff4.BC.eggs_ts,plot = FALSE)
plot(eggs4thdiff_pacf, main = "4th differencing of Egg deposition Time Series PACF plot")
```



- There is one significant lag and kind of a decreasing trend observed in PACF. We can consider the model {ARIMA(1,4,1)} from ACF and PACF plot.

EACF plot of 4th differencing

Hide

```
eacf(diff4.BC.eggs_ts,ar.max = 2, ma.max = 2)
```

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

- I have considered ar.max and ma.max as “2” because projection matrix results are shown invalid if the values are increased.
- Candidate models from EACF plot are {ARIMA(0,4,1), ARIMA(0,4,2), ARIMA(1,4,1), ARIMA(1,4,2)}

BIC table of 4th differencing

Hide

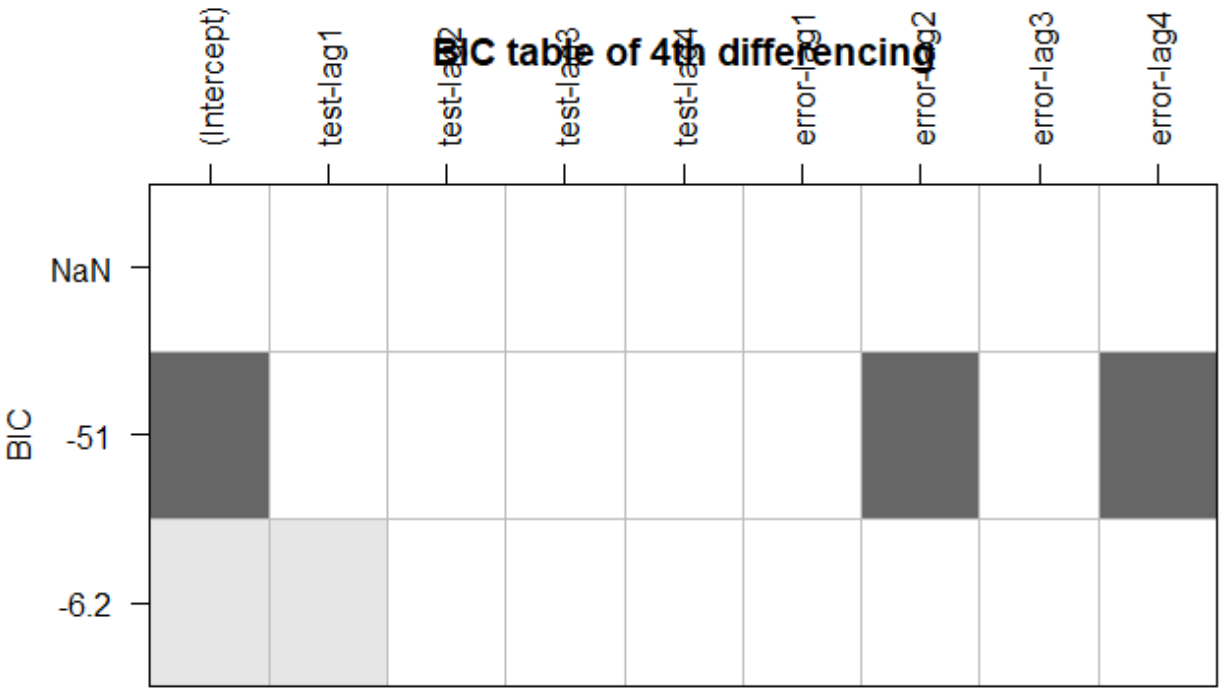
```
BIC_eggs.ols = armasubsets(y=diff4.BC.eggs_ts,nar=4,nma=4,y.name='test',ar.method='ols')
```

```
model order:  6 singularities in the computation of the projection matrix results are only va
lid up to model order 55  linear dependencies foundnvmax reduced to  3
```

Hide

```
plot(BIC_eggs.ols,main="BIC table of 4th differencing")
```

```
NaNs producedNaNs producedNaNs produced
```



- From the BIC table we can consider the models {ARIMA(0,4,2), ARIMA(0,4,4)}
- I'm not including ARIMA(0,4,4) considering the length of the series
- The final set of possible models are {ARIMA(0,4,1), ARIMA(0,4,2), ARIMA(1,4,1), ARIMA(1,4,2)}

Parameter estimation

- We will be doing parameter estimation using two methods to see the consistency i.e. CSS (conditional sum of squares) and ML (maximum likelihood).

ARIMA(0,4,1) - CSS

Hide

```
model_041_css = arima(diff4.BC.eggs_ts,order=c(0,4,1),method='CSS')
coeftest(model_041_css)
```

```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ma1 -0.71507    0.19460 -3.6746 0.0002382 ***
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1
```

ARIMA(0,4,1) - ML

Hide

```
model_041_ml = arima(diff4.BC.eggs_ts,order=c(0,4,1),method='ML')
coeftest(model_041_ml)
```

```
z test of coefficients:

    Estimate Std. Error z value Pr(>|z|)
ma1 -0.93603    0.34371 -2.7233 0.006463 **
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1
```

- Both conditional sum of squares and maximum likelihood methods show us that this model is significant

ARIMA(0,4,2) - CSS

Hide

```
model_042_css = arima(diff4.BC.eggs_ts,order=c(0,4,2),method='CSS')
coeftest(model_042_css)
```

```
z test of coefficients:

    Estimate Std. Error z value Pr(>|z|)
ma1 -0.96402    0.32660 -2.9517  0.00316 **
ma2  0.44644    0.26991  1.6540  0.09812 .
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1
```

- only MA(1) is significant using conditional sum of squares method

ARIMA(0,4,2) - ML

Hide

```
model_042_ml = arima(diff4.BC.eggs_ts,order=c(0,4,2),method='ML')
coeftest(model_042_ml)
```

```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ma1 -1.70750    0.41520 -4.1124 3.915e-05 ***
ma2  0.90912    0.39415  2.3066   0.02108 *
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1
```

- using the maximum likelihood method, both MA(1) and MA(2) are significant.

ARIMA(1,4,1) - CSS

Hide

```
model_141_css = arima(diff4.BC.eggs_ts,order=c(1,4,1),method='CSS')
```

```
possible convergence problem: optim gave code = 1
```

Hide

```
coeftest(model_141_css)
```

```
z test of coefficients:

     Estimate Std. Error  z value  Pr(>|z|)
ar1 -0.728797   0.001113 -654.812 < 2.2e-16 ***
ma1 -2.570176   0.044918  -57.219 < 2.2e-16 ***
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1
```

- using the conditional sum of squares method, AR(1) and MA(1) is significant

ARIMA(1,4,1) - ML

Hide

```
model_141_ml = arima(diff4.BC.eggs_ts,order=c(1,4,1),method='ML')
coeftest(model_141_ml)
```

```
z test of coefficients:

     Estimate Std. Error  z value  Pr(>|z|)
ar1 -0.935358   0.089804 -10.4156 < 2.2e-16 ***
ma1 -0.937656   0.342549  -2.7373  0.006195 **
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1
```

- using the maximum likelihood method, AR(1) and MA(1) is significant

ARIMA(1,4,2) - CSS

Hide

```
model_142_css = arima(diff4.BC.eggs_ts,order=c(1,4,2),method='CSS')
```

```
possible convergence problem: optim gave code = 1
```

Hide

```
coeftest(model_142_css)
```

```
z test of coefficients:

      Estimate Std. Error   z value  Pr(>|z|)
ar1 -0.7353679  0.0012092 -608.1692 < 2.2e-16 ***
ma1 -2.9788233  0.4007513   -7.4331 1.061e-13 ***
ma2  1.6470485  0.8974066    1.8353   0.06645 .
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1
```

- MA(2) is insignificant using conditional sum of squares method and AR(1), MA(1) is significant

ARIMA(1,4,2) - ML

Hide

```
model_142_ml = arima(diff4.BC.eggs_ts,order=c(1,4,2),method='ML')
coeftest(model_142_ml)
```

```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ar1 -0.90516    0.11712 -7.7288 1.086e-14 ***
ma1 -1.64557    0.50634 -3.2499  0.001154 **
ma2  0.82574    0.48150  1.7149  0.086356 .
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1
```

- Again, both AR(1), MA(1) are significant using maximum likelihood method and MA(2) is insignificant
- The models with all coefficients significant are {ARIMA(0,4,1), ARIMA(0,4,2), ARIMA(1,4,1)} and ARIMA(1,4,2) was shown insignificant using CSS and ML method.

Sort score function for AIC and BIC score

Hide

```
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")')
  }
}
```

AIC and BIC score

Hide

```
sort.score(AIC(model_041_ml,model_042_ml,model_141_ml), score = "aic")
```

	df <dbl>	AIC <dbl>
model_141_ml	3	68.97142
model_042_ml	3	73.96107
model_041_ml	2	78.36419

3 rows

Hide

```
sort.score(BIC(model_041_ml,model_042_ml,model_141_ml), score = "bic" )
```

	df <dbl>	BIC <dbl>
model_141_ml	3	69.20974
model_042_ml	3	74.19939
model_041_ml	2	78.52308

3 rows

- The order of scoring in both AIC, BIC is the same. model_141_ml has the lowest AIC and BIC score which is better. I am also considering model_042_ml as it has the second lowest AIC & BIC score.

Residual analysis function

Hide

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

Diagnostic checks

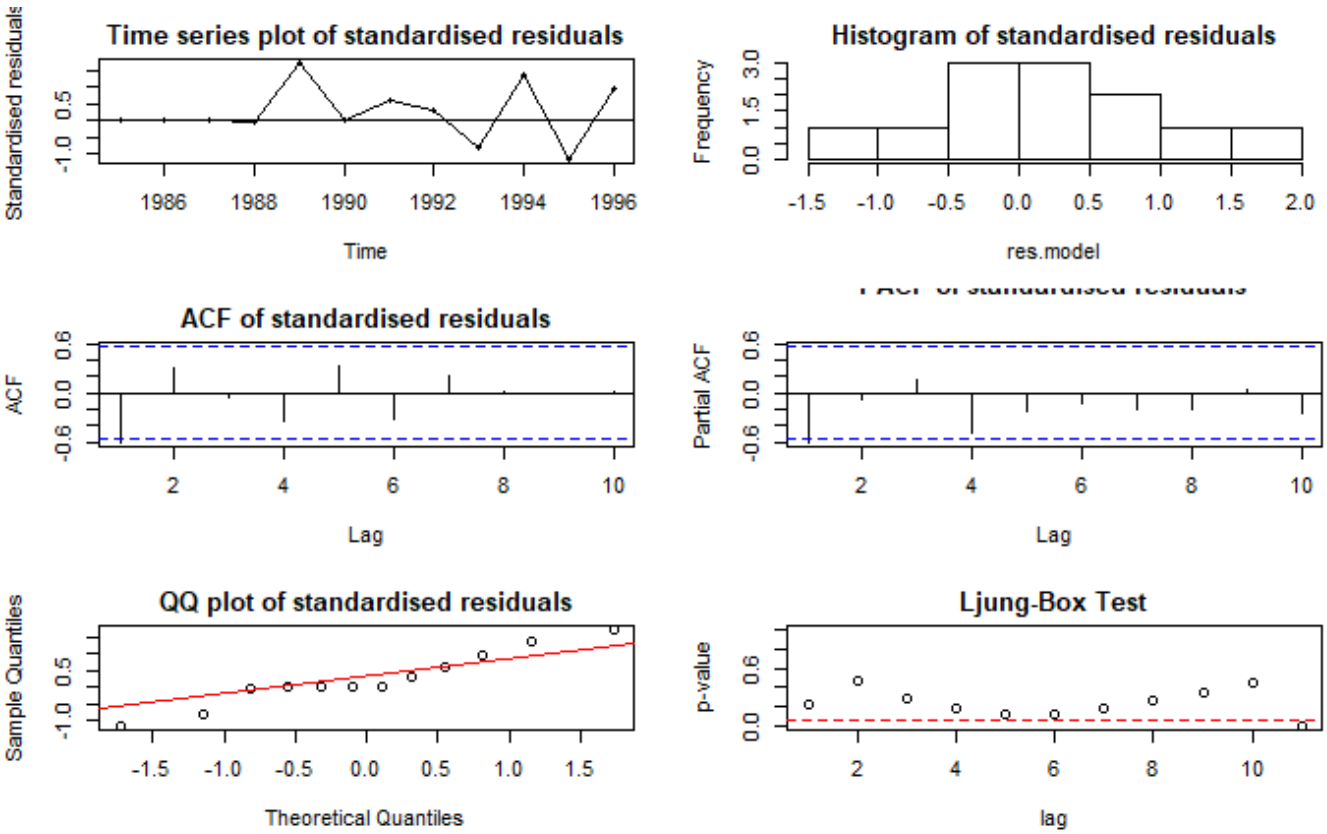
- Performing diagnostic checks for both the models model_042_ml and model_141_ml

Diagnostic check for model_042_ml

Hide

```
residual.analysis(model = model_042_ml)
```

```
    Shapiro-Wilk normality test

data:  res.model
W = 0.94842, p-value = 0.614
```



- Time series plot: There are at least three or four residuals mid-ending of the series with magnitudes larger than 1 and -1 which is unusual in a standard normal distribution. Ideally, we should go back to those years and try to learn what outside factors may have influenced unusually large drops or unusually large increases in the egg depositions.
- Histogram: The histogram is slightly left skewed which doesn't seem normally distributed.
- ACF plot: There is one slightly significant lag. We conclude that the graph shows statistically significant evidence of nonzero autocorrelation in the residuals.
- PACF plot: There is one slightly significant lag which shouldn't be the case, we will confirm this in Ljung-box test
- QQ plot: Very few data-points are falling on the qqline and the data-points at the beginning and end of the line are tailing off.
- Ljung-box test: The estimated ARIMA(0,4,2) model doesn't seem to be capturing the dependence structure of the egg depositions time series quite well. There is one data-point below the 0.05 significance level.

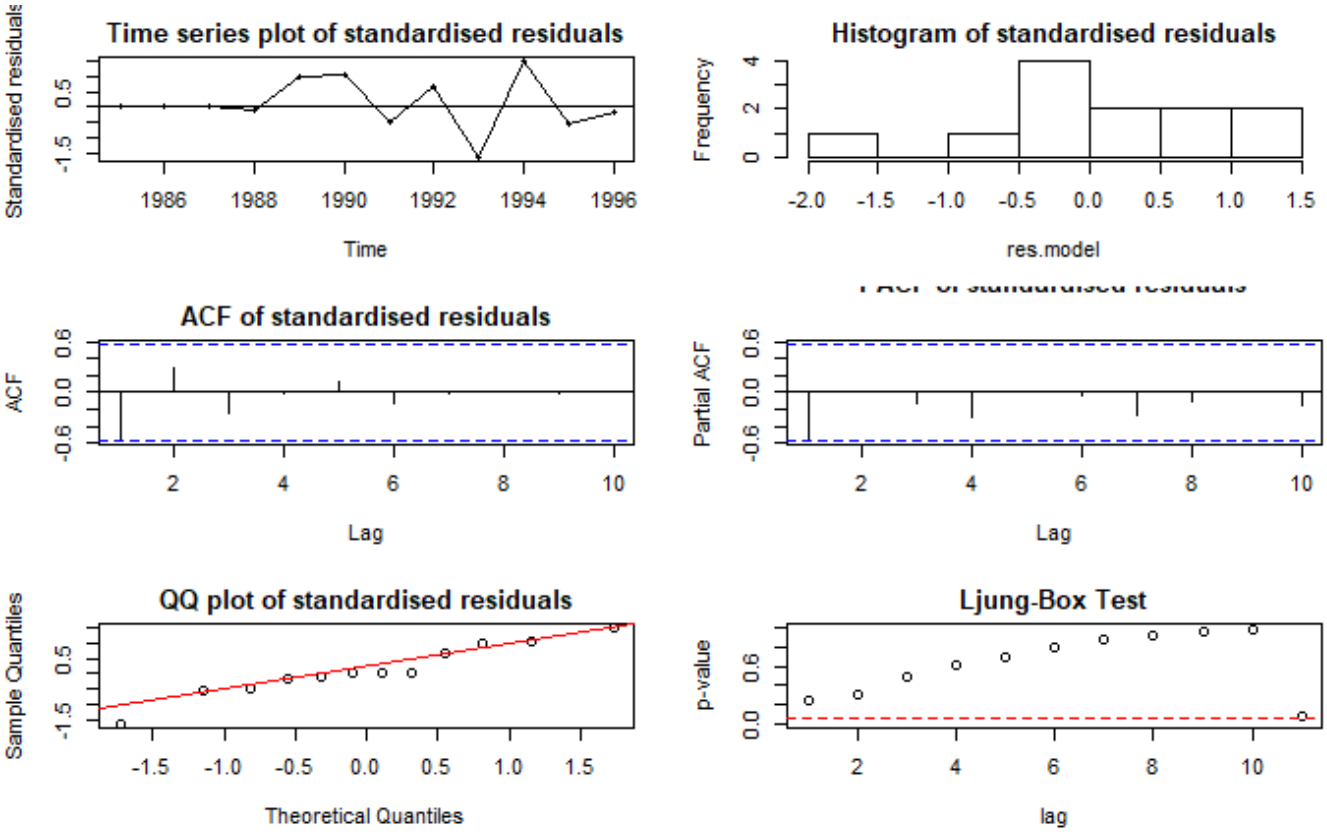
- Shapiro-Wilk test: The p-value is greater than our chosen alpha level of 0.05, hence we are not able to reject the null hypothesis that the data is normally distributed. This could be because we just have 16 datapoints in our dataset.

Diagnostic check for model_141_ml

Hide

```
residual.analysis(model = model_141_ml)
```

```
    Shapiro-Wilk normality test

data:  res.model
W = 0.95083, p-value = 0.6492
```



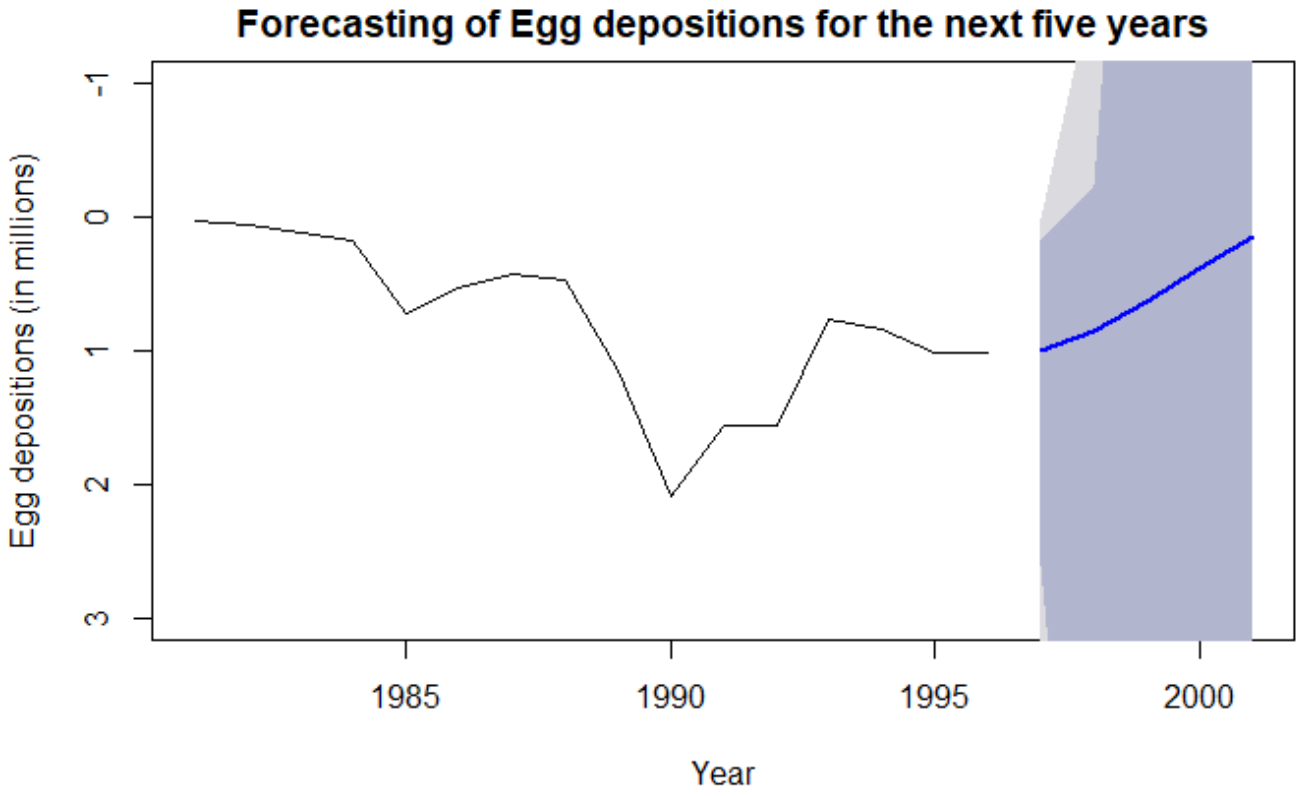
- Time series plot: There are at least two or three residuals towards the end of the series with magnitudes larger than 1 and -1. Ideally, we should go back to those years and try to learn what outside factors may have influenced unusually large drops or unusually large increases in the egg depositions.
- Histogram: The histogram is slightly left skewed which doesn't seem normally distributed.
- ACF plot: There is no trend nor any significant lags in the plot. We conclude that the graph does not show statistically significant evidence of nonzero autocorrelation in the residuals.
- PACF plot: There is no trend nor any significant lags in the plot. There could just be white-noise.
- QQ plot: Most of the data-points are falling on the qqline and there is just one residual which is tailed off in the beginning.
- Ljung-box test: The estimated ARIMA(1,4,1) model seems to be capturing the dependence structure of the egg depositions time series quite well.
- Shapiro-Wilk test: The p-value is greater than our chosen alpha level of 0.05, hence we are accepting the null hypothesis that the data is normally distributed.
- There is no problem in the residuals of ARIMA(1,4,1) model.

Forecasting

- Since there is no problem in the residuals of ARIMA(1,4,1) model, i am considering this as my final model for forecasting. The lambda value is mentioned in the below function considering box-cox transformation applied.

Hide

```
model141fit=Arima(eggs_ts,c(1,4,1),lambda=.45)
plot(forecast(model141fit,h=5),ylim=c(3,-1),xlab='Year',ylab='Egg depositions (in millions)',
main='Forecasting of Egg depositions for the next five years')
```



Conclusion

- Based on the ACF, PACF, EACF, BIC table and diagnostic checks, we have considered ARIMA model (1,4,1). Even, model ARIMA(0,4,2) could have been used for forecasting as apart from few diagnostic checkpoints it was a really good model. However, there seems to be an increasing trend in the egg depositions based on our forecasting of ARIMA model (1,4,1) for the next 5 years (1997-2001).