MATH 1318 - Time Series Assignment

Assignment 1

Anirudhda Pardhi - s3807109

Introduction

Ozone is a naturally occurring molecule. An ozone molecule is made up of three oxygen atoms. It has the chemical formula O3.The ozone layer is the common term for the high concentration of ozone that is found in the stratosphere around 15–30km above the earth's surface. It covers the entire planet and protects life on earth by absorbing harmful ultraviolet-B (UV-B) radiation from the sun.Prolonged exposure to UV-B radiation is linked to skin cancer, genetic damage and immune system suppression in humans and animals, and lower yielding agricultural crops.As per the provided data in this report we will do the analysis of changes in Ozone layer thickness in time period of 1927 to 2016. With the help of time series modelling techniques we aim to predict forecast for next 5 years 2017-2021.

Importing required libraries and reading the data

```
library(readr)
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(plyr)
```

```
##
## Attaching package: 'plyr'
## The following objects are masked from ’package:Hmisc’:
##
##     is.discrete, summarize
```

```r
library(tidyr)
library(magrittr)
```

```
##
## Attaching package: ’magrittr’

## The following object is masked from ’package:tidyr’:
##
##     extract
```

```r
library(dplyr)
```

```
##
## Attaching package: ’dplyr’

## The following objects are masked from ’package:plyr’:
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from ’package:Hmisc’:
##
##     src, summarize

## The following objects are masked from ’package:stats’:
##
##     filter, lag

## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(TSA)
```

```
##
## Attaching package: ’TSA’

## The following object is masked from ’package:readr’:
##
##     spec

## The following objects are masked from ’package:stats’:
##
##     acf, arima
## The following object is masked from 'package:utils':
##
##     tar
```

```
library(tseries)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
data1 <- read.csv("/Users/ADMIN/Desktop/Sem 3/Time Series/Assignment/data1.csv",header=FALSE)
class(data1)
```

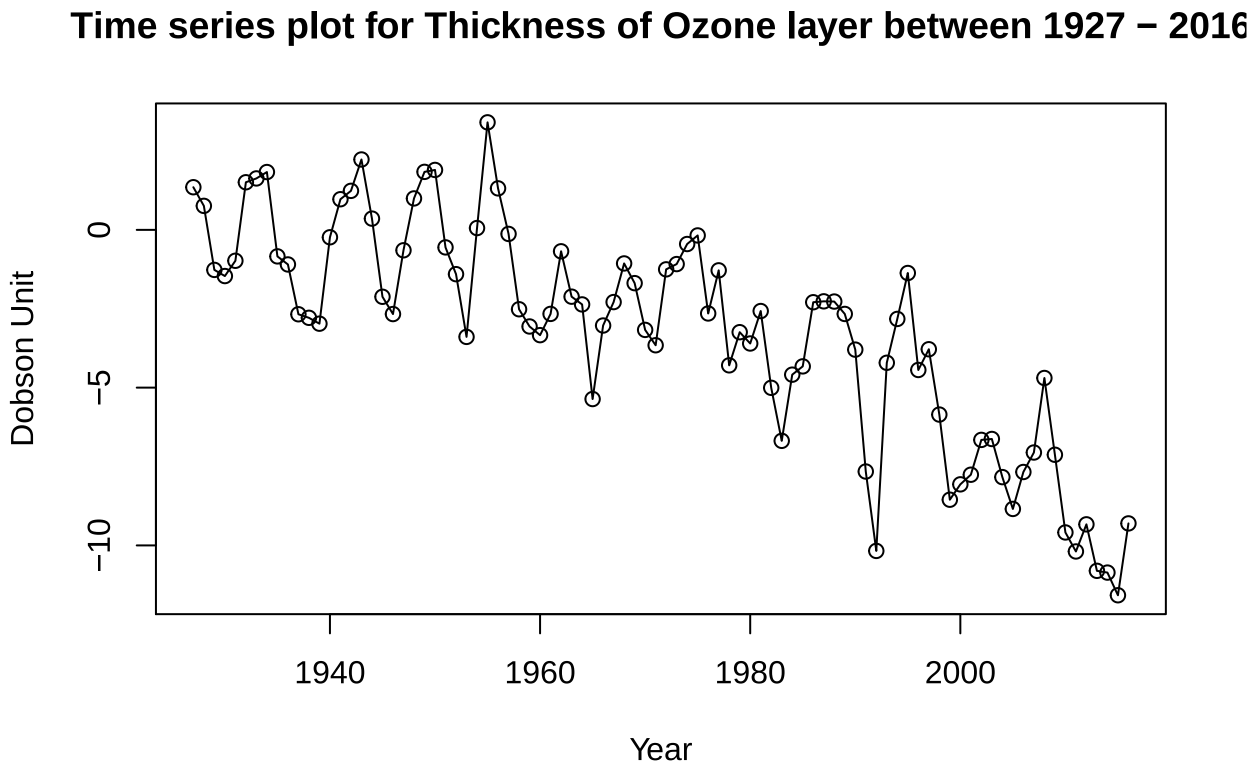
```
## [1] "data.frame"
```

Time series dataplot

```
tsdata <- ts(as.vector(data1), start=1927, end=2016, frequency = 1)
class(tsdata)
```

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

```
plot(tsdata,ylab='Dobson Unit',xlab='Year',type='o',
     main = "Time series plot for Thickness of Ozone layer between 1927 - 2016")
```



From the above time series plot we can observe overall downward trend in the thickness of ozone layer between time period 1927-2016 time series plot. In the above graph -negative Dobson value indicates decrease in the thickness of ozone layer thickness and vice-versa positive Dobson value indicates relative increase in thickness.

Key points obtained from above graph :

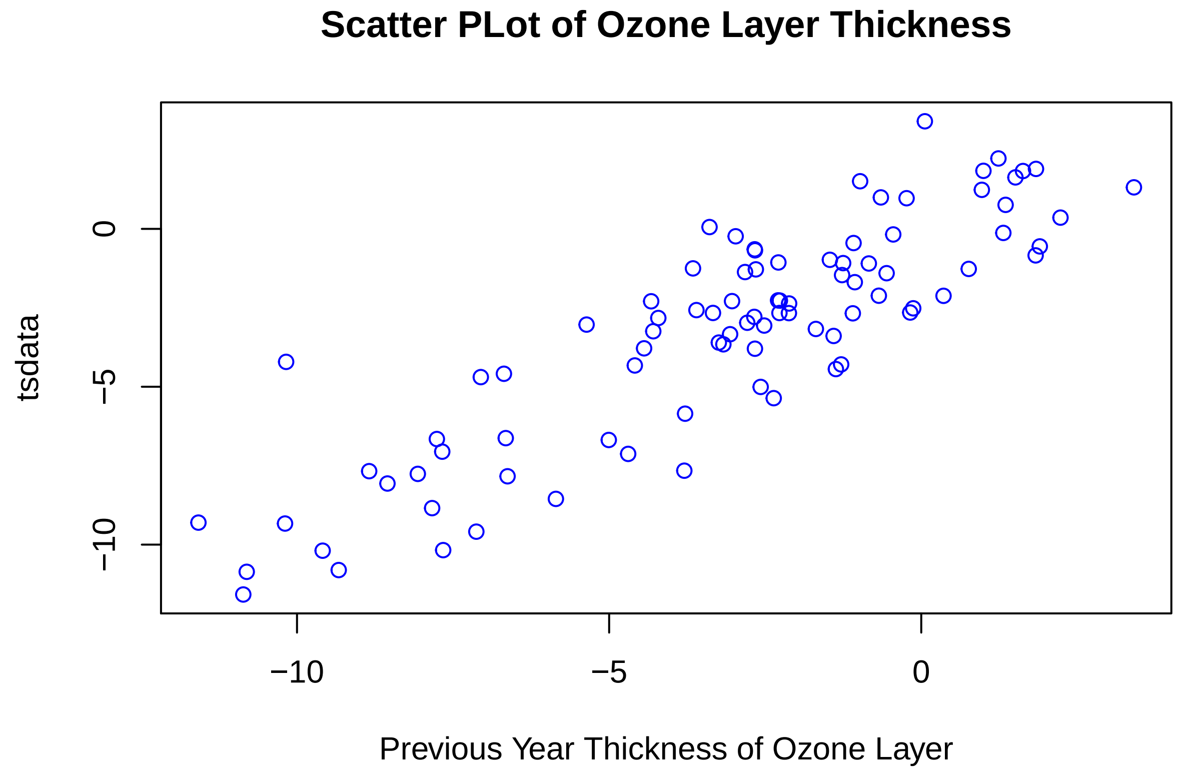
Trend : Gradually downward trend can be clearly observed. Seasonality: There seem to be no sign of seasonality present. Change in Variance: There seems to be no change in variance. Intervention: There seems to be no specific sign of intervention. Behavior: Autoregressive (AR) and Moving Average (MA) behavior .

Check for correlation

```
y = tsdata
x = zlag(tsdata)
index = 2:length(x)
cor(y[index],x[index])
```

```
## [1] 0.8700381
```

```
plot(y=tsdata,x=zlag(tsdata),col=c("blue"),
     xlab = "Previous Year Thickness of Ozone Layer",main = "Scatter PLot of Ozone Layer Thickness")
```



From the above scatterplot we can see that strong correlation exist among the observations which indicates every single observation significantly depends on previous year's observation. Calculated Value of correlation is 0.87 which also support strong correlation factor.

Linear modelling

```
lmmodel = lm(tsdata~time(tsdata))
summary(lmmodel)
```

```
##
## Call:
## lm(formula = tsdata ~ time(tsdata))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7165 -1.6687  0.0275  1.4726  4.7940
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  213.720155  16.257158   13.15   <2e-16 ***
## time(tsdata)  -0.110029   0.008245  -13.34   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.032 on 88 degrees of freedom
## Multiple R-squared:  0.6693, Adjusted R-squared:  0.6655
## F-statistic: 178.1 on 1 and 88 DF,  p-value: < 2.2e-16
```

Intercept value = 213.72 Slope value = -0.11

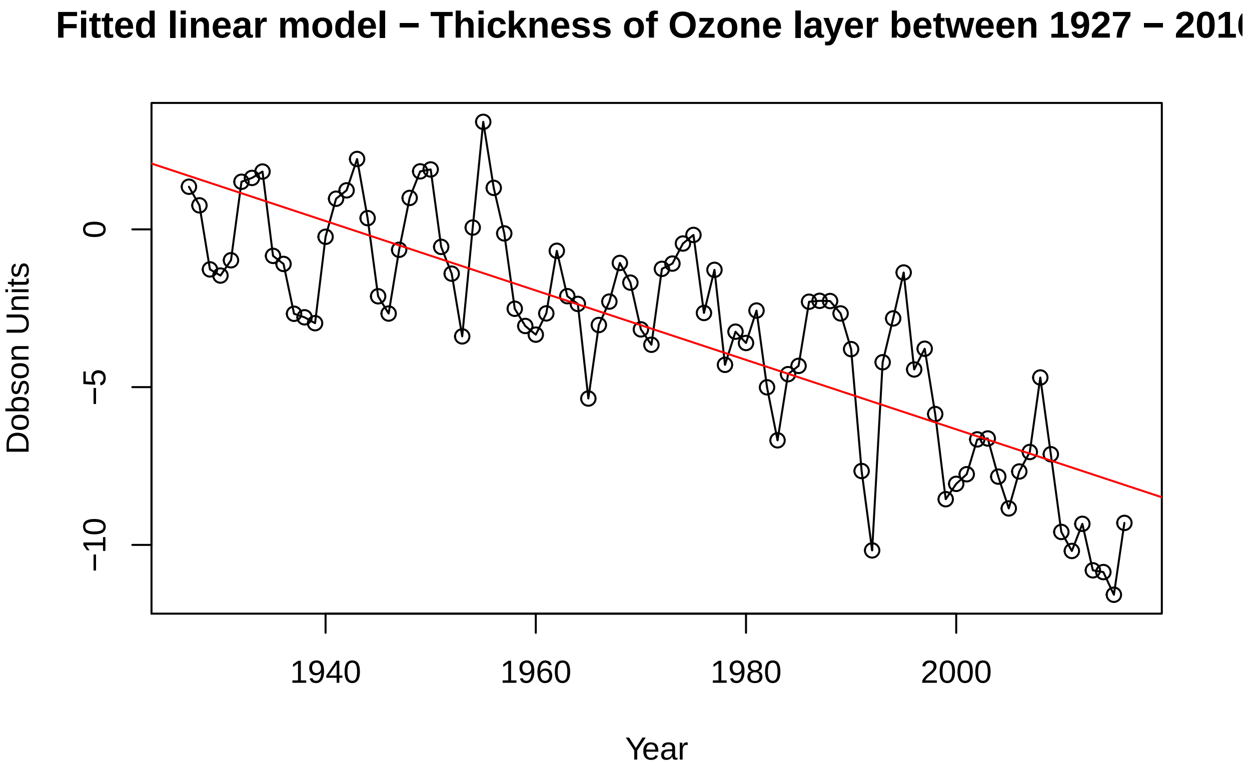
Both slope and intercept p value are less than 0.05 which indicates they are statistically significant at 5% level of significance.

p value of linear model is less than 0.05 which indicates our model is statistically significant at 5% level of significance.

R-squared value or Coefficient of determination gives us information about goodness of fit of a model. In our case its value is 0.66 which is not so good or in other word our model is partially significant.

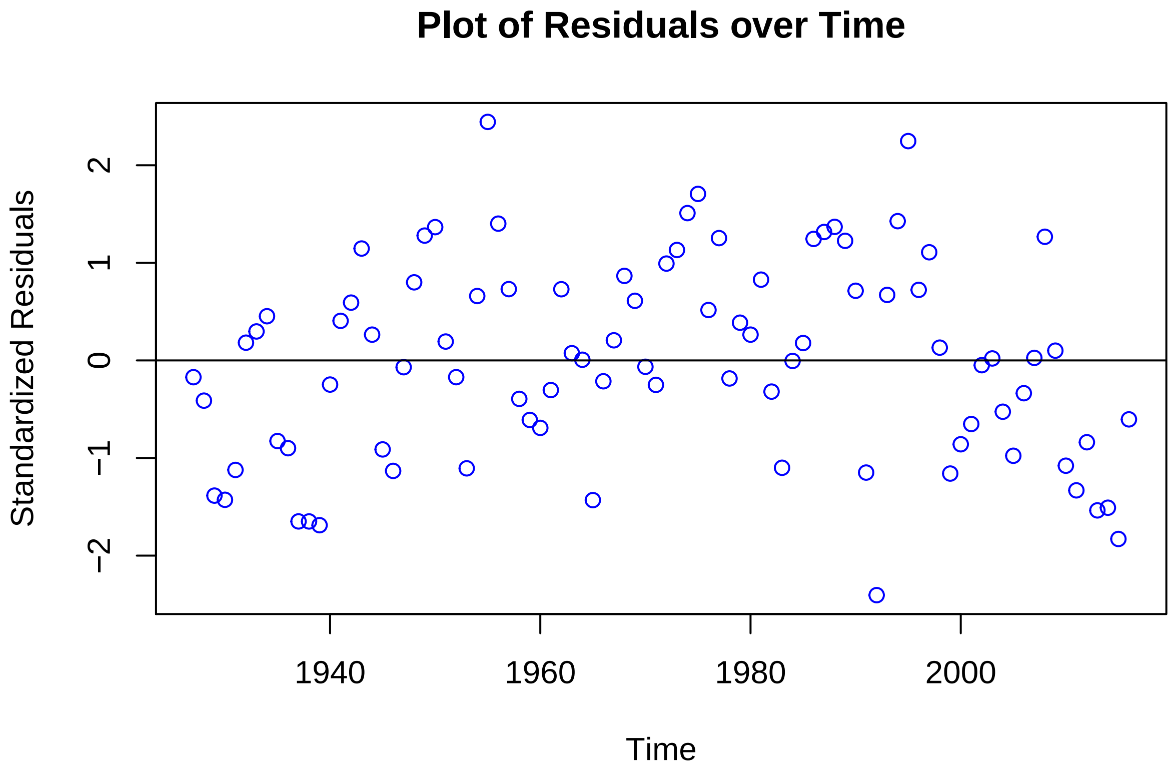
Plot for Linear modelling

```
plot(tsdata, ylab=’Dobson Units’,xlab=’Year’,type=’o’,
     main = "Fitted linear model - Thickness of Ozone layer between 1927 - 2016")
abline(lmmodel,col="Red")
```

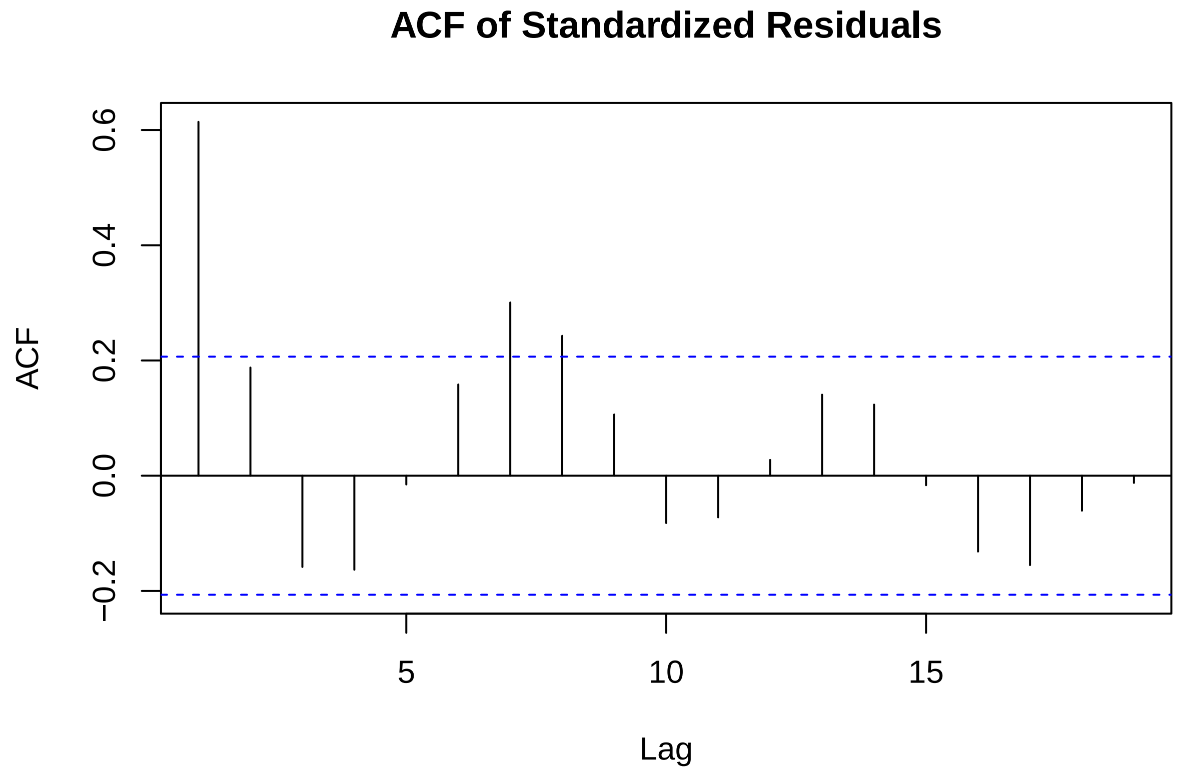


## Residual analysis of linear model

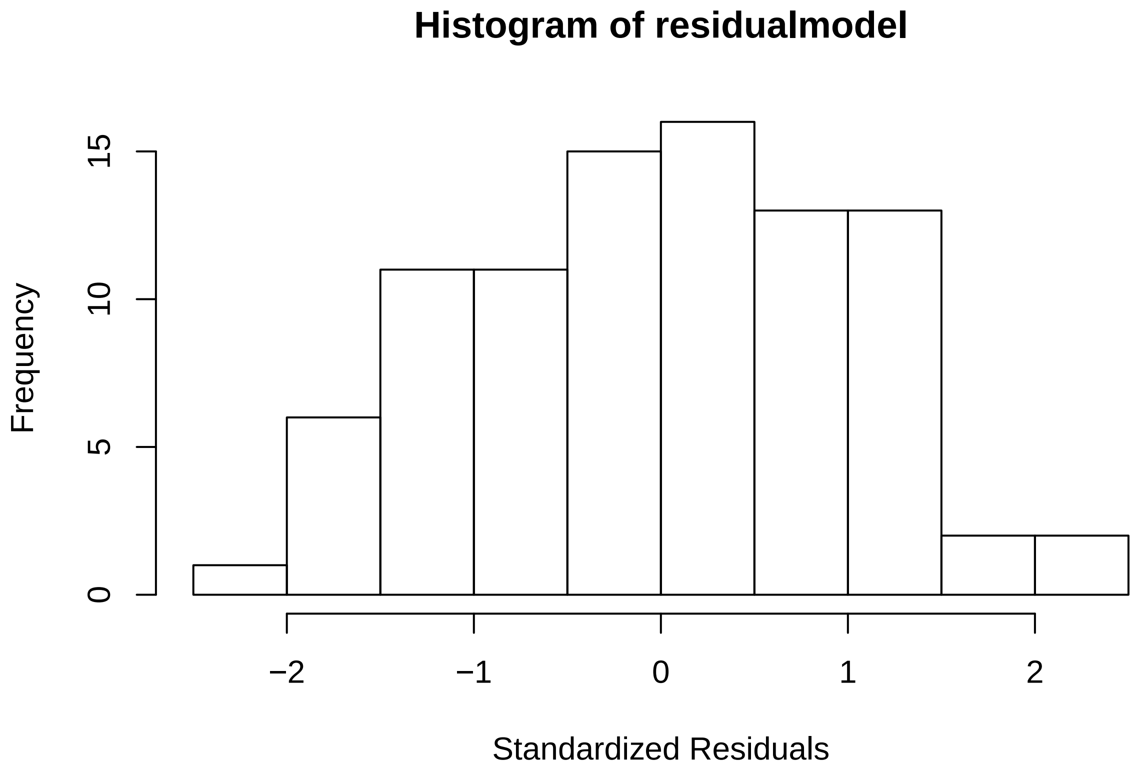
```
#Scatterplot for residual model
residualmodel = rstudent(lmmodel)
plot(y = residualmodel, x = as.vector(time(tsdata)),
    xlab = 'Time', ylab='Standardized Residuals',type='p',col=c("blue"),main = "Plot of Residuals over '
abline(h=0)
```



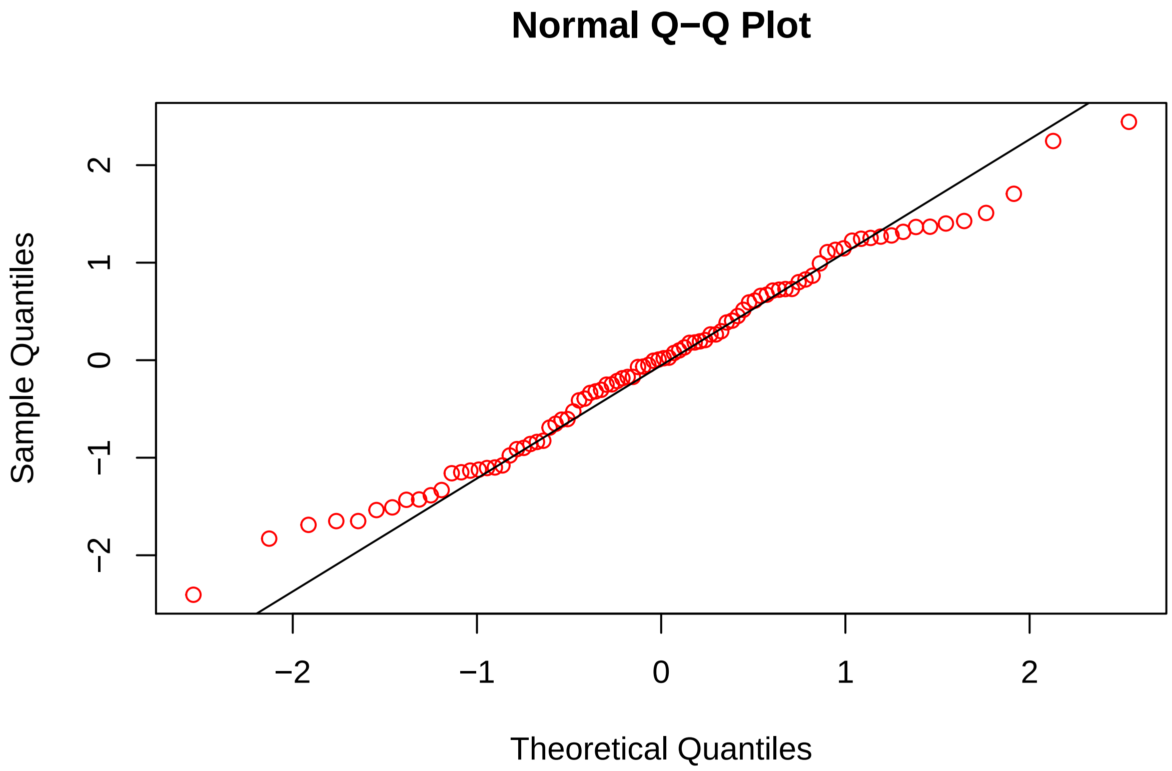
```
#ACF of Standardized Residuals
acf(residualmodel,main="ACF of Standardized Residuals")
```



```
#Histogram of Standardized Residuals
hist(residualmodel,xlab='Standardized Residuals')
```



```
#QQplot of Standardized Residuals
qqnorm(residualmodel,col=c("red"))
qqline(residualmodel)
```



```
#shapiro test
shapiro.test(residualmodel)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residualmodel
## W = 0.98733, p-value = 0.5372
```

Key points from Residual Analysis of Linear model :

1) Plot of Residuals over time: At the first glance of scatterplot, we can say there is some sort of randomness but if we closely observe its seems that above x-axis (x=0) there are more data points than below which indicates that the scatterplot is not completely random.

2) Autocorrelation Function (ACF): In ACF we check for early lags.Before lag = 5 we can observe that 2 correlation values are slightly touching significant confidence boundaries hence we can comprehend that stochastic component of time series is not complete white noise.

3) Quantile-quantile (QQ-plot): In QQ plot we can observe that the observation points are deviating away from both extreme ends of line which indicates it is not completely showing white noise behavior.

4) Histogram: Histogram plot for residual is slightly symmetric.

5) Shapiro-Wilk Test: Obtained p value = 0.53 is greater than 0.05 which is implying that the distribution of the data are not significantly different from normal distribution. Thus we can assume the normality.

We want our residual to be White noise so that our model contains most of the meaningful data or information but since in this case Residuals are not showing completely white noise behavior we can comprehend that our linear model is not suitable for our time series data.

Quadratic modelling

```
t = time(tsdata)
t2 = t^2
quad_model = lm(tsdata ~ t + t2)
summary(quad_model)
```

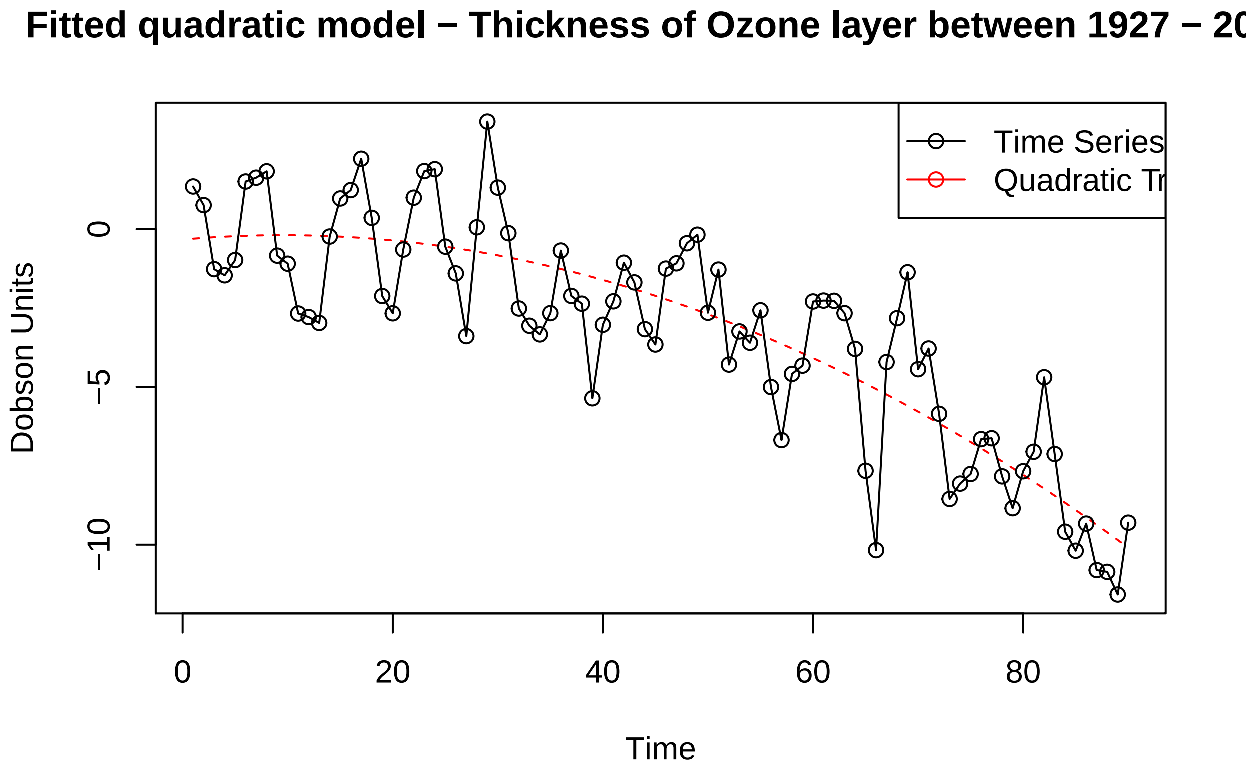
```
##
## Call:
## lm(formula = tsdata ~ t + t2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1062 -1.2846 -0.0055  1.3379  4.2325
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.733e+03  1.232e+03  -4.654 1.16e-05 ***
## t            5.924e+00  1.250e+00   4.739 8.30e-06 ***
## t2          -1.530e-03  3.170e-04  -4.827 5.87e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.815 on 87 degrees of freedom
## Multiple R-squared:  0.7391, Adjusted R-squared:  0.7331
## F-statistic: 123.3 on 2 and 87 DF,  p-value: < 2.2e-16
```

p-value of linear model is less than 0.05 which indicates our model is statistically significant at 5% level of significance.

R-squared value or Coefficient of determination gives us information about goodness of fit of a model. In our case its value is 0.73 which is good and more significant, fitting than the linear model.

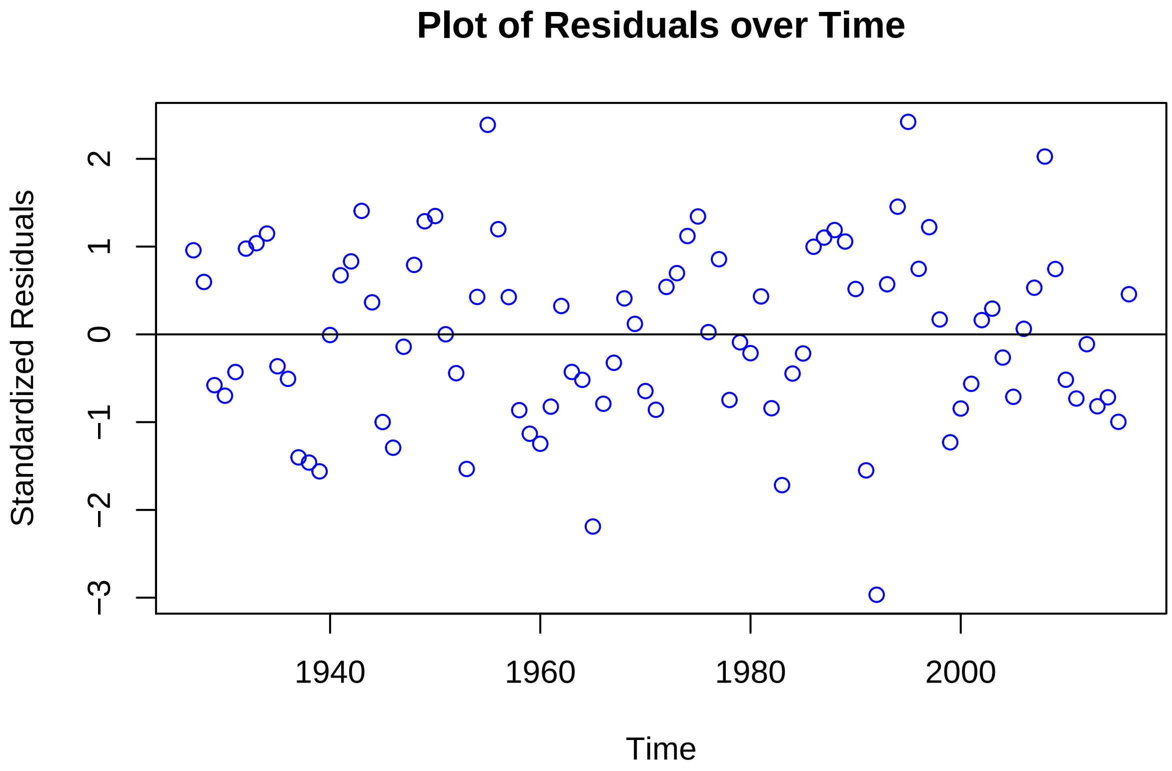
Plot of quadratic model

```
plot(ts(fitted(quad_model)), ylim = c(min(c(fitted(quad_model),
   as.vector(tsdata))), max(c(fitted(quad_model),as.vector(tsdata)))),
  ylab='Dobson Units' , main = "Fitted quadratic model - Thickness of Ozone layer between 1927 - 2016",
     type="l",lty=2,col="Red")
lines(as.vector(tsdata),type="o")
legend ("topright", lty = 1, pch = 1, col = c("black","red"), text.width = 15,
        c("Time Series Plot","Quadratic Trend Line"))
```

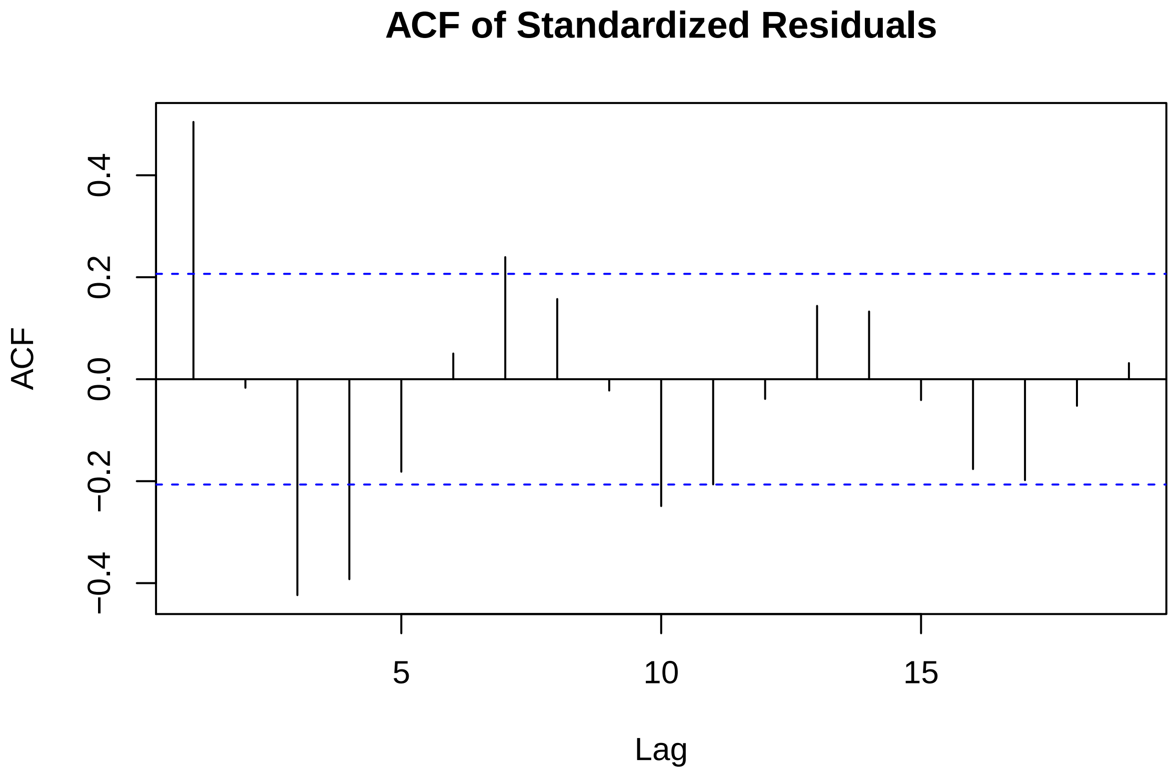


## Residual analysis of quadratic model

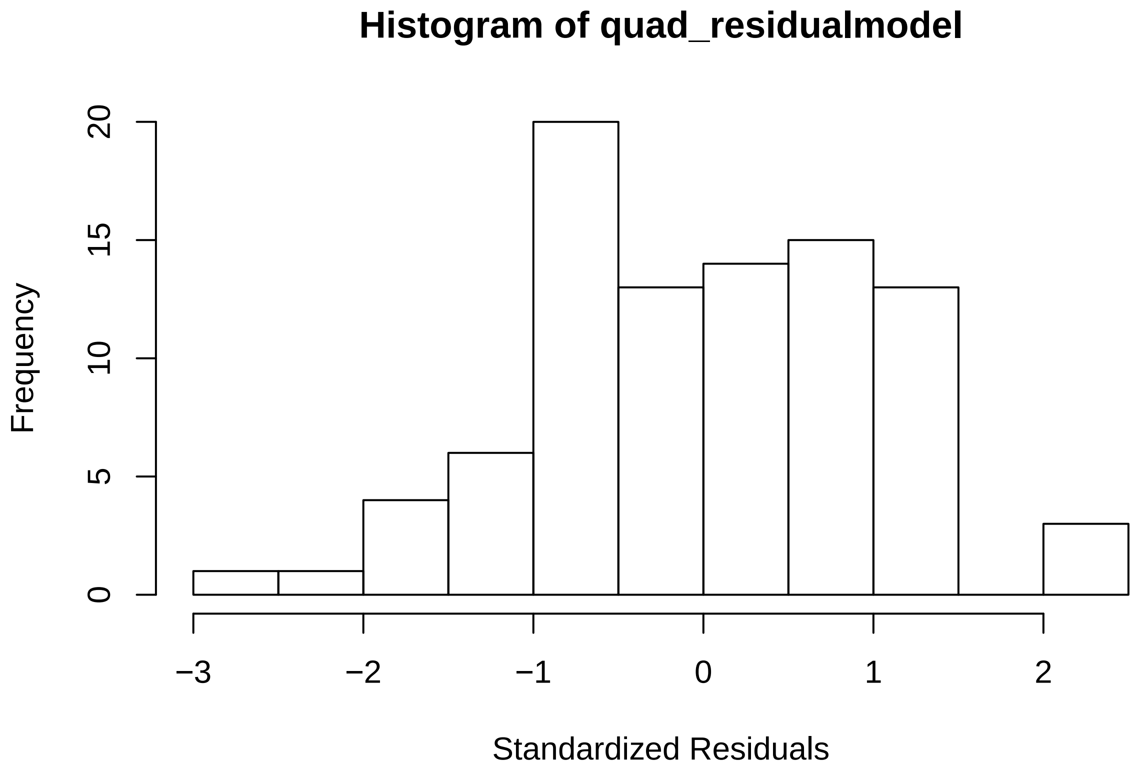
```
quad_residualmodel = rstudent(quad_model)
plot(y = quad_residualmodel, x = as.vector(time(tsdata)),
     xlab = 'Time', ylab='Standardized Residuals',type='p',col=c("blue"),main = "Plot of Residuals over
abline(h=0)
```



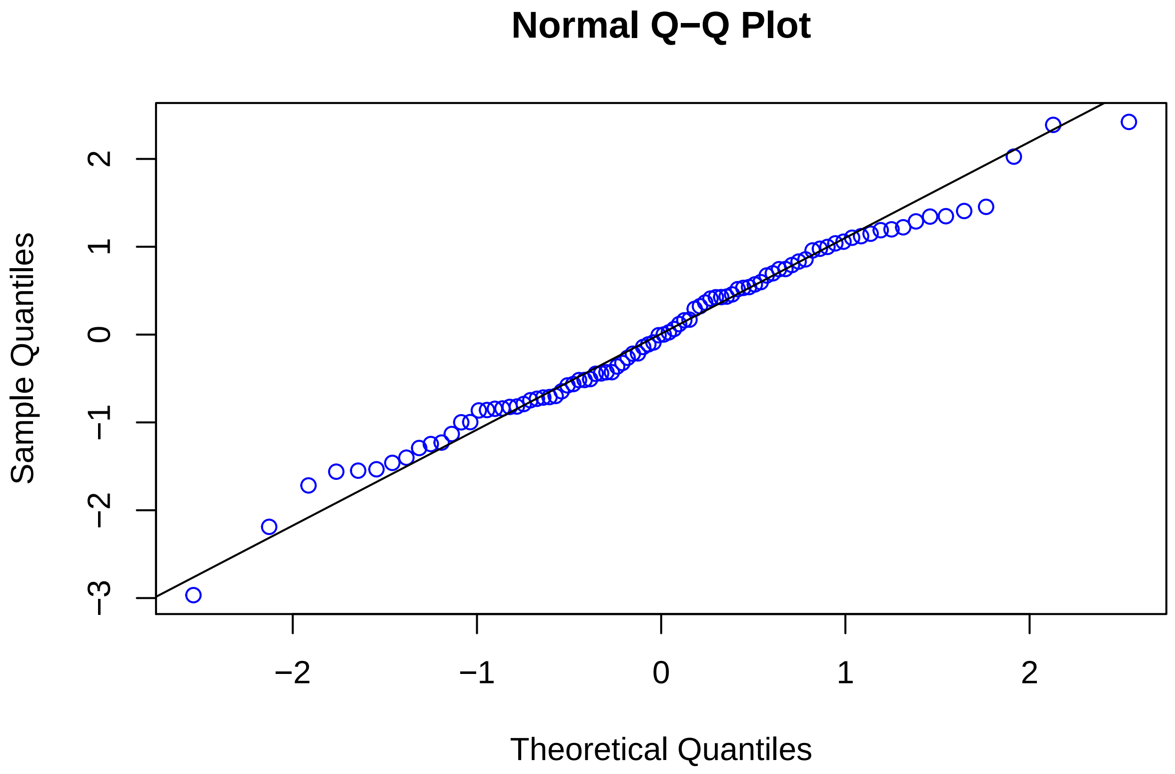
```
#ACF of Standardized Residuals
acf(quad_residualmodel,main="ACF of Standardized Residuals")
```



```
#Histogram
hist(quad_residualmodel,xlab='Standardized Residuals')
```



```
#QQplot
qqnorm(quad_residualmodel,col=c("blue"))
qqline(quad_residualmodel)
```



```
#shapiro test
shapiro.test(quad_residualmodel)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  quad_residualmodel
## W = 0.98889, p-value = 0.6493
```

1) Plot of Residuals over time: On the first glance of scatterplot we can say there is randomness in graph.
2) Autocorrelation Function (ACF): In ACF we check for early lags.Before lag = 5 we can observe that 2 correlation values are crossing significant confidence boundaries hence we can comprehend that stochastic component of time series is not complete white noise.
3) Quantile-quantile (QQ-plot): In QQ plot we can observe that very few observation points are deviating away from extreme ends of line which indicates it is not completely showing white noise behavior.
4) Histogram: Histogram plot for residual is moderately symmetric.
5) Shapiro-Wilk Test: Obtained p value 0.64 is greater than 0.05 which is implying that the stochastic component of the quadratic model is normally distributed.

We want our residual to be white noise so that our model contains most of the meaningful data or information but since in this case Residuals are not showing completely white noise behavior we can comprehend that our quadratic model is not completely significant for our time series data.

If we look at the r squared value of quadratic model, it is evident that Quadratic model is more significant than linear model.

Harmonic model

```
har_model=harmonic(tsdata,0.45)
harmonic_model=lm(tsdata ~ har_model)
summary(harmonic_model)
```

```
##
## Call:
## lm(formula = tsdata ~ har_model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3520 -1.8905  0.4837  2.3643  6.4248
##
## Coefficients: (1 not defined because of singularities)
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -2.970e+00  4.790e-01  -6.199 1.79e-08 ***
## har_modelcos(2*pi*t)         NA         NA      NA       NA
## har_modelsin(2*pi*t)  5.462e+11  7.105e+11   0.769    0.444
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.522 on 88 degrees of freedom
## Multiple R-squared:  0.006672,   Adjusted R-squared:  -0.004616
## F-statistic: 0.5911 on 1 and 88 DF,  p-value: 0.4441
```

p value of harmonic model 0.44 is greater than 0.05 which indicates our model is statistically insignificant at 5% level of significance.

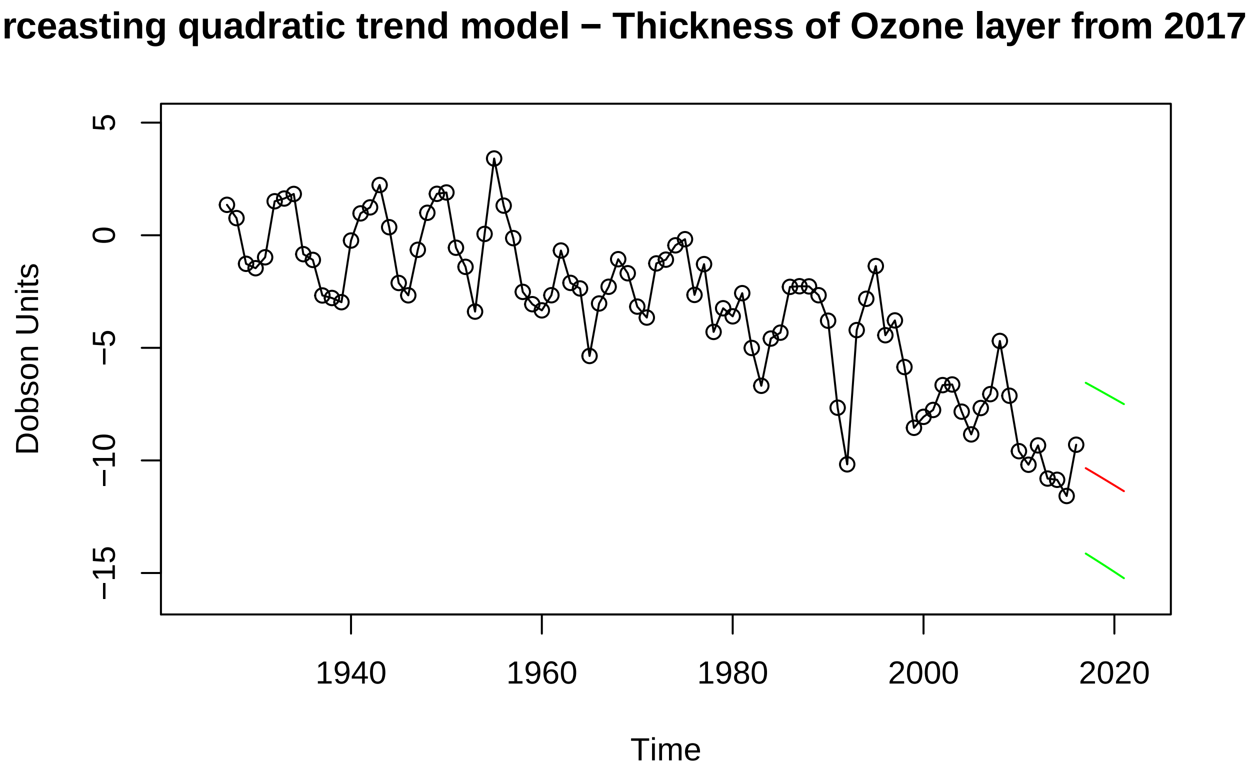
R-squared value or Coefficient of determination gives us information about goodness of fit of a model. In our case its value is 0.006 which is very poor. Hence we can conclude that harmonic model is insignificant.

Forecasting

```
t = c(2017, 2018, 2019, 2020, 2021)
t2 = t^2
new = data.frame(t,t2)
forecast = predict(quad_model, new, interval = "prediction")
print(forecast)
```

```
##         fit       lwr       upr
## 1 -10.34387 -14.13556 -6.552180
## 2 -10.59469 -14.40282 -6.786548
## 3 -10.84856 -14.67434 -7.022786
## 4 -11.10550 -14.95015 -7.260851
## 5 -11.36550 -15.23030 -7.500701
```

```
plot(tsdata, xlim = c(1924,2022), ylim = c(-16, 5), type="o", ylab='Dobson Units' ,
  main = " Forceasting quadratic trend model - Thickness of Ozone layer from 2017 to 2021",)
  lines(ts(as.vector(forecast [,1]), start = c(2017,1), frequency = 1), col="red", type="l")
  lines(ts(as.vector(forecast [,2]), start = c(2017,1), frequency = 1), col="green", type="l")
lines(ts(as.vector(forecast [,3]), start = c(2017,1), frequency = 1), col="green", type="l")
```

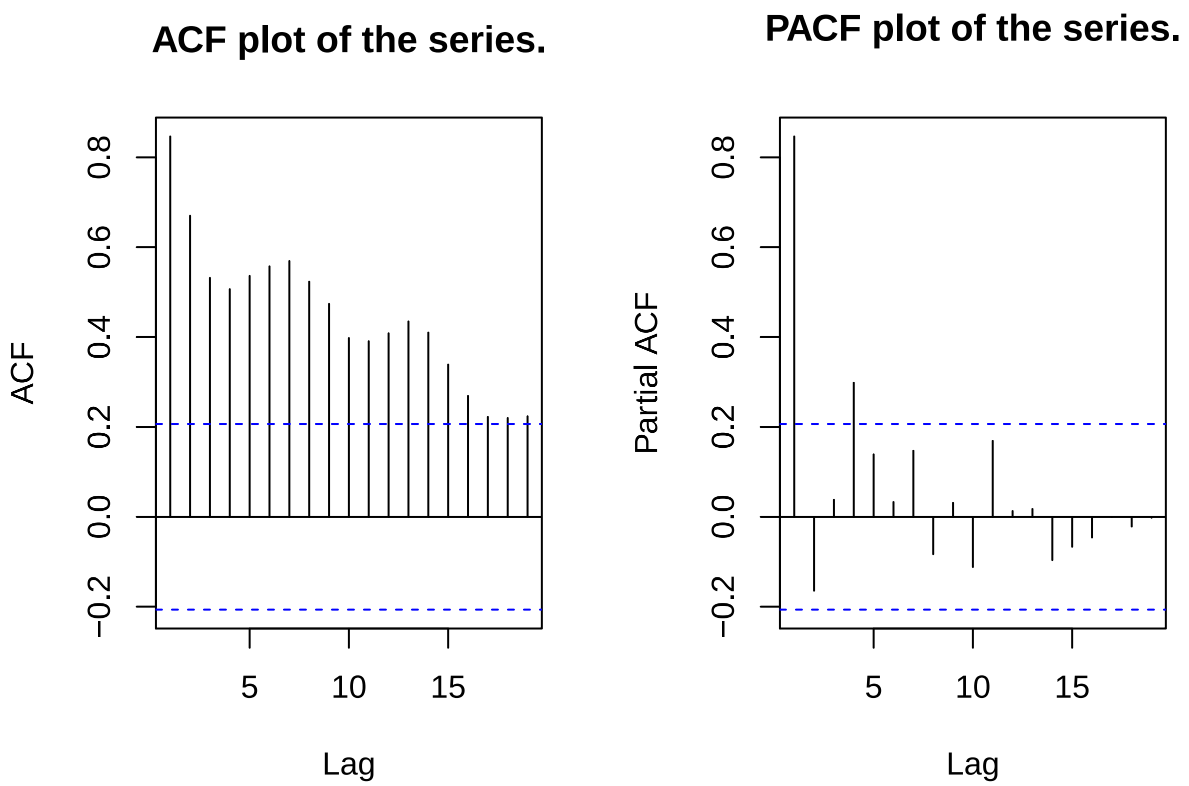


As per the earlier discussion we reach on to the conclusion that quadratic model is the best suited among linear model, quadratic model and harmonic model. So we did forecasting for next 5 years 2017-2021 with quadratic model. As per the above graph we can see that it is showing downward trend which implies there will be a gradual decrease in the thickness of ozone layer in upcoming 5 years. Red line is the forecasting line and blue lines are showing 5% forecast limit range.

Task 2

In this section we are going to find possible solution set of ARIMA models using model specification tool such as ACF-PACF, EACF, BIC table.

```
par(mfrow=c(1,2))
acf(tsdata, main ="ACF plot of the series.")
pacf(tsdata, main ="PACF plot of the series.")
```



By looking at the ACF plot we can say that graph is fluctuating and decreasing in nature which indicates that the series is Non-stationary.

```
adf.test(tsdata,k=4, alternative = c("stationary"))
```

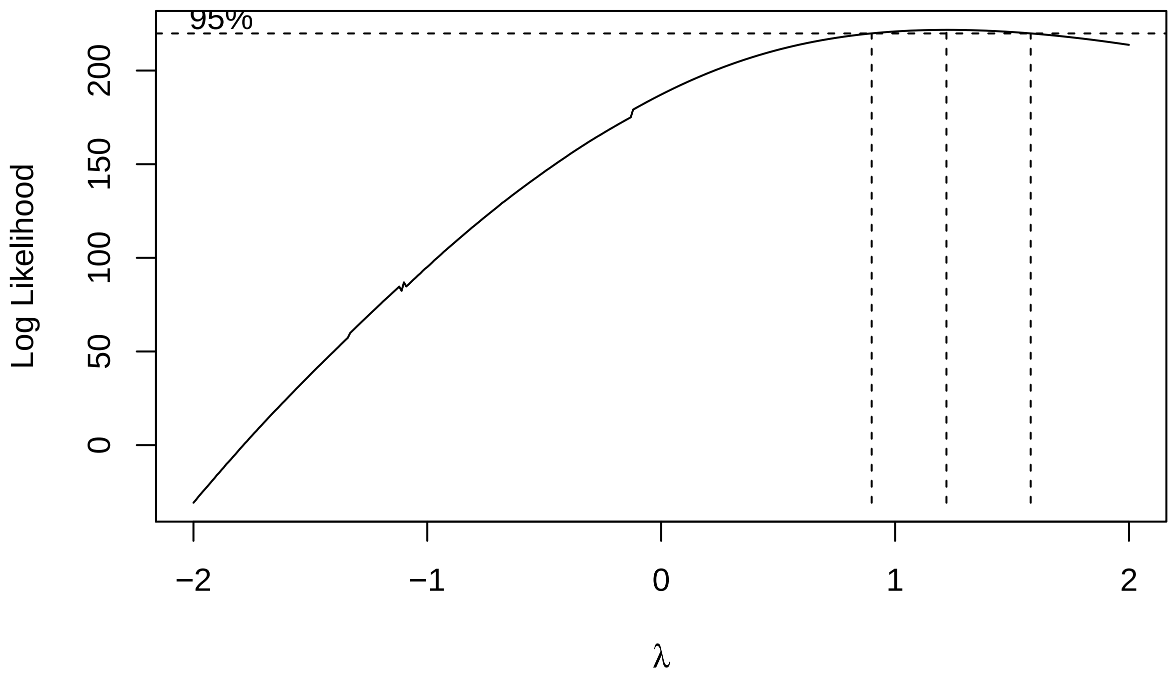
```
##
##  Augmented Dickey-Fuller Test
##
## data:  tsdata
## Dickey-Fuller = -3.2376, Lag order = 4, p-value = 0.0867
## alternative hypothesis: stationary
```

In our series downward trend is present and with the help of ACF plot also it is evident that our series is Non-stationary. We performed Dickey-Fuller test and obtained p-value 0.08 which is greater than significant value 0.05 and confirm that our series is Non-stationary.

Trend presence screens out the possibility of MA (change in variance) so it is good practice to convert non-stationary series to stationary time series.

In order to do so we can use log likelihood vs lambda graph to find optimal value of lambda so that we can check which transformation will be best suited. Adding a constant positive value does not affect series autocorrelation value.Since our time-series data contains negative value so first we need to add some constant in order to make series positive.

```
# Box-cox Transformation
tsdata_positive = tsdata + abs(min(tsdata))+1
tsdata_transform = BoxCox.ar(tsdata_positive, lambda = seq(-2,2, 0.01))
```



```
tsdata_transform$ci
```

```
## [1] 0.90 1.58
```

```
lambda= mean(tsdata_transform$ci)
lambda
```

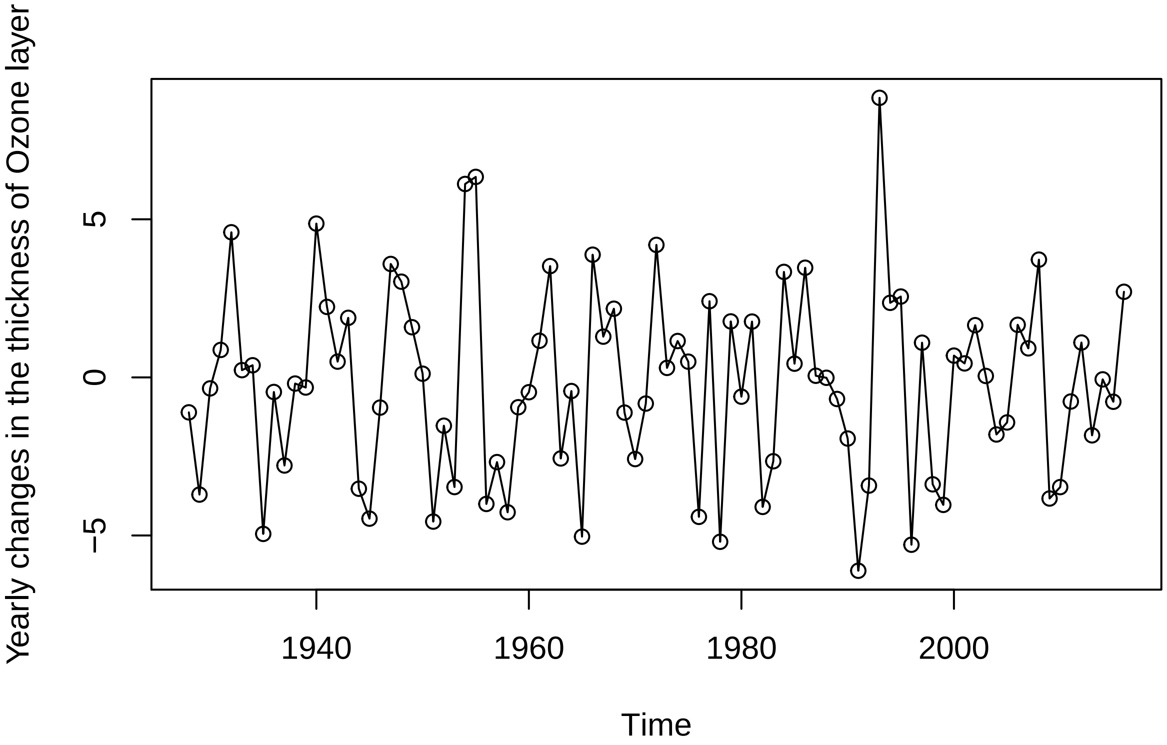
```
## [1] 1.24
```

```
BC.tsdata = (tsdata_positive^lambda-1)/lambda
```

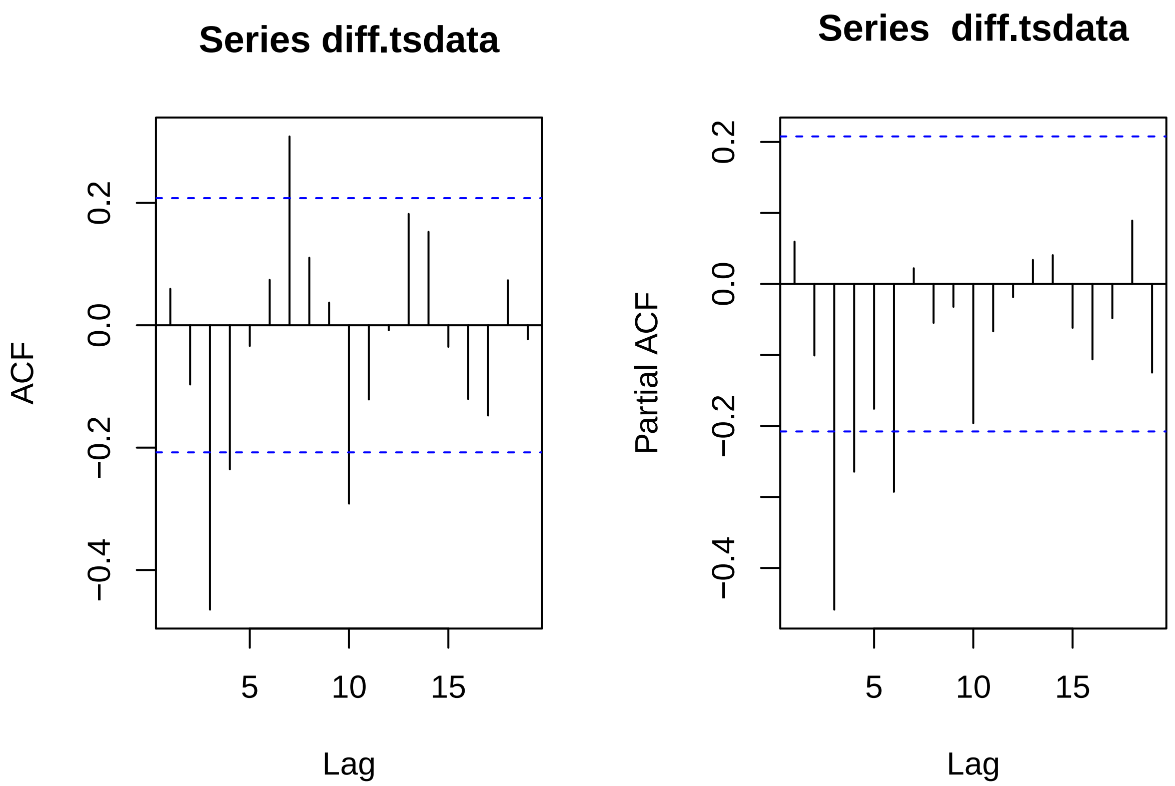
In our case value of lambda is close to 1 which indicates data is already normally distributed and box-cox transformation will not be much helpful for us so will use Difference method.

Differencing method - Transforming non-stationary series to stationary

```
# Differencing method
diff.tsdata = diff(BC.tsdata,differences=1)
plot(diff.tsdata,type='o',
     ylab='Yearly changes in the thickness of Ozone layer')
```



```r
par(mfrow=c(1,2))
acf(diff.tsdata)
pacf(diff.tsdata)
```



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

After differencing transformation method application on our series we can observe that now trend is absent. In ACF we check for early lags.Before lag = 5 we can observe that 2 correlation values are crossing significant confidence boundaries and no gradually decreasing pattern is found hence we can conclude that now series is Stationary. To confirm this we will perform Dickey-Fuller unit root Test.

```r
adf.test(diff.tsdata)
```

```
## Warning in adf.test(diff.tsdata): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff.tsdata
## Dickey-Fuller = -7.2585, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

Obtained p-value 0.01 is less than significant value 0.05 which confirm that our series is Stationary.

EACF table

```
eacf(diff.tsdata)
```

```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 o o x x o o x o o x o  o  o  o
## 1 x o x o o o x o o x o  o  o  o
## 2 o o x o o o x o o x o  o  o  o
## 3 x o x o o x o o o o o  o  o  o
## 4 x o o x o x o o o o o  o  o  o
## 5 x x x x o x o o o o o  o  o  o
## 6 o o o x o o o o o o o  o  o  o
## 7 x o o x o o o o o o o  o  o  o
```

EACF gives the probable list of ARMIMA models.If we look from top left we can see that below is the list of probable models :

(0.1,1) , (0,1,3) (1,1,1) , (1,1,3) (2,1,1) , (2,1,3) (3,1,1) , (3,1,3)

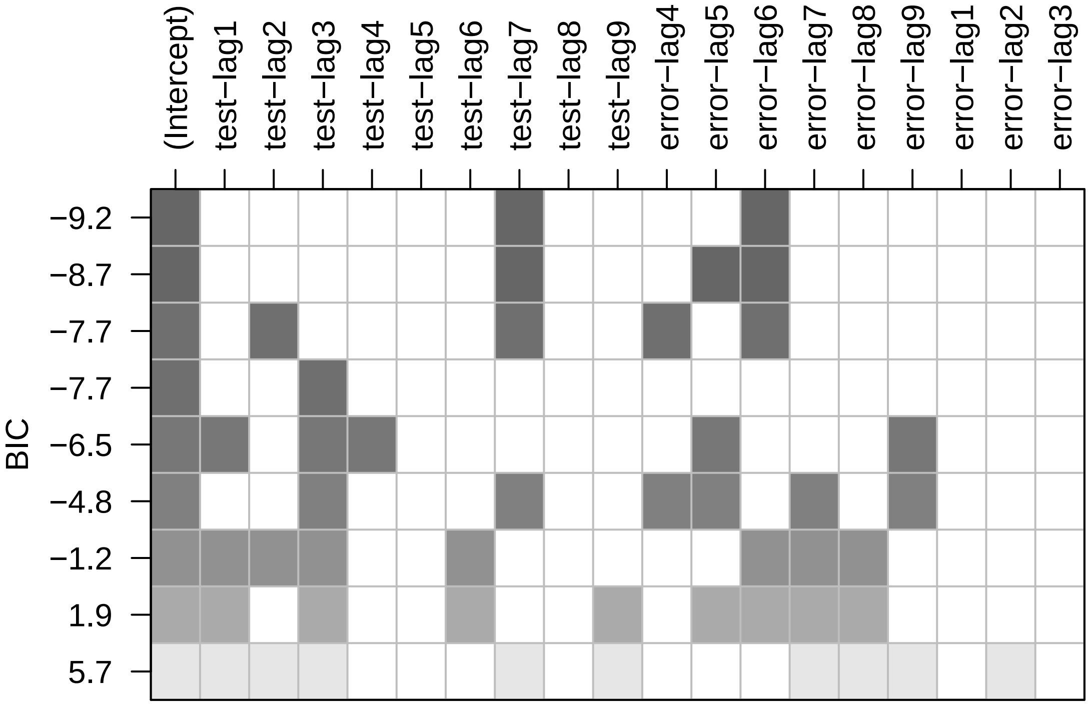
BIC table

```
tsdata_BIC = armasubsets(y=diff.tsdata,nar=9,nma=9,y.name=’test’,ar.method=’ols’)
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 3 linear dependencies found
```

```
## Reordering variables and trying again:
```

```
plot(tsdata_BIC)
```



Conclusion

In ACF graph which gives AR value, we can observe that 2 correlation values were crossing significant confidence boundaries around lag = 5. Similarly in PACF graph which gives MA values, we can observe that 3 correlation values were crossing significant confidence boundaries around lag = 5. Since we have used differencing method so value of I = 1 hence maximum value of AR could be 2 and maximum value of MA could be 3 in ARIMA model.

Based on the ACF and PACF graph along with the earlier discussed candidates from BIC table and EACF table we can conclude that the 2 best possible ARIMA models are :

(2,1,3) and (3,1,3).