Assignment 1

Namita Chhibba

24 March 2018

Introduction

The following Time Series Analysis is concerning changes in the thickness of the ozone layer during the time period from 1927 to 2016. The analysis is focussed on fitting the best trend model for the series and predict the values for the thickness of the ozone layer for next 5 years based on the analysis.

Loading Packages

```
## Loading required package: leaps
```

```
## Loading required package: locfit
```

```
## locfit 1.5-9.1    2013-03-22
```

```
## Loading required package: mgcv
```

```
## Loading required package: nlme
```

```
## This is mgcv 1.8-23. For overview type 'help("mgcv-package")'.
```

```
## Loading required package: tseries
```

```
##
## Attaching package: 'TSA'
```

```
## The following objects are masked from 'package:stats':
##
##      acf, arima
```

```
## The following object is masked from 'package:utils':
##
##      tar
```

Importing data

```
##           V1
## 1  1.3511844
## 2  0.7605324
## 3 -1.2685573
## 4 -1.4636872
## 5 -0.9792030
## 6  1.5085675
```

Introducing years to the data
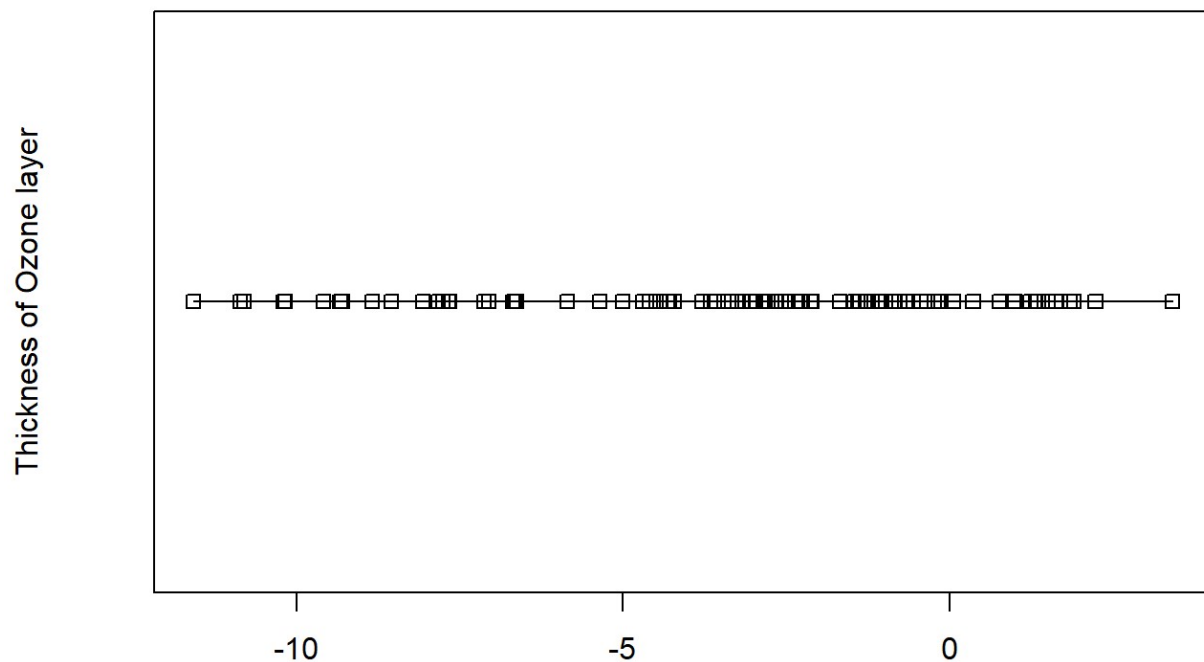
Checking the class of the data and plotting it

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

```
## Warning in plot.window(xlim, ylim, log, ...): graphical parameter "type" is
## obsolete
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): graphical
## parameter "type" is obsolete
```

```
## Warning in title(xlab = xlab, ylab = ylab, ...): graphical parameter "type"
## is obsolete
```

The above plot doesn't hold any significance. As though the years have been added but information is not conveyed to R as time.

Converting data to time series object

```
## [1]  1.3511844  0.7605324 -1.2685573 -1.4636872 -0.9792030  1.5085675
```
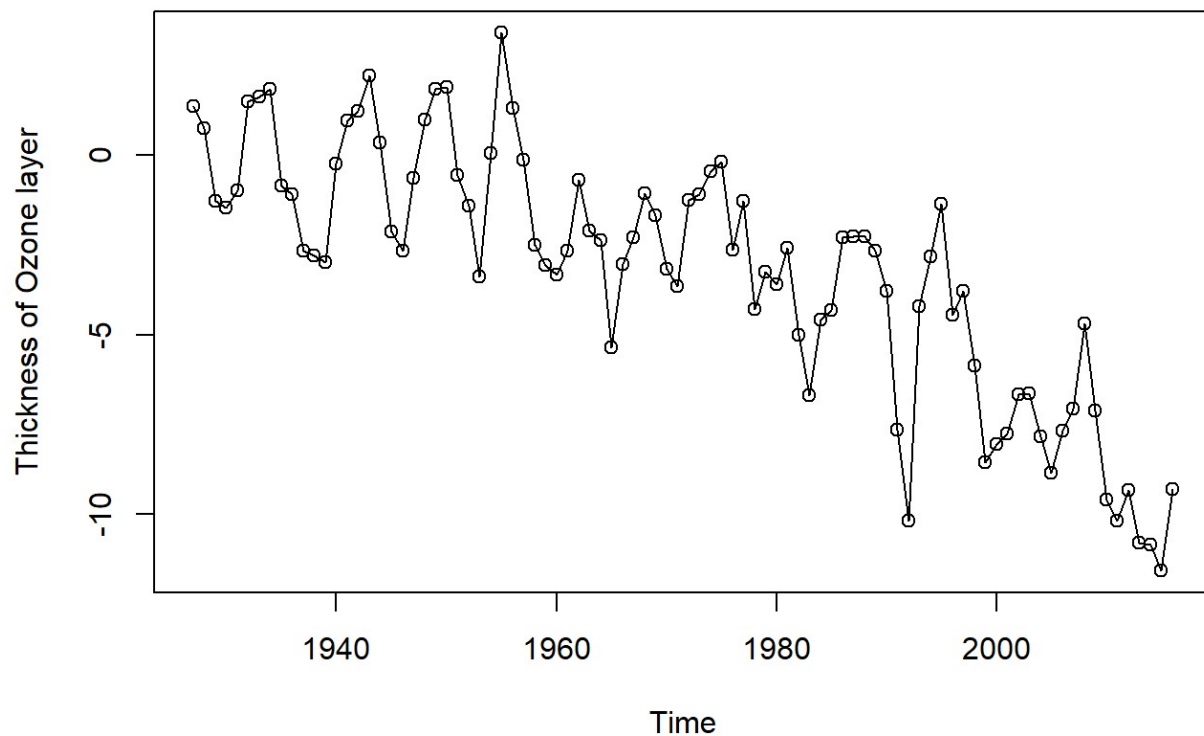
The positive values represent increase in the thickness of the ozone and negative values represent decrease in it.

ts() function converts the given observations through time having starting date and end date. Frequency is set to 1 as it is annual data.

# Descriptive Analysis

Plotting the time series

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

Interpretation:

As we can see the class is time series object and thereby the graph plotted is the time series plot.The discussion of the produced plot is based on 4 aspects.

1. Trend: There is clear display of downward trend, implying that width of the ozone layer has reduced over years.

2. Seasonality: There is no clear indication of seasonality. As there is no repeating pattern seen.

3. Changing variance:Not very clear view of changing variance.

4. Autoregressive behaviour or Moving Average: It seems that succeeding measurements are related to one another as the values that are neighbors in time tend to be similar in size. If so, we might be able to use one year's ozone width value to help forecast next year's width.

The following code chunk generates a scatter plot to investigate the relationship between pairs of consecutive width values:

Scatter plot of neighboring ozone width measurements
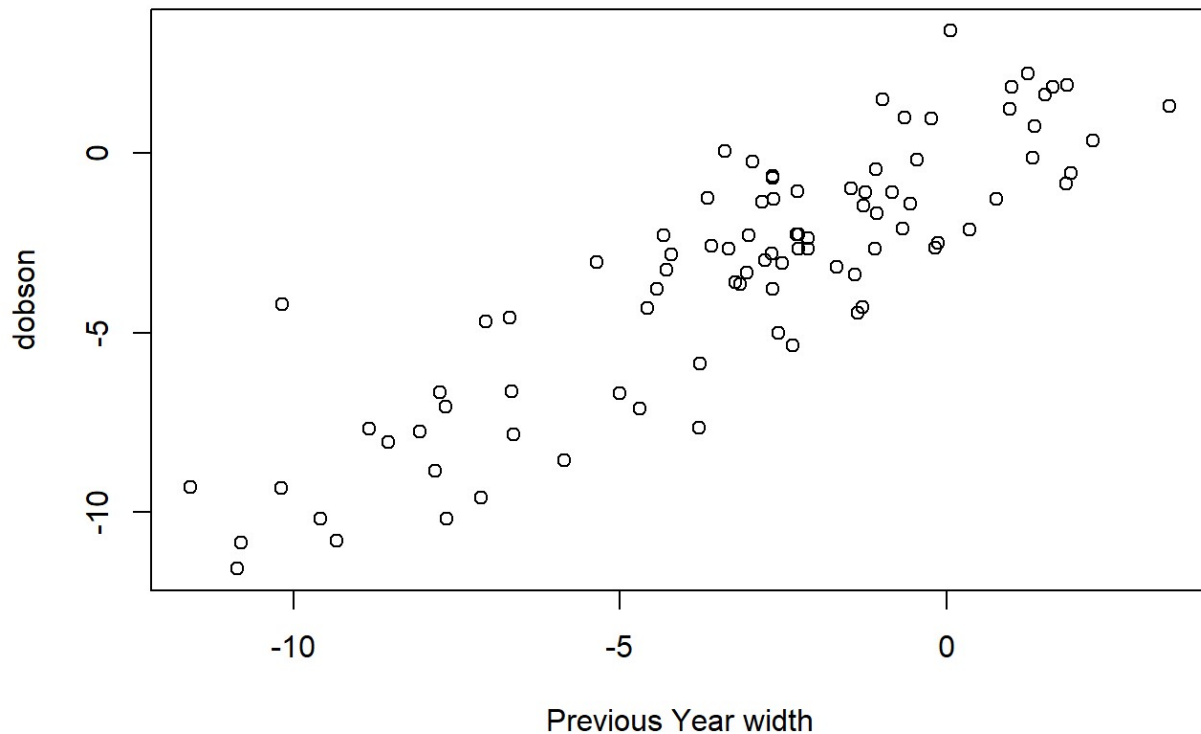
We observe a strong upward trend. The amount of correlation is quite high and can be calculated as follows.

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

As expected, we observe a high correlation between ozone width in one year with the succeeding year.

Modelling

Now to decide which models to apply. As the data displays the trend factor so let's begin by applying trend models

I) Linear trend model

```
## 
## Call:
## lm(formula = data1.ts ~ time(data1.ts))
## 
## 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(data1.ts)  -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
```

1. Both the coefficients (Intercept and linear component) are significant.
2. F-statistic value is significant implying the overall significance of the above linear trend model.
3. R-squared value is moderate implying the linear model is good but there is scope of improvement.

Let's fit the trend line to check whether the model is good fit or not.

Fitted linear trend line to ozone width data
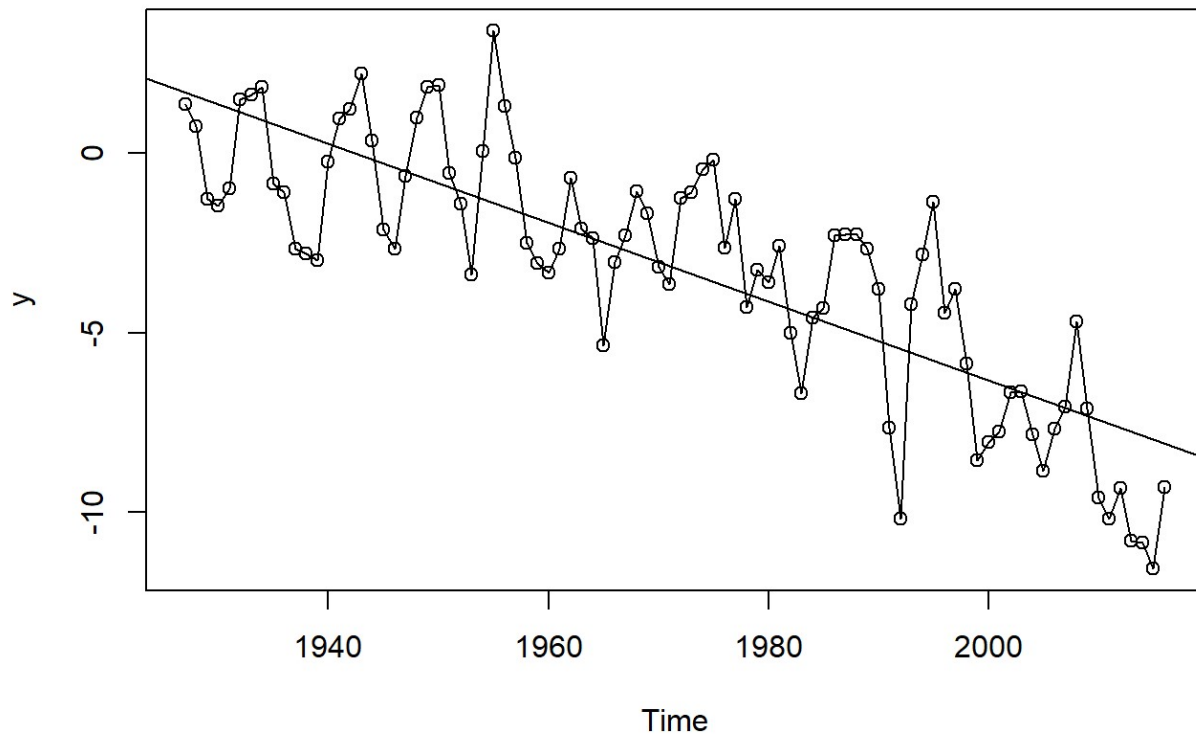
The trend line fits fine but is not the optimal one as the distance between the data points and the trend line is quite considerable at many time points.

So we can think of trying the quadratic model next.

II) Quadratic Model:

```
##
## Call:
## lm(formula = data1.ts ~ 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
```

1. All the coefficients are significant.
2. F-statistic value is also significant.
3. R-squared value has improved over linear model. So the quadratic model seems to be better fitting model.

Plotting the fitted quadratic curve along with the observed ozone width series.

Fitted quadratic curve to ozone width data
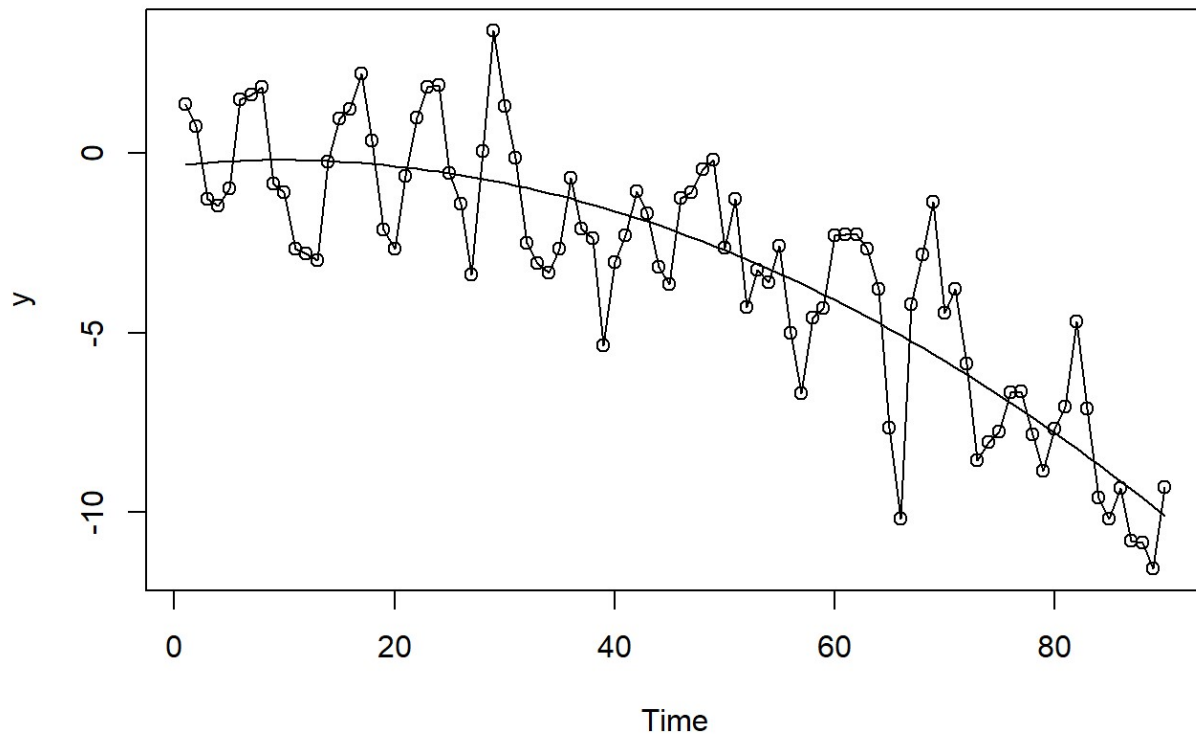
The quadratic line seems to fit better to the data in comparison to the linear trend line.

Just to check if there is some presence or effect of cyclical or seasonal trend, we convert the annual data with frequency 1 to the one with higher frequency (=4). And try apply the seasonal models.

Converting annual data to seasonal data with frequency equal to 4.

```
## [1]  1.3511844  0.7605324 -1.2685573 -1.4636872 -0.9792030  1.5085675
```
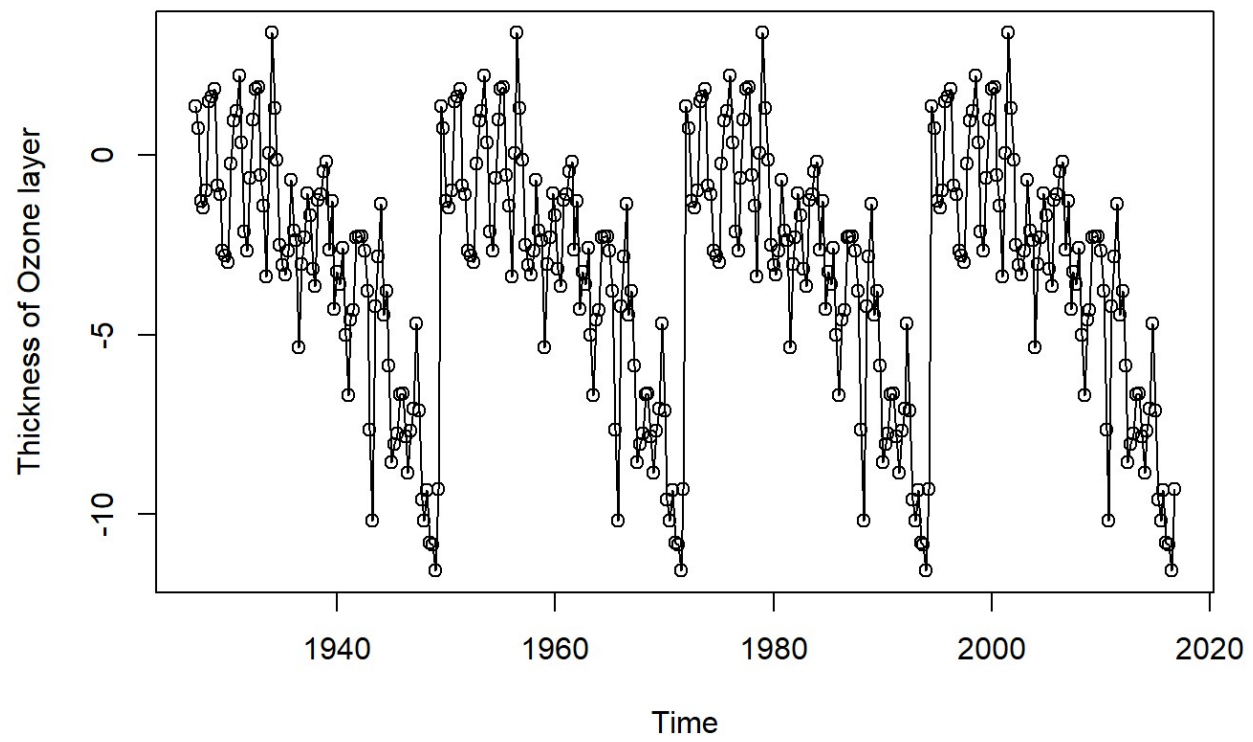
III) Seasonal model

```
##
## Call:
## lm(formula = data2.ts ~ quarter. - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3605 -1.8203  0.5494  2.3792  6.6261
##
## Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
## quarter.1Q  -3.2189     0.3704  -8.691  < 2e-16 ***
## quarter.2Q  -3.1856     0.3704  -8.601 2.55e-16 ***
## quarter.3Q  -3.2189     0.3704  -8.691  < 2e-16 ***
## quarter.4Q  -3.1856     0.3704  -8.601 2.55e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.514 on 356 degrees of freedom
## Multiple R-squared:  0.4565, Adjusted R-squared:  0.4504
## F-statistic: 74.76 on 4 and 356 DF,  p-value: < 2.2e-16
```

All quarterly coefficients are significant. So effects of all quarters are significant in that model.

To see how this seasonal model fits the data, we plot the model along with series and check the results

Fitted model to ozone width series
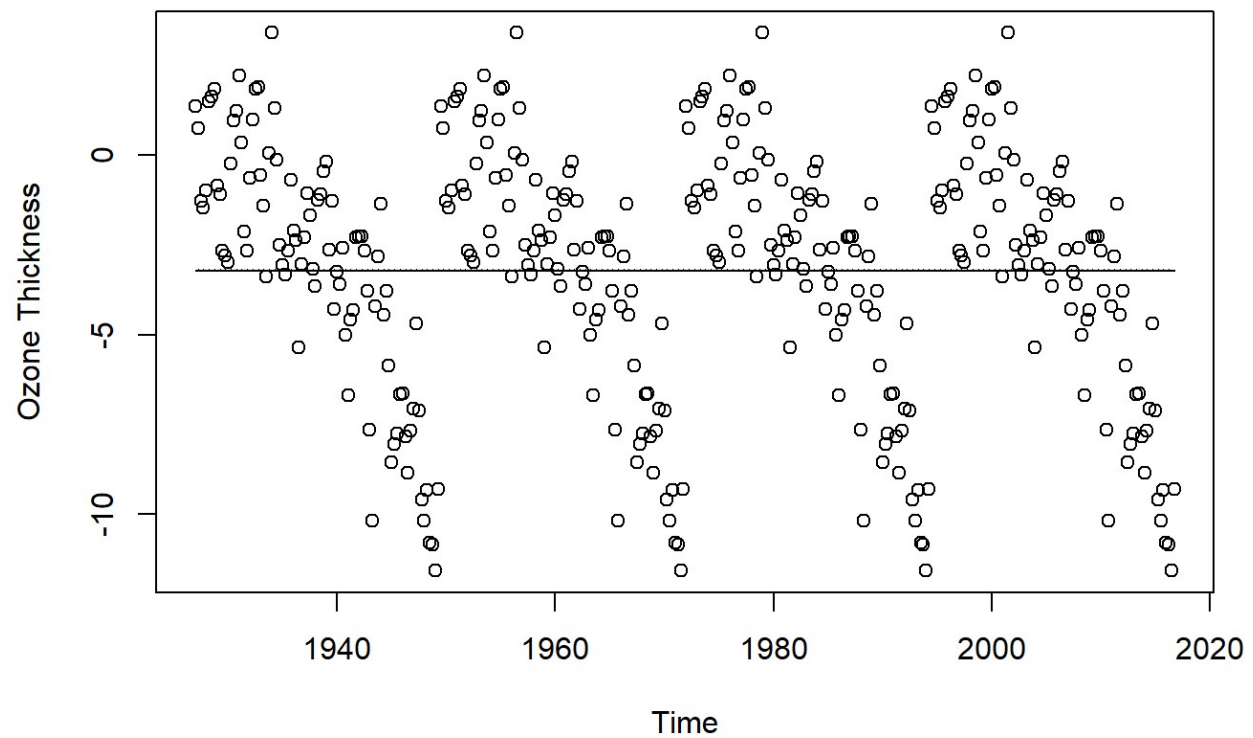
As clearly seen that model doesn't befit the data plot implying that the cyclic model is not an appropriate model for modelling ozone depletion series.

Let's try fitting the harmonic model to the data.

IV) Harmonic Model

```
## 
## Call:
## lm(formula = data2.ts ~ har.)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -8.3771 -1.8036  0.5384  2.3626  6.6094 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     -3.202e+00  1.849e-01  -17.32    <2e-16 ***
## har.cos(2*pi*t) -2.467e-13  2.615e-01    0.00         1    
## har.sin(2*pi*t)  1.318e-13  2.615e-01    0.00         1    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.509 on 357 degrees of freedom
## Multiple R-squared:  3.176e-27,  Adjusted R-squared:  -0.005602 
## F-statistic: 5.669e-25 on 2 and 357 DF,  p-value: 1
```

The harmonic coefficients are not significant, implying that harmonic model is not the suitable one for modelling and forecasting ozone layer width.

Inference from modelling results:

Considering the statistical results and plot outputs of the above models we choose linear trend model and quadratic model for the diagnostic testing. Asseasonal model and harmonic model do not come up as the suitable models for the given data series.

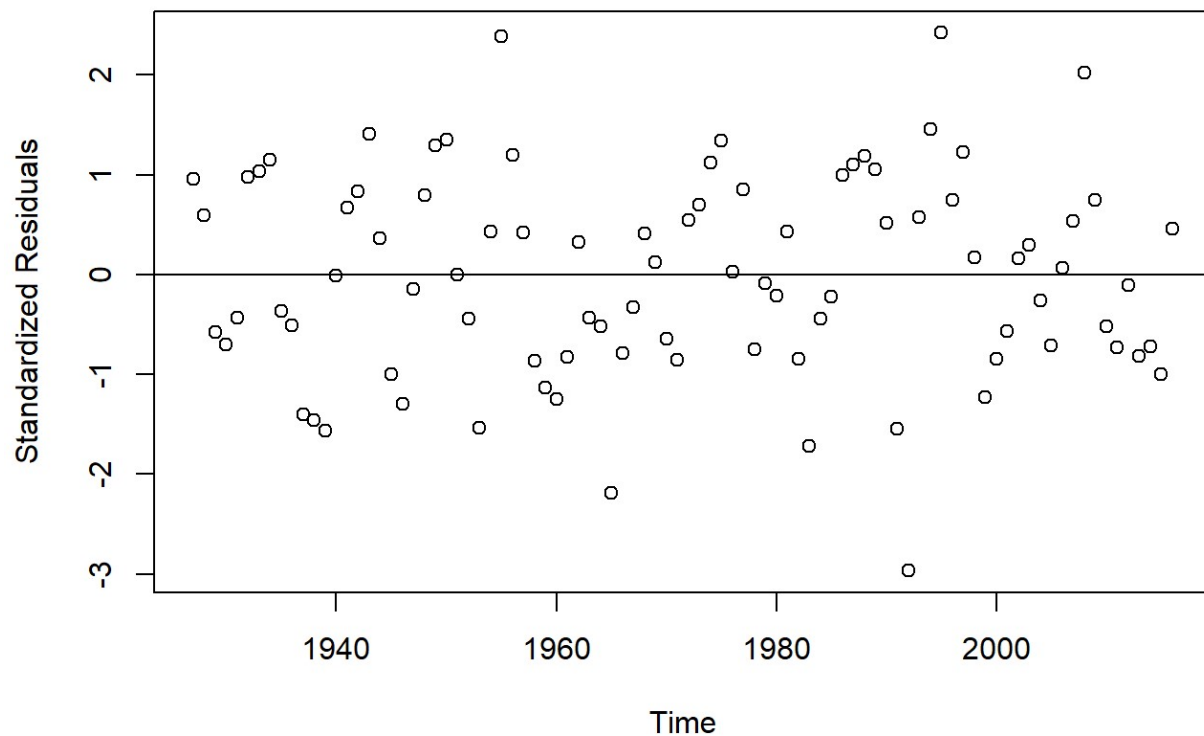# Diagnostic Testing

A) Residual Analysis:

If the trend model is reasonably correct, then the residuals should behave roughly like the true stochastic component (white noise). Let's check it for the models we applied above.

1. Residual Plot for linear model:
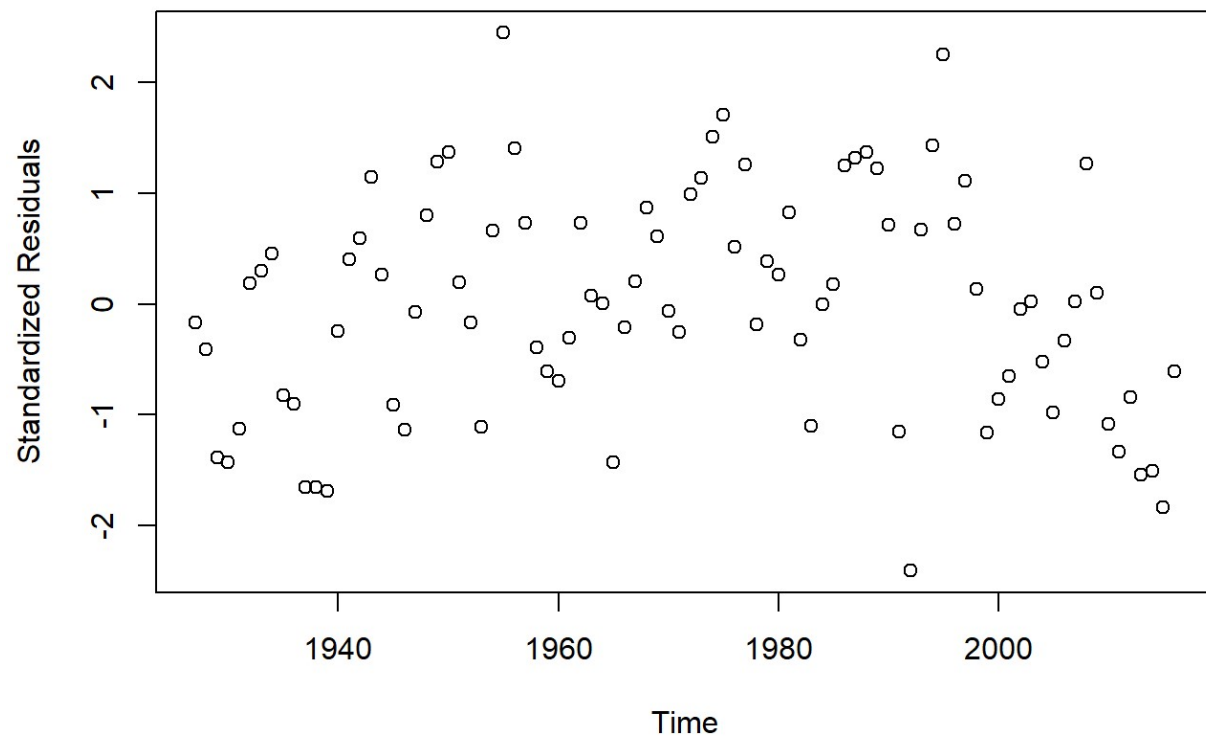
As seen the residual plot is random but it still shows hints of autoregressive behaviour so let's further check for quadratic model to see if it is better in terms of residuals and R^2 too.

2. Residual Plot for quadratic model

Still the residuals are not completely random. Autoregressive behaviour is very apparent for quadratic model.
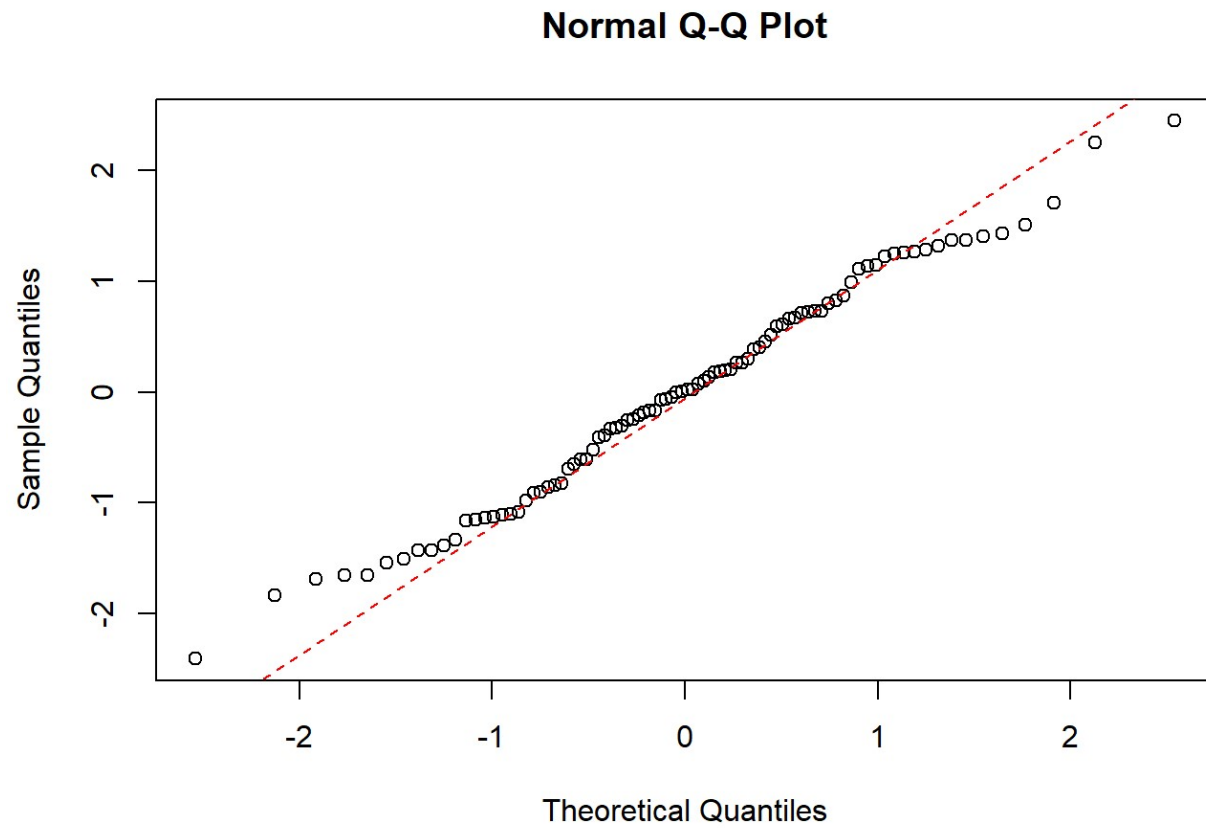
B) R-squared value

Another measure of diagnostic testing is R-squared value

1. For linear model: 0.6655
2. For quadratic model: 0.7331
3. For Seasonal model: 0.4504
4. For harmonic model: insignificant

C) QQ plot

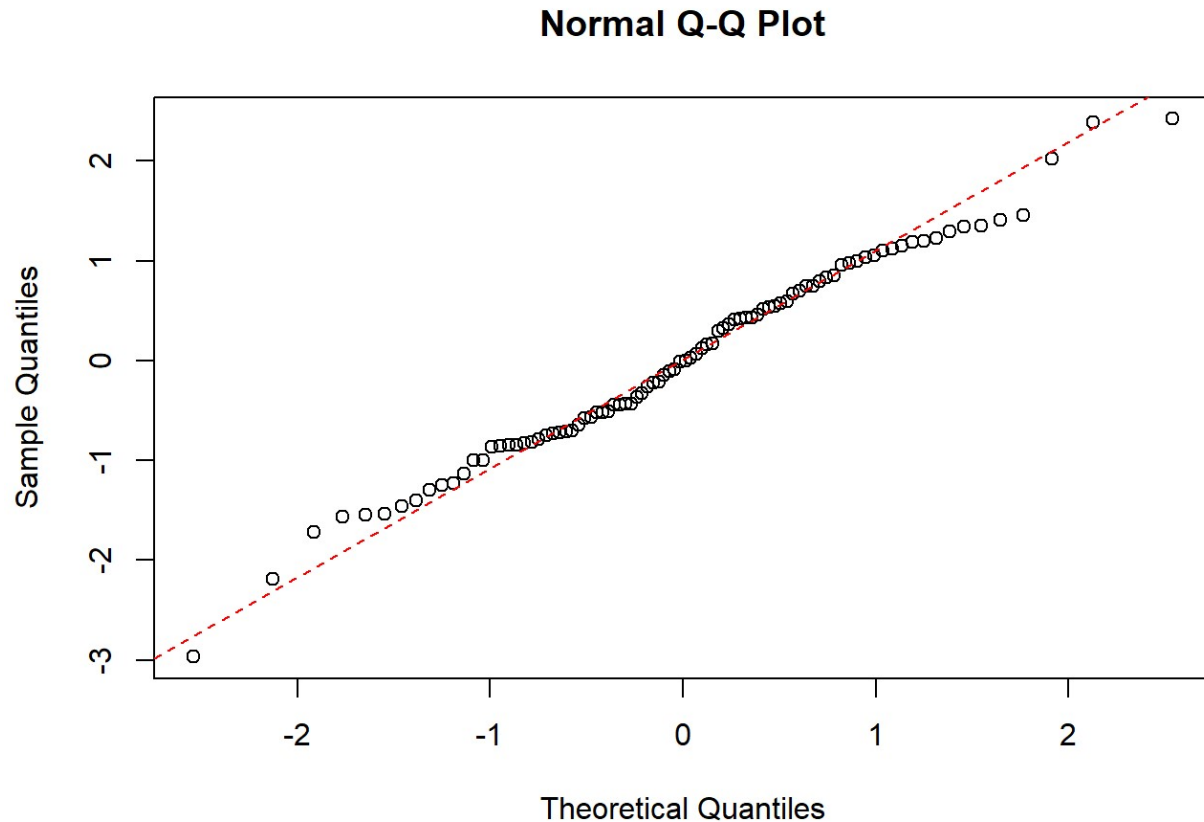Third measure of diagnostic testig is quantile-quantile (QQ) plot.

1. QQ plot for linear model

Normal Q-Q Plot

The QQ plot for the linear model shows much variation from the normal at the ends.

2. QQ plot for quadratic model

Normal Q-Q Plot

The straight-line pattern here supports the assumption of a normally distributed stochastic component in this model. Much better than the linear model QQ plot.

D) Shapiro Wilk Test

Fourth measure of diagnostic testing is Shapiro Wilk normality test.

1. Shapiro Wilk Test for linear model

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

As the p-value for the linear model is greater than alpha, hence the data is normally distributed.

2. Shapiro Wilk test for quadratic model

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

As the p-value for the quadratic model is greater than alpha , hence the data is normally distributed. The p value is even greater than the linear model case. Therefore, it's even better test result.

E) ACF plot

The fifth measure of diagnosis is the acf plot.

1. ACF plot for linear model

ACF of standardized residuals
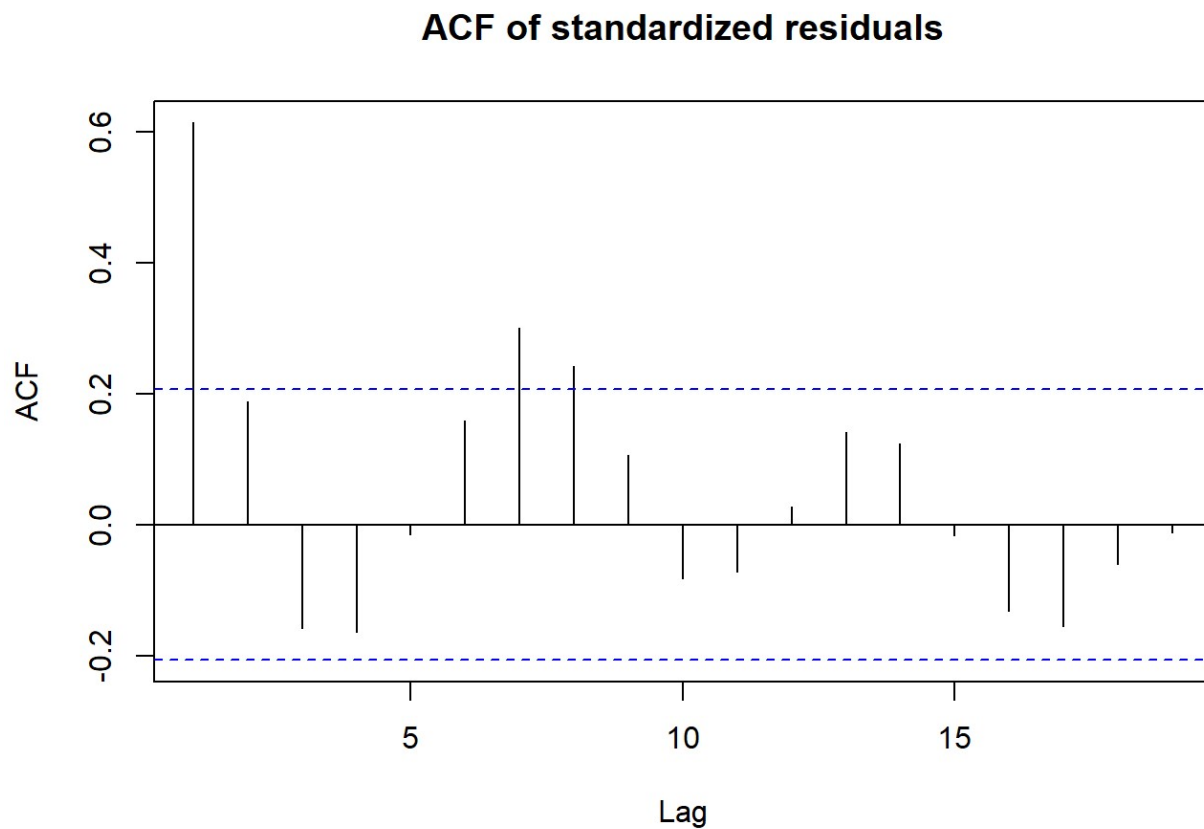
There is presence of significant auto correlation for some lags. This implies still significant auto correlation left in our residuals as could be seen by presence of close by successive points in the residual plot for the linear model

2. ACF plot for quadratic model

ACF of standardized residuals
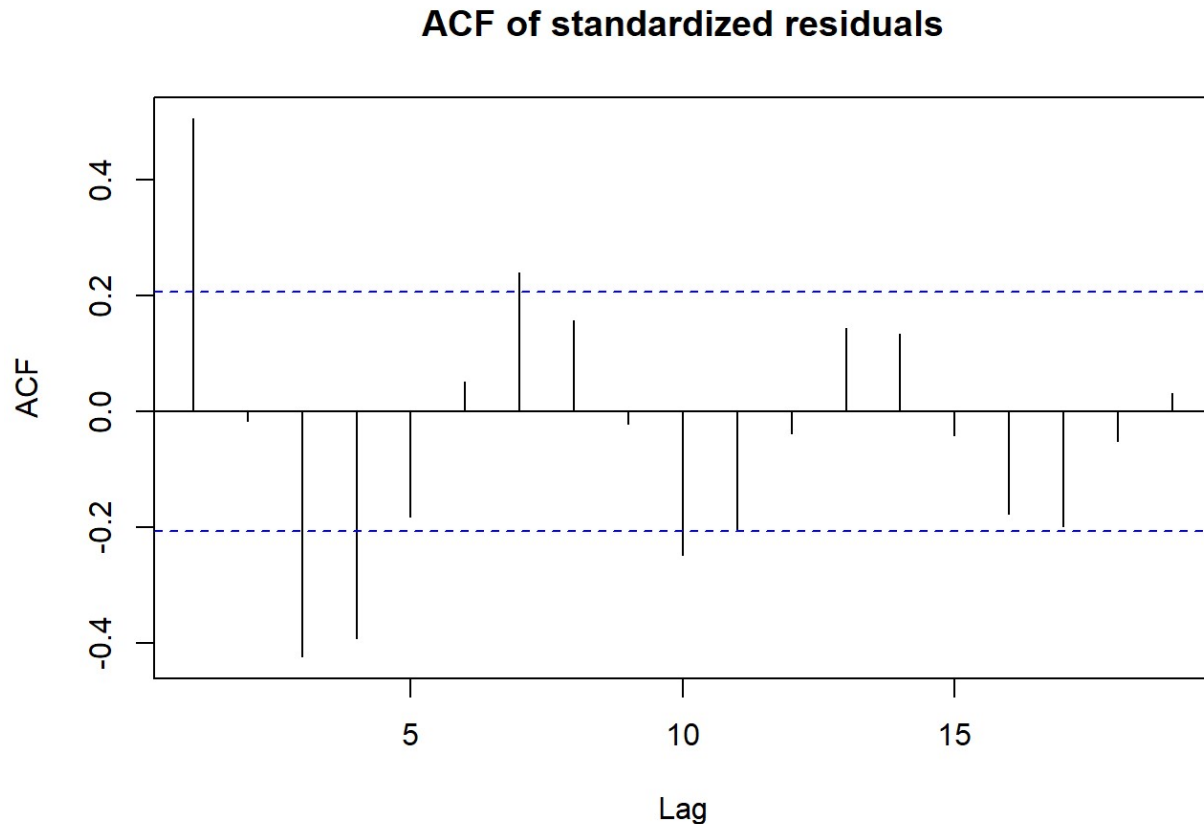
There is presence of significant auto correlation for some lags even for the quadratic model. Even quadratic model couldn't cope with it.

Interpretation from Diagnostic Testing:

1. Based on residual testing residual plot for linear trend model is more random as compared to quadratic model's residual plot.

2. Based on R-squared value quadratic model shows better results in comparison to the linear model's results. R-square is higher for quadratic model. Hence, better.

3. Considering QQ plot also the quadratic model shows better results for the condition of normally distributed stochastic component.

4. Following the Shapiro Wilk normality test, though both linear and quadratic models show p-value greater than alpha but value for quadratic model is higher. Thereby we can choose quadratic model over linear model in that regards.

5. Based on ACF plots output both linear trend model and quadratic model fail as both show the presence of significant autocorrelation. Hence, the autoregressive behaviour is held and is uncaptured by the trend model.

Conclusion for best fit model:

Based on the diagnosis quadratic model seems to be the best fit considering R value and normality condition. But we have to keep in mind that auto correlation is still not removed from the residuals and would hamper the predictions done through these models.

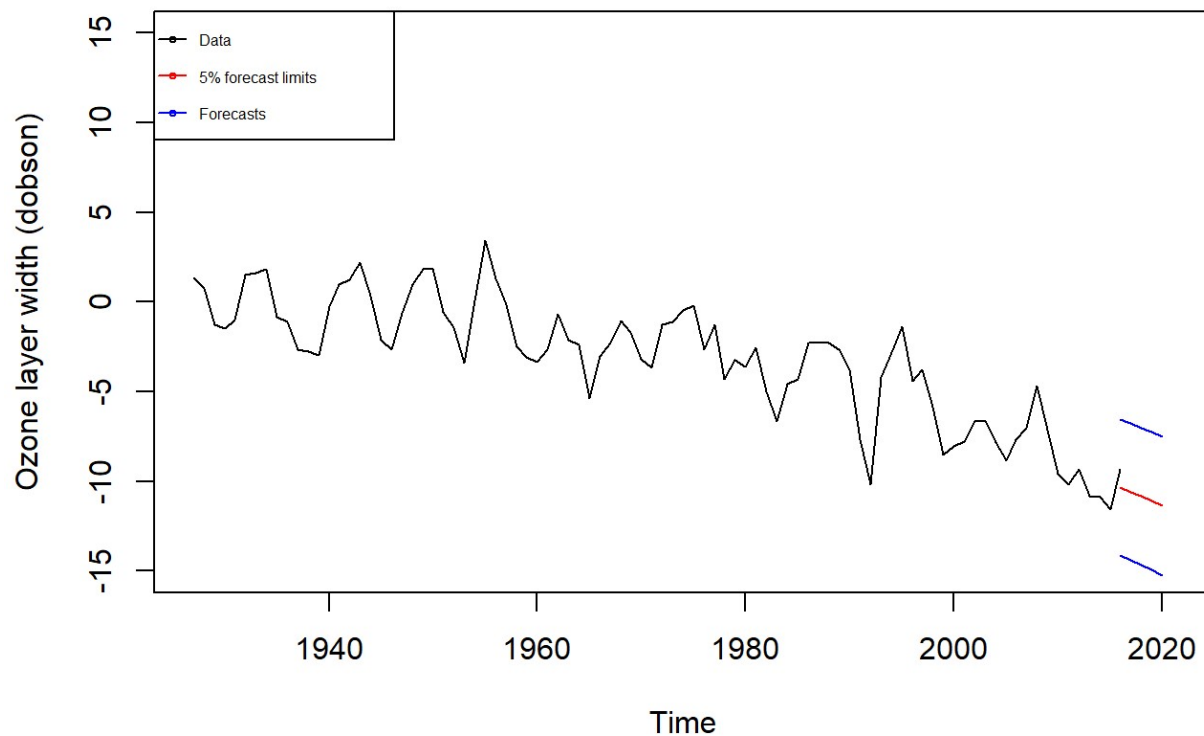# Forecasting: Predictions of ozone layer thickness for the next 5 years.

Using quadratic trend model

    a. Five step ahead forecasts for the ozone width

```
##         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
```

The predicted values through the quadratic model show decreasing trend. Implying that the width of the ozone layer is going to decrease in next five years.

The graphical representation is as below:

Summary:

In this task we dealt with the dataset representing yearly changes in the thickness of ozone layer during the year 1927 to 2016 in Dobson unit. The goal was to find the best fitting trend model to the dataset and give predictions of yearly changes for the next 5 years. 4 models were applied viz. linear trend model, quadratic trend model, seasonal model and the harmonic model and their statistical outputs and the plots were obtained. Based on which the suitable models were picked and taken forward for the diagnostic testing. As seen, out of all the four models considered, only two models namely linear trend model and quadratic trend model were found suitable for the further analysis and were thereby taken up for diagnostic testing. Our diagnostic phase involved 5 tests-

1. Residual analysis
2. R-squared value
3. QQ plot
4. Shapiro Wilk test
5. ACF plots

Though it was seen that none of the two models dealt with autocorrelation properly and the residual plots still hold the autocorrelated values. But based on other measures it was found that quadratic model is the most suitable model and hence was picked for the forecasting. The forecasted values seem to follow the downward trend.

Appendix:

Loading Packages

install.packages("TSA") library(TSA)

Importing data

data1 <- read.csv("C:/Users/user/Desktop/Namita Chhibba/Semester III/Time Series/Assignment 1/data1.csv", header= FALSE) head(data1)

Introducing years to the data

rownames(data1) <- seq(from=1927, to=2016)

Checking the class of the data and plotting it

class(data1) plot(data1, type='o', ylab='Thickness of Ozone layer')

Converting data to time series object

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

Descriptive Analysis

Plotting the time series

class(data1.ts) plot(data1.ts, type='o', ylab='Thickness of Ozone layer')

Scatter plot of neighboring ozone width measurements

plot(y=data1.ts,x=zlag(data1.ts),ylab='dobson', xlab='Previous Year width', main = "Scatter plot of neighboring ozone width measurements")

y = data1.ts # Read the ozone width data into y x = zlag(data1.ts) # Generate first lag of the series index = 2:length(x) # Create an index to get rid of the first NA value in x cor(y[index],x[index]) # Calculate correlation between numerical values in x and y

Modelling

Now to decide which models to apply. As the data displays the trend factor so let's begin by applying

trend models

    I. Linear trend model

model1 =lm(data1.ts~time(data1.ts)) summary(model1)

plot(data1.ts, type='o', ylab='y', main = "Fitted linear trend line to ozone width data") abline(model1)

    II. Quadratic Model:

t=time(data1.ts) t2= t^2 model2 =lm(data1.ts~t+t2) summary(model2)

Plotting the fitted quadratic curve along with the observed ozone width series.

plot(ts(fitted(model2)), ylim = c(min(c(fitted(model2), as.vector(data1.ts))), max(c(fitted (model2),as.vector(data1.ts)))),ylab='y' , main = "Fitted quadratic curve to ozone width data") lines (as.vector(data1.ts),type="o")

Converting annual data to seasonal data with frequency equal to 4.

data2.ts <- ts(as.vector(data1), start=c(1927,1), end= c(2016,4), frequency=4) head(data2.ts) quarter.=season(data2.ts) # season function creates the 4 variables. We apply the linear model to those 4 columns.

plot(data2.ts, type='o', ylab='Thickness of Ozone layer')

    III. Seasonal model

model3=lm(data2.ts~quarter.-1) # -1 removes the intercept term summary(model3)

plot(ts(fitted(model3),freq=4,start=c(1927,1)),ylab='Ozone Thickness',type='l', ylim=range(c(fitted (model3),data2.ts)),main="Fitted model to ozone width series") # ylim ensures that the y axis range fits the raw data and the fitted values points(data2.ts)

    IV. Harmonic Model

har.=harmonic(data2.ts,1) # calculate cos($2 \pi t$) and sin($2 \pi t$) model4=lm(data2.ts~har.) summary (model4)

Diagnostic Testing

    1. Residual Plot for linear model:

res.model1=rstudent(model1) plot(y=res.model1, x=as.vector(time(data1.ts)), xlab='Time', ylab='Standardized Residuals', type='p')

    2. Residual Plot for quadratic model

res.model2= rstudent(model2) plot(y=res.model2, x=as.vector(time(data1.ts)), xlab='Time', ylab='Standardized Residuals', type='p') abline(h=0)

1. QQ plot for linear model

y = rstudent(model1) qqnorm(y) qqline(y, col = 2, lwd = 1, lty = 2)

2. QQ plot for quadratic model

y = rstudent(model2) qqnorm(y) qqline(y, col = 2, lwd = 1, lty = 2)

1. Shapiro Wilk Test for linear model

y = rstudent(model1) shapiro.test(y)

2. Shapiro Wilk Test for quadratic model

y = rstudent(model2) shapiro.test(y)

1. ACF plot for linear model

acf(rstudent(model1), main = "ACF of standardized residuals")

2. ACF plot for quadratic model

acf(rstudent(model2), main = "ACF of standardized residuals")

Forecasting: Predictions of ozone layer thickness for the next 5 years.

Using quadratic trend model

a. Five step ahead forecasts for the ozone width

t=c(2017, 2018, 2019, 2020, 2021) t1=t t2=t^2 new = data.frame(t1 , t2) forecasts2 = predict(model2, new, interval = "prediction") print(forecasts2)

The graphical representation is as below:

plot(data1.ts, xlim = c(1927,2021), ylim = c(-15, 15), ylab = "Ozone layer width (dobson)") lines(ts (as.vector(forecasts2[,1]), start = 2016), col="red", type="l") lines(ts(as.vector(forecasts2[,2]), start = 2016), col="blue", type="l") lines(ts(as.vector(forecasts2[,3]), start = 2016), col="blue", type="l")

legend("topleft", lty=1, pch=1, col=c("black","red","blue","red"), text.width = 18, c("Data","5% forecast limits", "Forecasts"), cex=0.5, y.intersp=2)