Assignment 1 Time Series

Wesley Nderi

25/03/2018

Introduction

Contextual background

The environment can be argued to be one of the most important resources to mankind. Over the years, urbanisation, globalisation and other human activities have had considerable effects on Mother Earth. The ice caps are melting, temperatures are reaching unprecented extremes, animal species are becoming extinct, islands are sinking and the list goes on and on. There is numerous scientific evidence to the effect that the human civilisation has had a devastating impact on the environment.

Methodology

The first task in this assignment is to analyze yearly changes in the ozone layer recorded from 1927 to 2016. The thickness of the ozone layer for the purposes of this task is measured in Dobson units. The dataset contains 90 records representing a recording each year. The second task is to find the best fitting trend model and the third task is to give predictions of yearly changes for the next five years.

This dataset was provided by Mr.Haydar Demirhan.

Task 1: Analysing yearly changes in the ozone layer

```
#Loading the required packages
library(TSA)
```

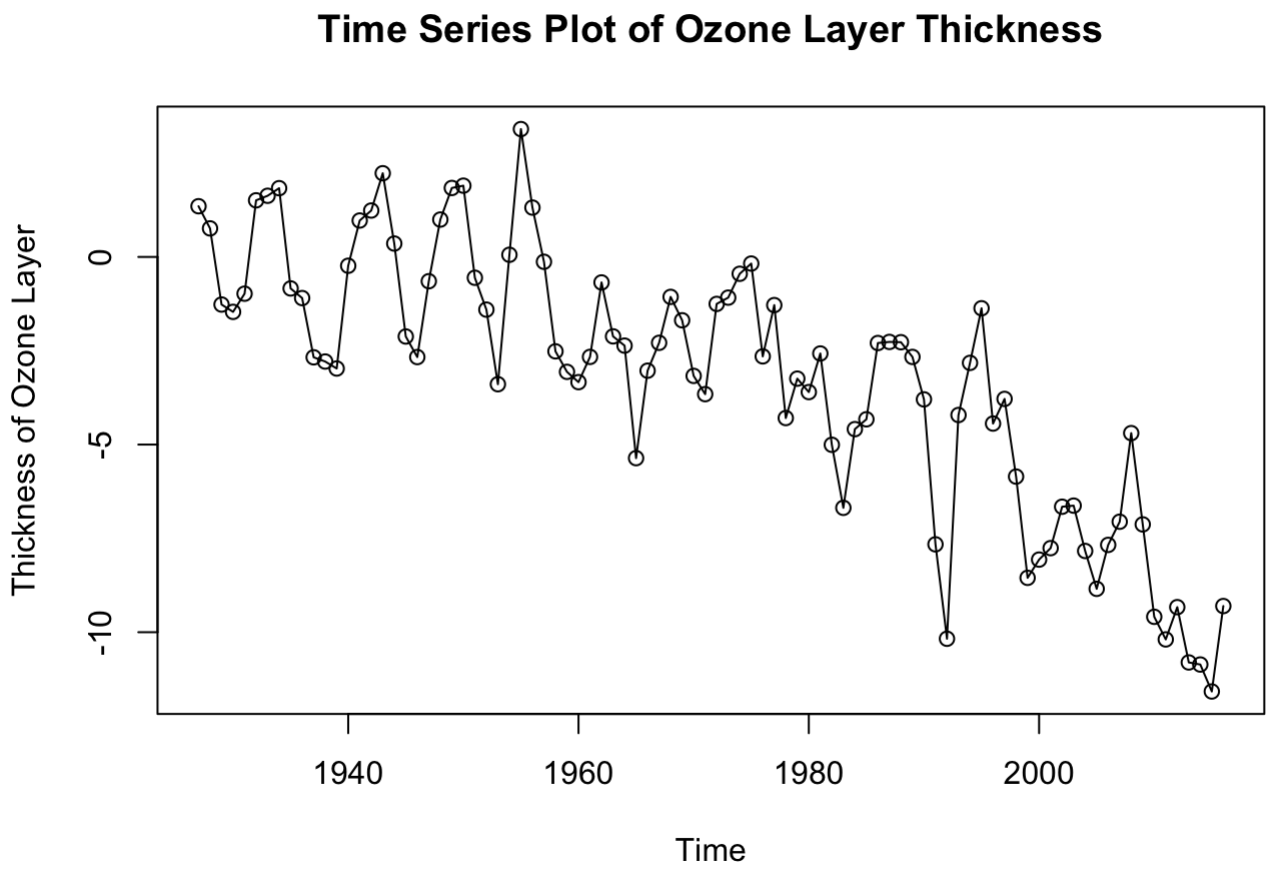
```
#Read the data into R
OzoneData<-read.csv("/Users/wes/Downloads/data1.csv", header = FALSE)
rownames(OzoneData) <- seq(from=1927, to=2016)
colnames(OzoneData) <- c("Thickness")

#The dataset still appears to be a dataframe.
class(OzoneData)
```

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

```
#Converting the dataframe into a ts object
OzoneData <- ts(as.vector(OzoneData), start=1927, end=2016)

#A visualisation of the data
plot(OzoneData,type='o',ylab='Thickness of Ozone Layer', main="Time Series Plot of Oz
one Layer Thickness")
```



What can we observe from the above plot in terms of the following?

a)Trend: There is a general decreasing trend.

b)Behaviour: It appears that the above series is auto-regressive as there are numerous succeeding points.

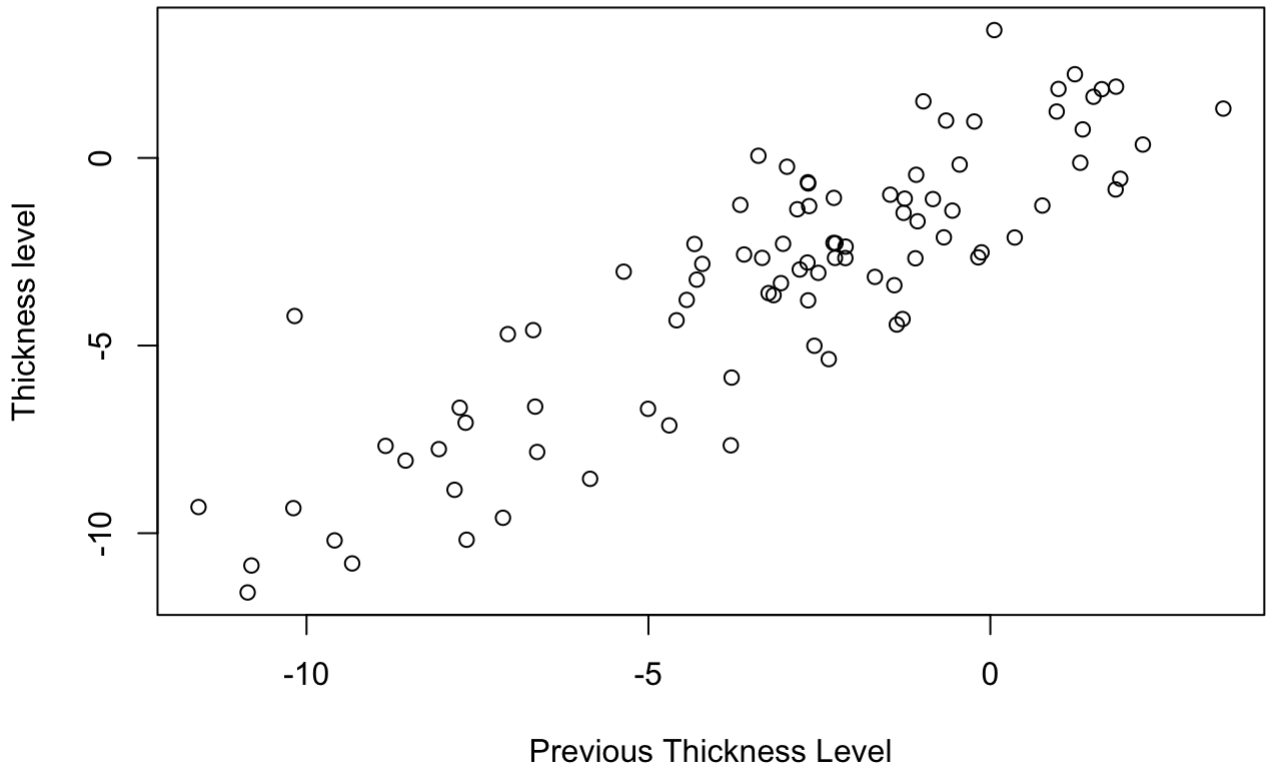
c)Seasonality: There is no obvious seasonality.

d)Changing variance: It is possible that there is a change in variance depicted by the movement from high values to low values.

We can investigate whether there is a relationship in the neighbouring measurements of ozone thickness.

```
#Scatter plot of Neighboring Thickness Levels
plot(y=OzoneData,x=zlag(OzoneData),ylab = 'Thickness level', xlab = 'Previous Thickne
ss Level', main="Scatter plot of Thickness Levels vs Previous Thickness levels in Dob
son units")
```

Scatter plot of Thickness Levels vs Previous Thickness levels in Dobson u



It appears that there is a correlation between the values of neighbouring years as evidenced by the upward trend. This means that low values tend to be followed by similarly low values, middle-sized values by middle sized and high values by high values. Further, we can also calculate the correlation value.

```
#Finding the value of the correlation
y = OzoneData
x = zlag(OzoneData)
index = 2:length(x)
cor(y[index],x[index])
```

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

The scatterplot above and correlation function indicate that there is a strong correlation between the thickness levels between succeeding years. The visualisation does not exactly give an indication of a changing variance.

Modelling Techniques

Task 2: Finding a suitable trend model

Using a process of elimination, we shall aim to distinguish which time series model fits this data best.

a) Linear Model

```
model1<-lm(OzoneData~time(OzoneData))
summary(model1)
```

```
##
## Call:
## lm(formula = OzoneData ~ time(OzoneData))
##
## 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(OzoneData)  -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
```

From the model summary above :

a. The coefficients are both significant. The slope and intercept are recorded as β^1= -0.11 and β^0 as 213.72.

b. The Std Error represents the standard deviations of the coefficients and assume a white noise stochastic component which is rarely true for time series.

c. The t-values are the estimated regression coefficients, each divided by their respective standard errors. If the stochastic component is normally distributed white noise, these ratios can be used to check the significance of the regression coefficients.

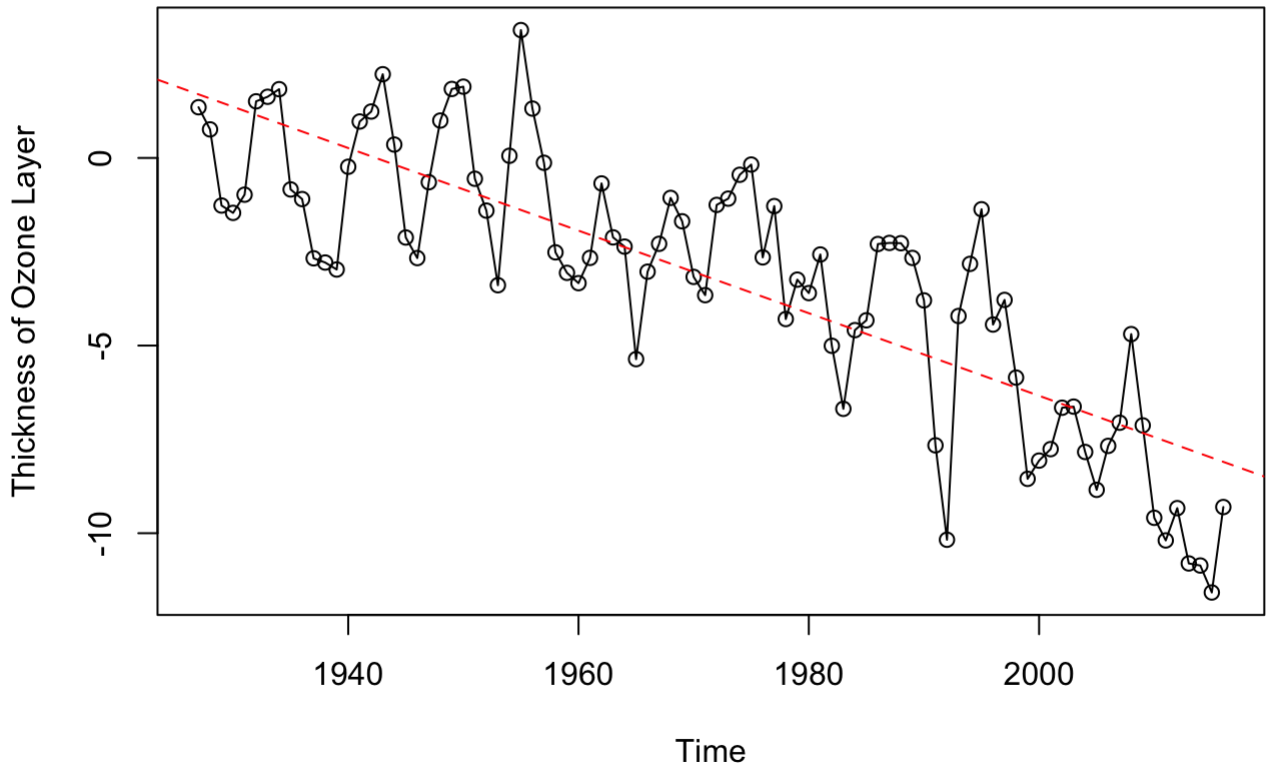
d. The F statistic is the overall significance of the model and in this case, is significant

e. R squared is defined as the square of the sample correlation coefficient between the observed series and the estimated trend. It is also the fraction of the variation in the series that is explained by the estimated trend. According to the summary, about 67% of the variation in the Ozone data time series is explained by this linear trend.The adjusted R Squared value is good enough but could possibly be improved.

The underlying assumption of this linear model therefore is that there must be a **normally distributed white noise stochastic component**. We can analyse the residuals which should behave like the stochastic component if the trend model is correct.

Fit of Linear Regression Model

```
#Fit of Linear Regression Model
plot(OzoneData,type='o',ylab='Thickness of Ozone Layer')
abline(model1,lty=2,col="red")
```

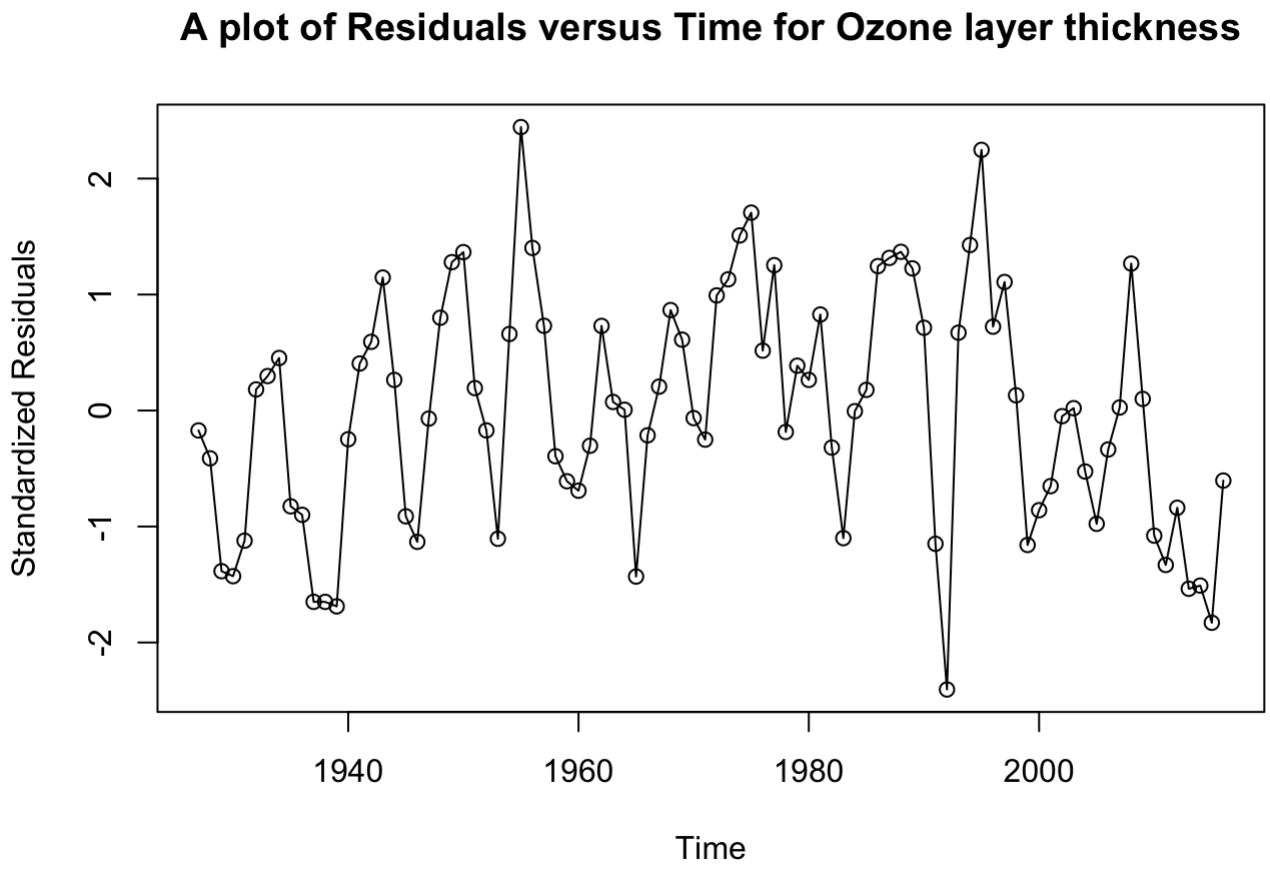


This is a significant fit for the series however we can see that multiple points are not captured by the linear regression. These points fall above and below the line.

Residuals of the Linear Model

As aforementioned, the residuals should behave like the stochastic component if the trend model is correct. Further, if the stochastic component is white noise, then the residuals should behave roughly like independent normally distributed random variables with zero mean and standard deviation s.

```
#Standardising the residuals
res.model1 = rstudent(model1)
plot(y = res.model1, x = as.vector(time(OzoneData)),xlab = 'Time', ylab='Standardized
 Residuals',type='o', main = "A plot of Residuals versus Time for Ozone layer thickne
ss")
```



From this visualisation of residuals:

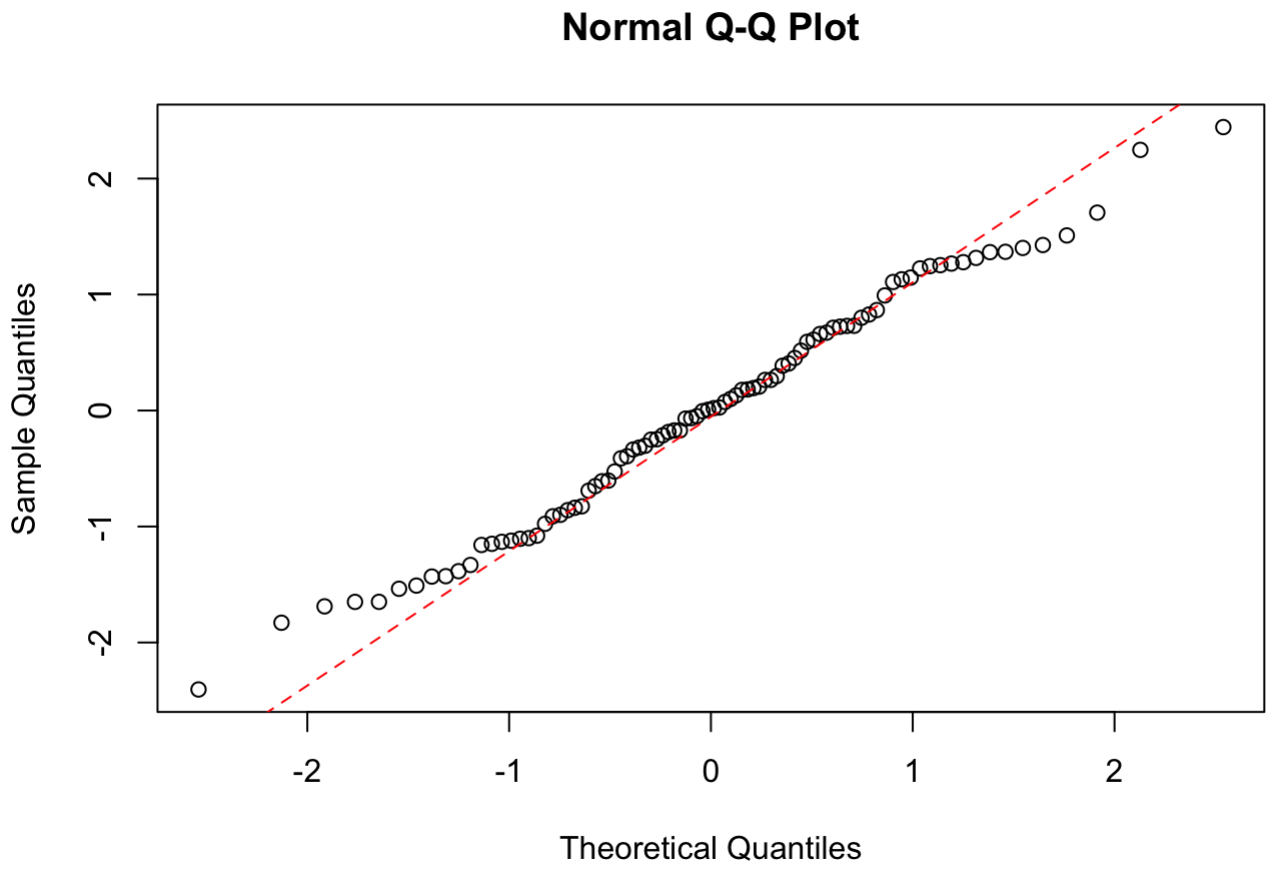
a. Residuals still do not capture a change in variance.

b. There does not seem to be a departure from randomness.

Checking for normality of the linear model residuals using a QQ plot

Non-normality can be assessed using a quantile-quantile (QQ) plot. With normally distributed values, a QQ plot looks approximately like a straight line.

```
qqnorm(res.model1)
qqline(res.model1, col = 2, lwd = 1, lty = 2)
```



As expected we can see a departure from normality in the end tails of the distribution.

Checking for normality of the linear model residuals using the Shapiro-Wilk Test

This test calculates the correlation between the residuals and the corresponding normal quantiles. The lower the correlation, the lower the evidence of normality. Similarly, the higher the correlation, the more evidence of normality.

```
#Shapiro-Wilk normality test
shapiro.test(res.model1)
```

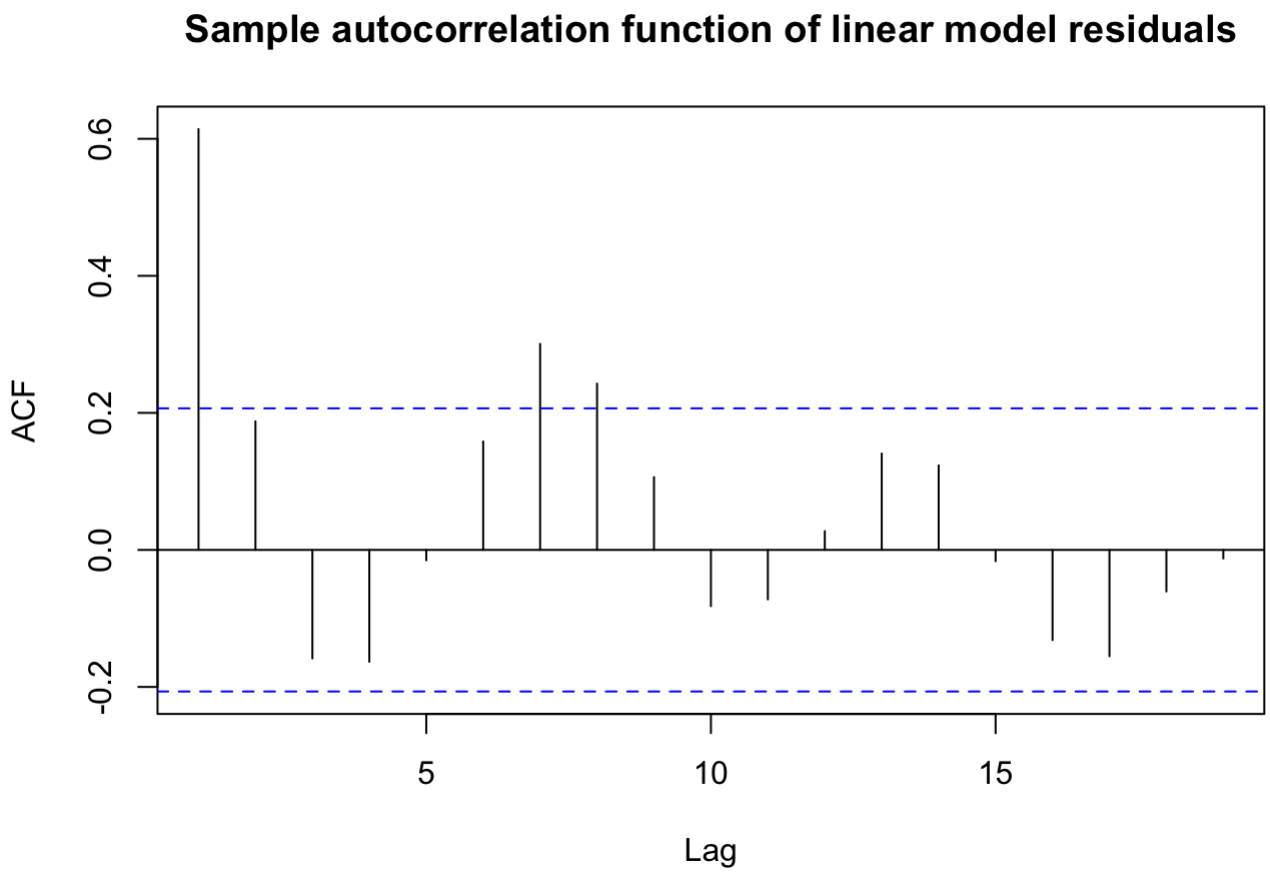
```
##
##  Shapiro-Wilk normality test
##
## data:  res.model1
## W = 0.98733, p-value = 0.5372
```

We cannot reject the null hypothesis that the stochastic component of this model is normally distributed.

Analysing possible dependence in the linear model residuals

We can use the sample autocorrelation function for the standardized residuals to infer whether this is a white noise process or not.

```
#ACF of Linear model
acf(res.model1, main="Sample autocorrelation function of linear model residuals")
```



Lag 1, 5 and 6 autocorrelations exceed two standard errors above zero. This is not what we expect from a white noise process. It is therefore reasonable to infer that the stochastic component of the series is **not white noise**. This means that there is significant dependence in the stochastic component which violates the underlying assumption of independence.

b) Quadratic model

```
t = time(OzoneData)
t2 = t^2
model2 = lm(OzoneData~t + t2)
summary(model2)
```

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

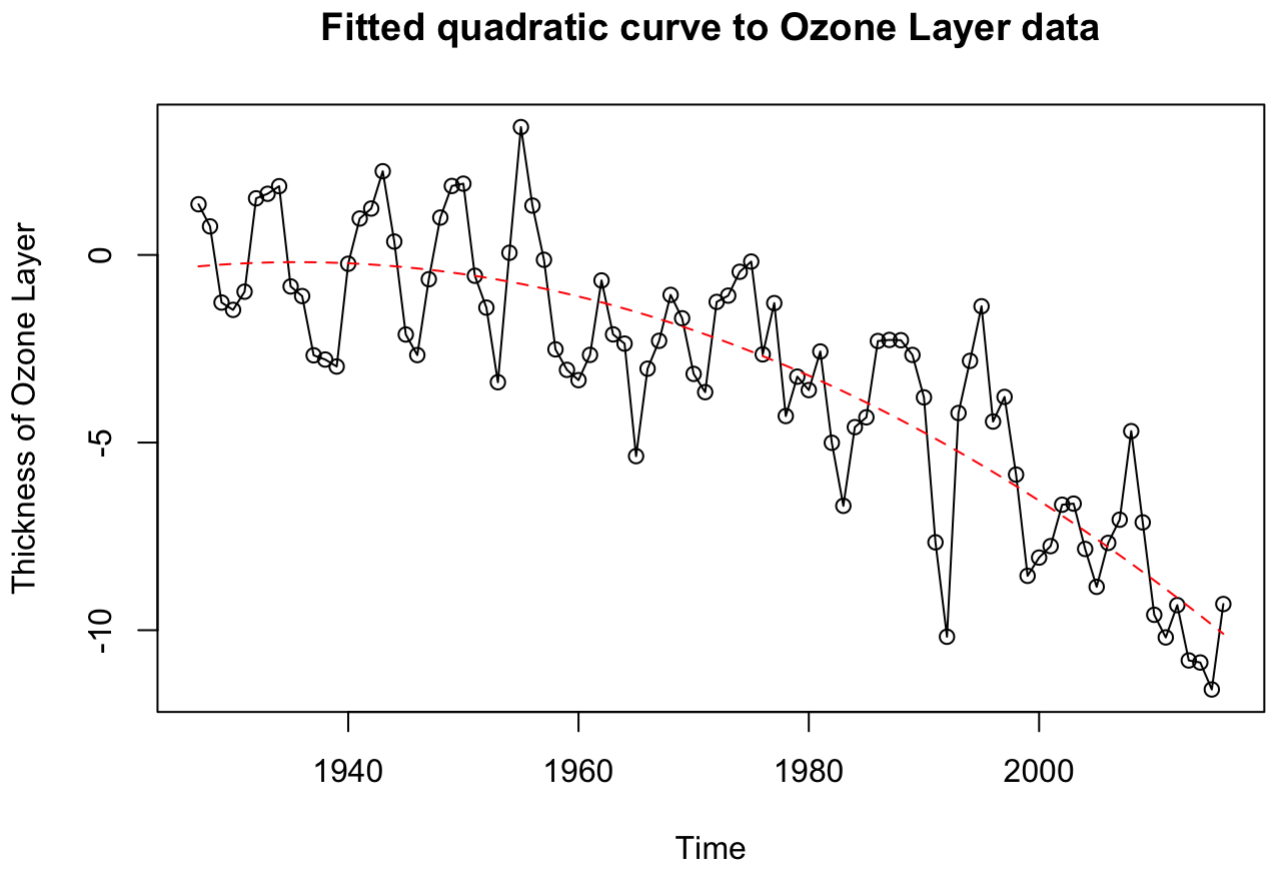
From the model summary above :

a. The quadratic component is significant.

b. According to the R squared summary above, about 73% of the variation in the Ozone data time series is explained by this quadratic trend.

c. The F statistic is significant.

```
#Fit of the Quadratic model
plot(OzoneData,type='o',ylab='Thickness of Ozone Layer', main="Fitted quadratic curve
 to Ozone Layer data")
points(t,predict.lm(model2), type="l", lty=2,col="red")
```



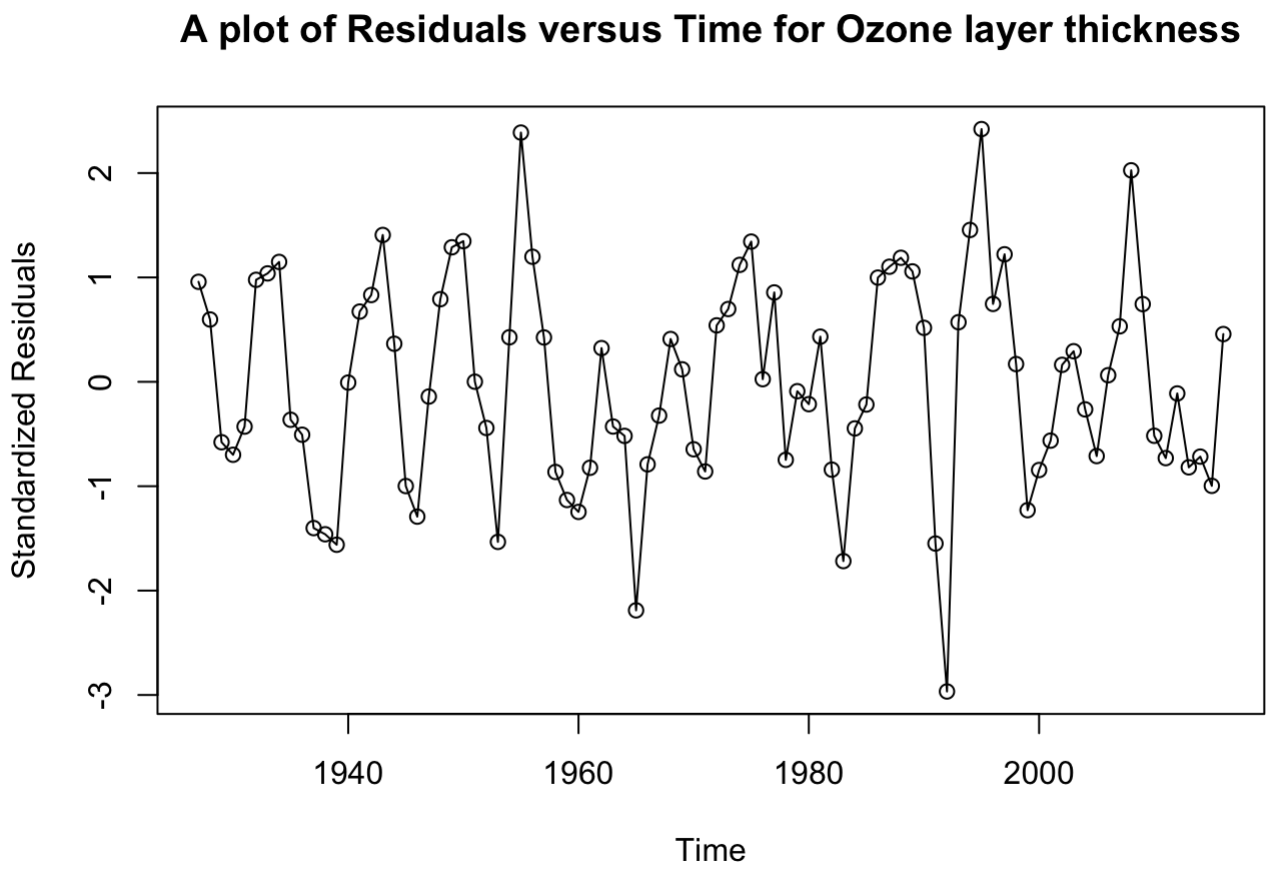
In comparison to the linear model summary and fit :

- The R squared value has improved from accounting for 66% of the variation to 73%.
- We observe that the quadratic model has a much better fit than the linear model which explains the improvement in the adjusted R squared figure.

Residuals of the Quadratic model

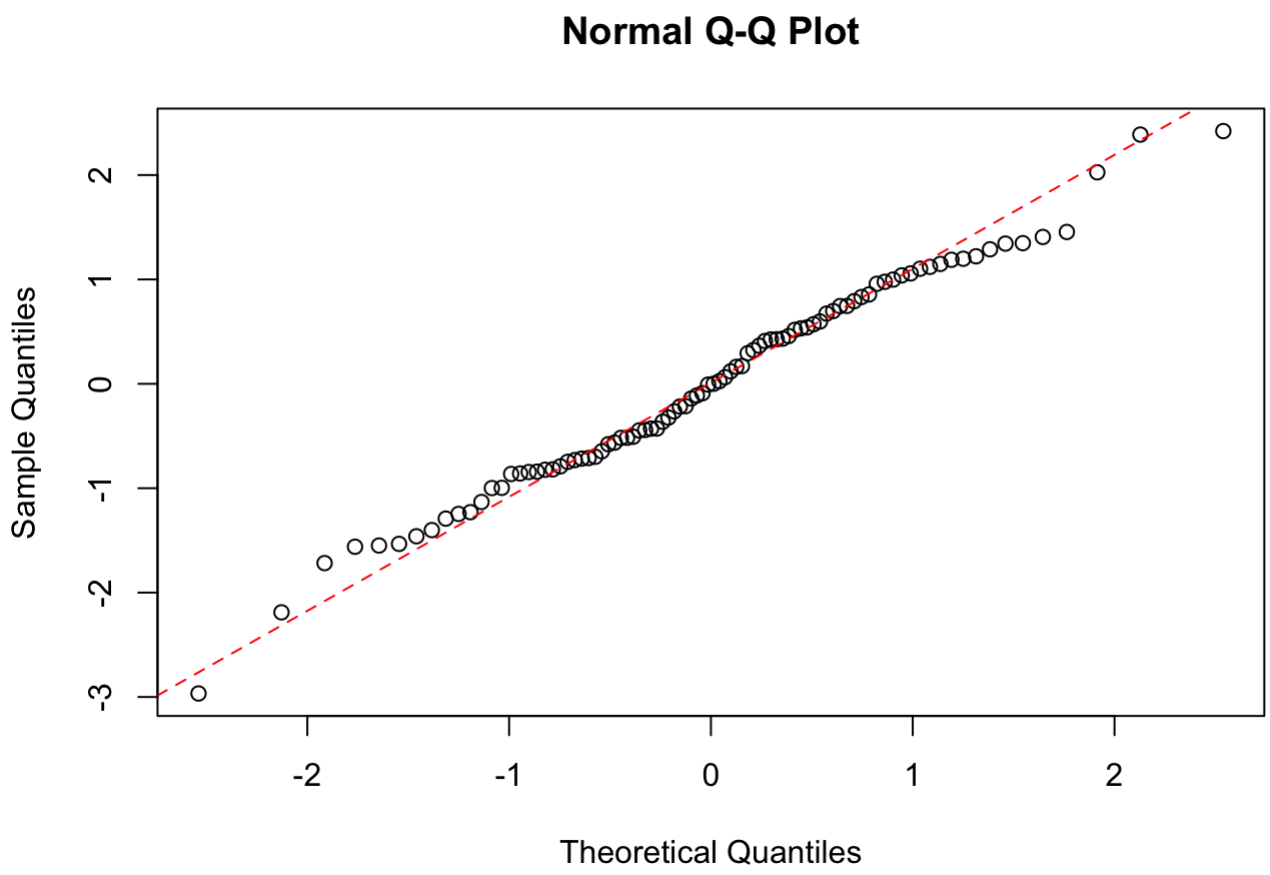
As mentioned when analysing the residuals of the linear model, we can ascertain that the trend model is correct if the stochastic component is white noise and the residuals behave like independent normally distributed random variables.

```
#Standardising Residuals
res.model2 = rstudent(model2)
plot(y = res.model2, x = as.vector(time(OzoneData)),xlab = 'Time', ylab='Standardized
 Residuals', main="A plot of Residuals versus Time for Ozone layer thickness",type=
'o')
```



Checking for normality of the quadratic model residuals using a QQ Plot

```
qqnorm(res.model2)
qqline(res.model2, col = 2, lwd = 1, lty = 2)
```



In comparison to the linear model, there is less departure from normality.

Checking for normality of the quadratic model residuals using the Shapiro-Wilk test

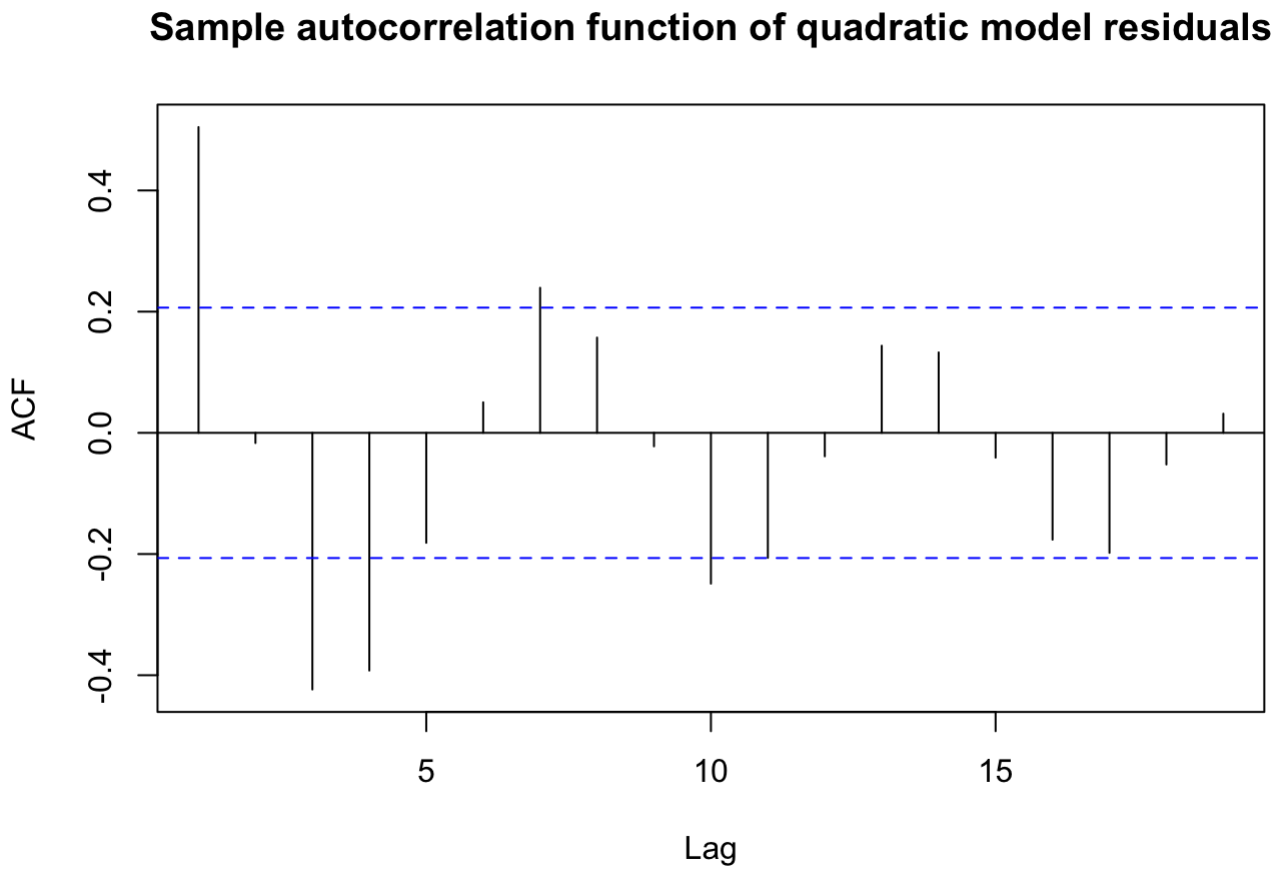
```
#Shapiro-wilk normality test
shapiro.test(res.model2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.model2
## W = 0.98889, p-value = 0.6493
```

Therefore, we cannot reject the null hypothesis that the stochastic component of this model is normally distributed. In addition, it is important to highlight that the score in the quadratic model(0.98889) is **higher** in the score in the linear model(0.98733). It is for this reason that we see less departure from normality in the QQ plot of the quadratic model.

Analysing possible dependence in the quadratic model residuals

```
#Auto Correlation Function of Quadratic model
acf(res.model2, main="Sample autocorrelation function of quadratic model residuals")
```

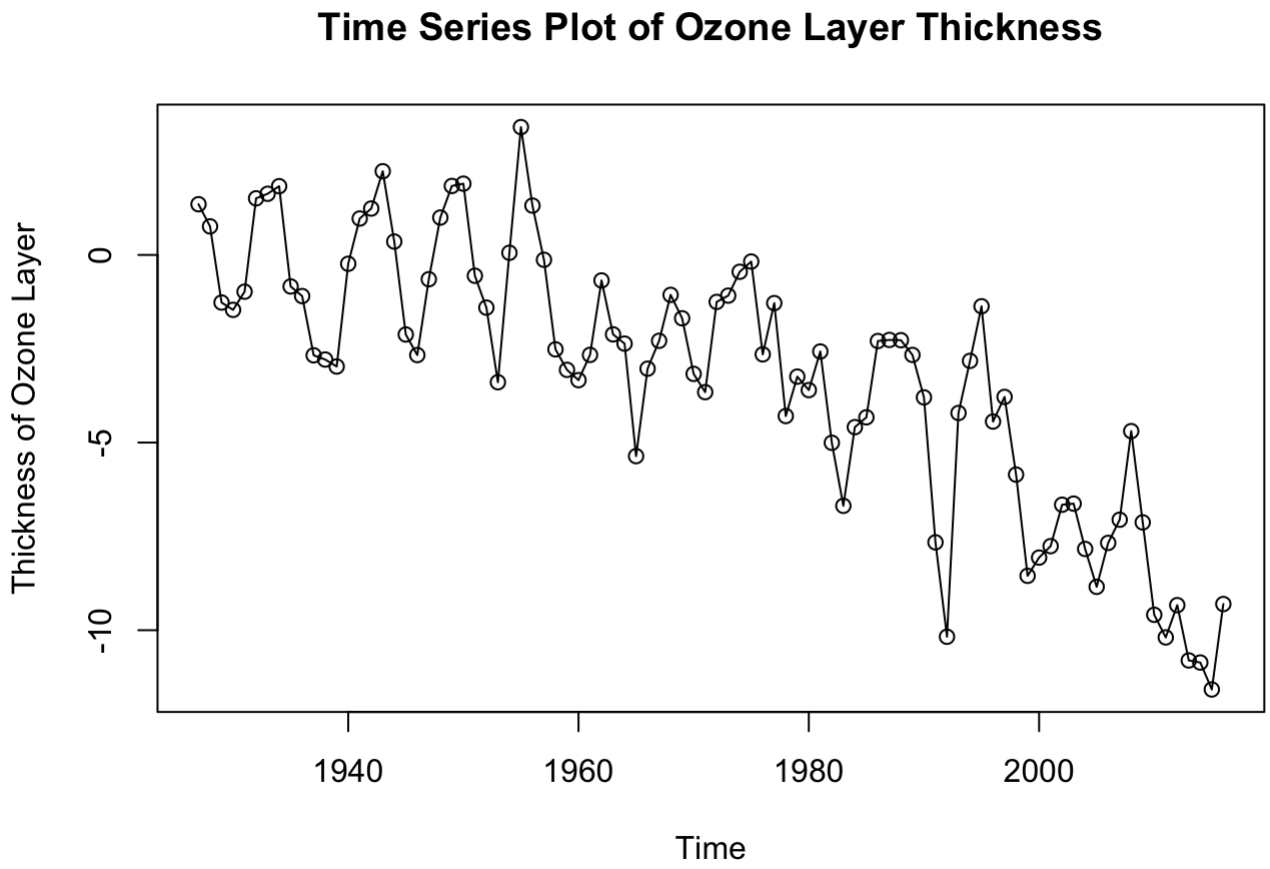


As can be seen above, there are several autocorrelations above and below two standard errors. It is therefore reasonable to infer that the stochastic component of the series is **not white noise**. This means that there is significant dependence in the stochastic component which violates the underlying assumption of independence.

c) Harmonic model

A cosine trend could not be fitted to this data because there is no evidence of seasonality observed in a time series plot of the Ozone layer data. The visualisation has been shown below.

```
plot(OzoneData,type='o',ylab='Thickness of Ozone Layer', main="Time Series Plot of Oz
one Layer Thickness")
```



Summary

The quadratic model seems to be the better model for the data, with a higher R squared value, higher Shapiro-Wilk values and better fit for the data. However, it does bear some inadequacies e.g. there is significant autocorrelation in its residuals.

Task 3: Prediction of yearly changes for the next five years

Using the quadratic model discussed earlier, we can estimate the ozone layer thickness values for the next 5 years based on the historical data available.

```
t10 = time(OzoneData)
t13 = t10^2

model5 <- lm(OzoneData~t10+t13)

#Reading in a vector for the next 5 years
t10 = c(2017,2018,2019,2020,2021)
t13 = t10^2

new = data.frame(t10, t13)

forecasts = predict(model5, new, interval="prediction")
print(forecasts)
```

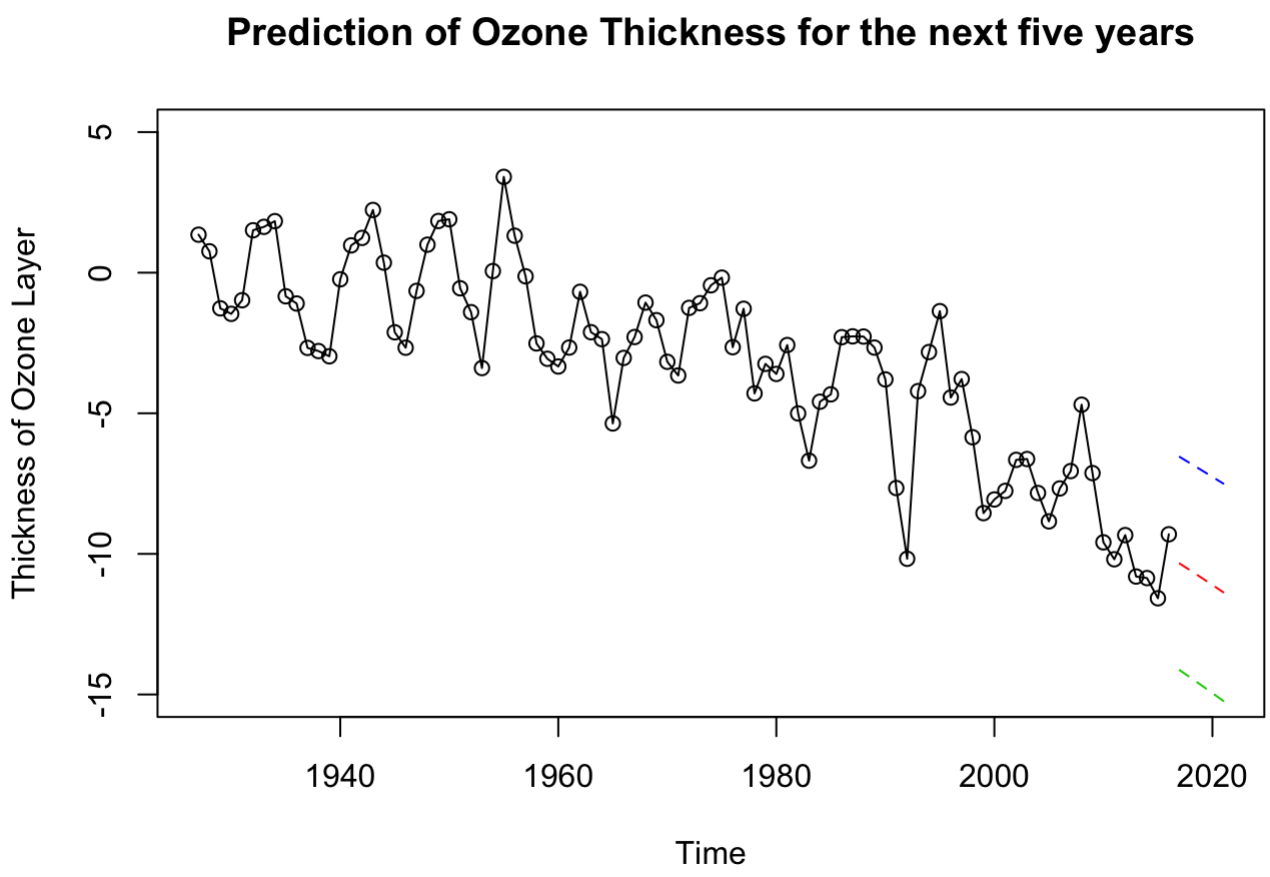
```
##         fit       lwr       upr
## 1 -10.34387 -14.13556 -6.552180
## 2 -10.59469 -14.40282 -6.786548
## 3 -10.84856 -14.67434 -7.022786
## 4 -11.10550 -14.95015 -7.260851
## 5 -11.36550 -15.23030 -7.500701
```

These are the predicition values for the next five years alongside a lower limit and upper limit.

A visualisation of these forecasts alongside the original plot

```
plot(OzoneData, xlim = c(1927,2021), ylim = c(-15,5),type = 'o', ylab = "Thickness of
 Ozone Layer", main = "Prediction of Ozone Thickness for the next five years")

lines(ts(as.vector(forecasts[,1]), start = 2017), col = 2, lwd = 1, lty = 2)#predicte
d value in red
lines(ts(as.vector(forecasts[,2]), start = 2017), col = 3, lwd = 1, lty = 2)#lower li
mit in green
lines(ts(as.vector(forecasts[,3]), start = 2017), col = 4, lwd = 1, lty = 2)#upper li
mit in blue
```



Conclusion

Although the quadratic model is better suited for this data, it still bears some inadequacy. There is still a significance in its residuals as illustrated by the sample auto correlation function that should not be there. The ideal for a sample ACF of residuals is that there is no significant correlation for any lag.

The possible explanation for this is that this is because this series is non-stationary. From its visualisation, there is a clear decereasing trend. This means that its probability laws that govern the behaviour of the process change over time.

To be able to find the best suited model, we must first convert this non stationary process into a stationary process through a process such as differencing.