Code ▾

Assignment 1-Time Series (MATH 1318)

Vamika Pardeshi-s3701024

March 31, 2019

Setup-

- Installed and loaded the necessary package to reproduce the report.

Hide

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

Introduction-

One of the global concerns, today, is the deterioration of Ozone layer. It results in increased amount of sun's ultraviolet light reaching Earth's surface, that is potential of damaging life on our planet. Over time, the enhanced human usage and production of ozone depleting chemicals have had noticeable impact on Earth. The given dataset deals with the same issue, i.e., the yearly changes in the thickness of Ozone layer from 1927 to 2016 in Dobson units. Dobson units measures what would be the physical thickness of Ozone layer if compressed in Earth's atmosphere. This report covers the analysis of the data, provides the best fitting trend model for the dataset and in addition to this, it also gives the predictions of yearly changes for the next 5 years.

Data Import-

Imported the dataset into R using read.csv() function.

Hide

```
Ozone_thickness <-read.csv("data1.csv", header = FALSE)
rownames(Ozone_thickness) <- seq(from=1927, to=2016)
colnames(Ozone_thickness) <- c("Thickness in DU")
class(Ozone_thickness)
```

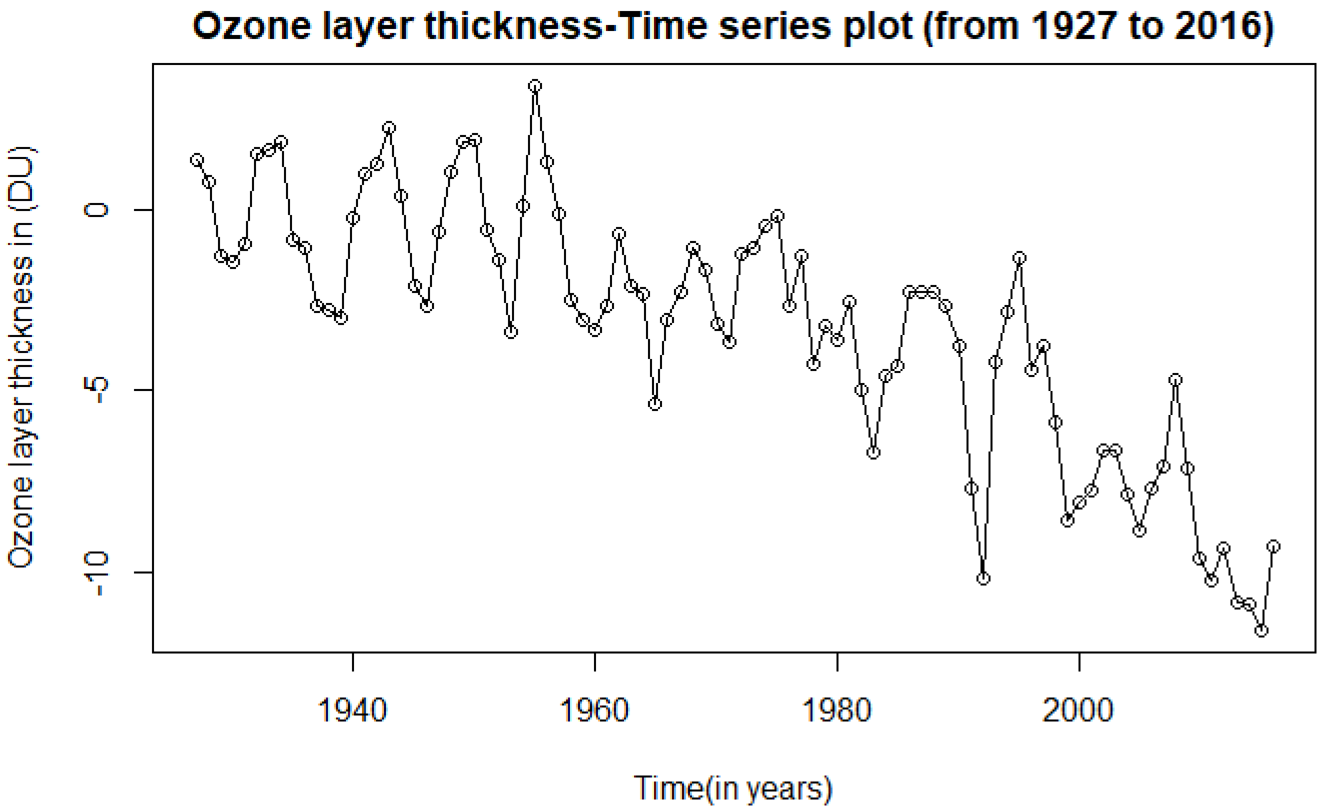
```
[1] "data.frame"
```

TASK 1- Analyzing the data

Plotted the time series graph in order to visualise the dataset.

Hide

```
#Since the class of the dataset is dataframe, so first conveting it into a ts object
Ozone_thickness <- ts(as.vector(Ozone_thickness), start=1927, end=2016)
plot(Ozone_thickness,type='o',ylab='Ozone layer thickness in (DU)',xlab='Time(in years)',main =
"Ozone layer thickness-Time series plot (from 1927 to 2016)")
```



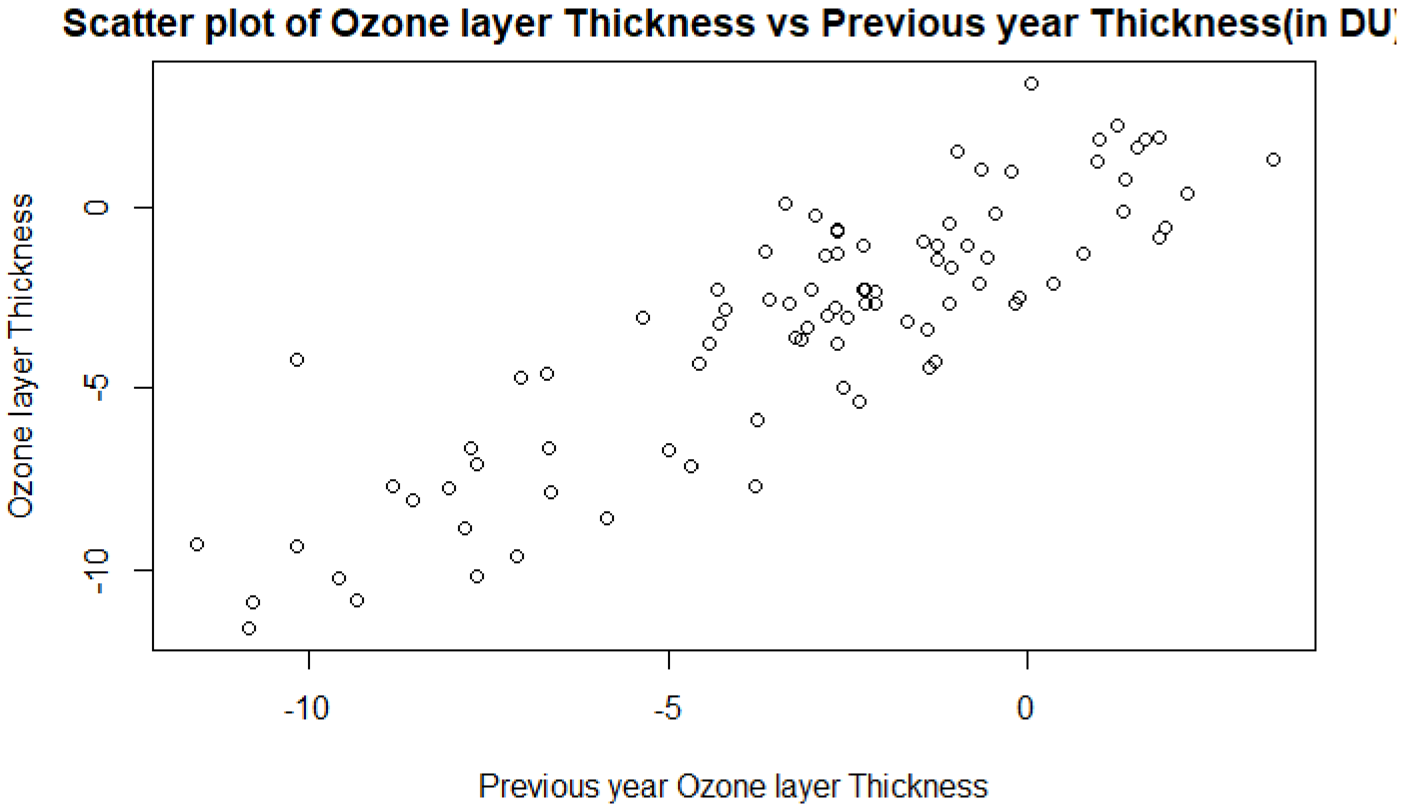
The above time series plot represents an overall negative movemnet, i.e., the thickness of ozone layer has been decreased over the years from 1927 to 2016. There is no seasonal trend that can be observed from the plot as the peaks and troughs is not appearing at regular intervals. Although, from the series it can be said that it follows an auto-regressive behaviour. The variance change can be depicted from the movement of values from high to low.

Scatter plot for neighbouring observations-

To examine if there is any relationship between the neighbouring observations, a scatter plot is plotted between the Ozone layer thickness and previous year oczone layer thiickness.

Hide

```
plot(y=Ozone_thickness,x=zlag(Ozone_thickness),ylab = 'Ozone layer Thickness', xlab = 'Previous
 year Ozone layer Thickness', main="Scatter plot of Ozone layer Thickness vs Previous year Thick
ness(in DU)")
```



The above scatter plot represents a positive trend and hence, it can be said that there is correlation between succeeding years. In other words, low values will be folllowed by likewise low values, middle sized values by likewise middle sized and those of high values by likewise high values.

The correlation value can also be calculated as follows-

Hide

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

```
[1] 0.8700381
```

The above result of correlation shows a strong correlation between Ozone layer thickness of succeeding years.

TASK 2- FINDING THE BEST FITTING TREND MODEL-

Before reaching to the final goal, i.e., giving predictions for the next five years, it is important to first find the best fitted model among LINEAR MODEL, QUADRATIC MODEL and HARMONIC MODEL, for the dataset and then carry forward with that particular model for the predictions.

1. LINEAR MODEL-

Hide

```
model_linear<-lm(Ozone_thickness~time(Ozone_thickness))
summary(model_linear)
```

```


Call:
lm(formula = Ozone_thickness ~ time(Ozone_thickness))

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(Ozone_thickness)  -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
```

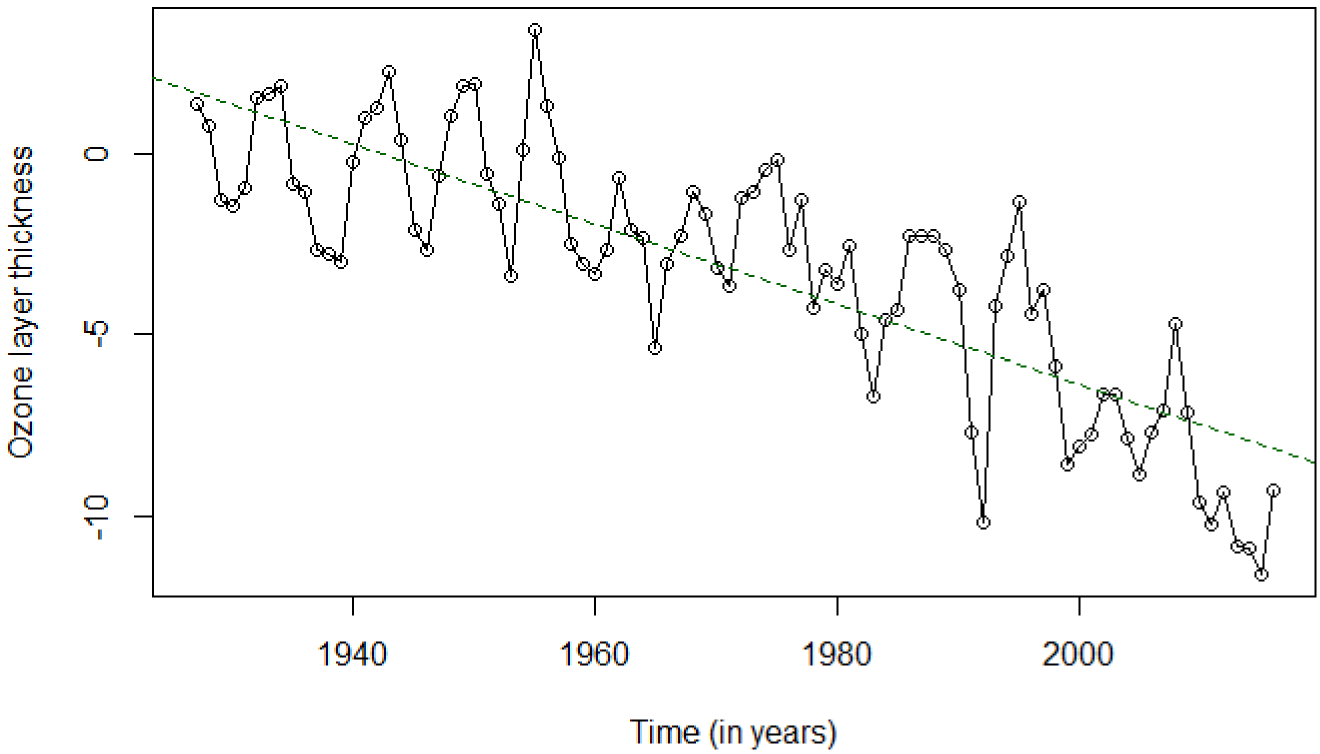
- The above summary shows that the slope and intercept coefficients are significant. The F-statistic value is also significant.
- The R-squared value indicates that around 67% of variation in the data can be explained by the linear trend.
- The adjusted R-sqaured value is also significant enough.

Therefore, for this model to be the best fitted, there should be a normally distributed white noise stochastic component, which can be analysed by the residuals, that should act like stochastic component if themodel is correct.

- SUPERIMPOSING THE LINE OF BEST FIT-

Hide

```
plot(Ozone_thickness,type='o',ylab='Ozone layer thickness', xlab='Time (in years)')
abline(model_linear,lty=2,col="dark green")
```



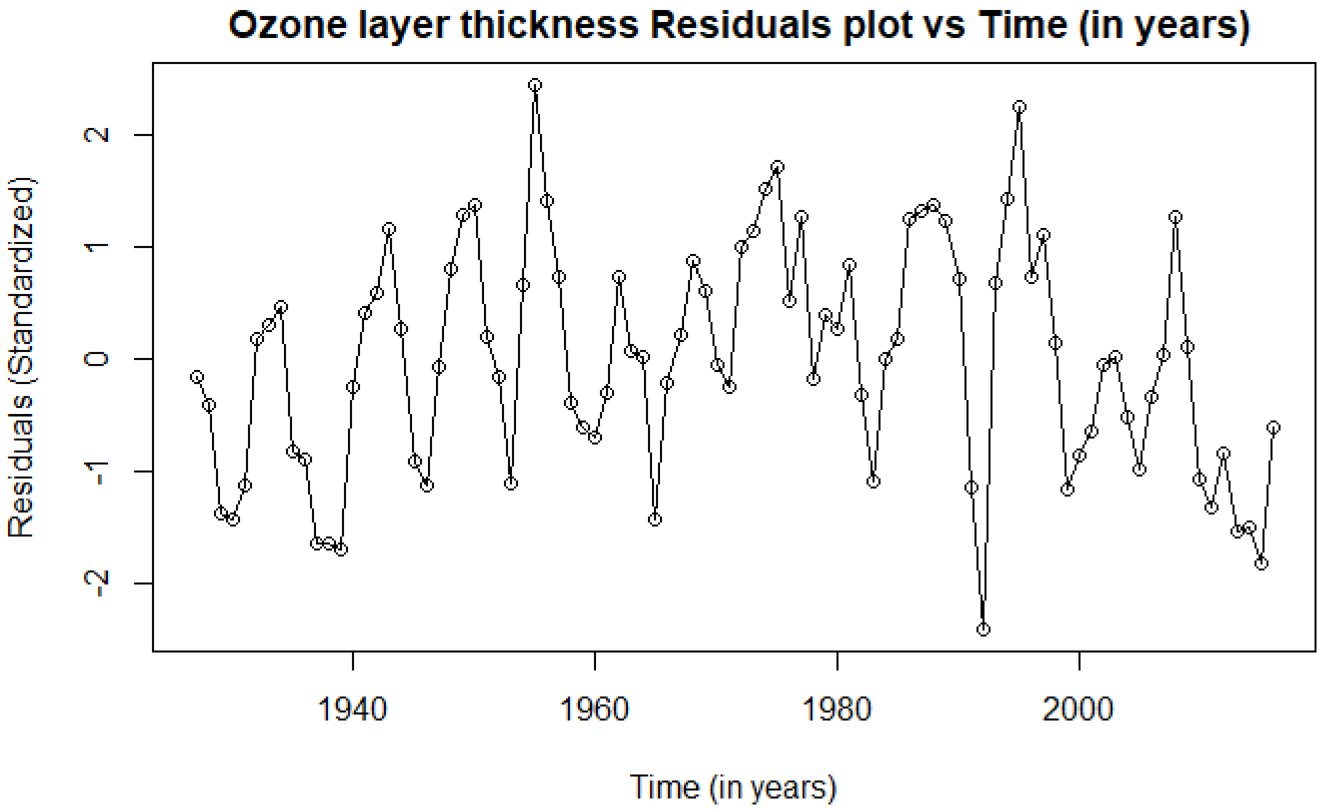
The above graph shows that the fit is significant for the series, but not capturing most of the points, that are falling above or below the fit line.

*LINEAR MODEL RESIDUALS-

The residuals should act like stochastic component, as mentioned before, if the model is correct. Also, in case the stochastic component is white noise, the residuals will behave like normally distributed random variables, having mean of Zero and Standard deviation as ‘s’.

Hide

```
residual_lm = rstudent(model_linear)
plot(y = residual_lm, x = as.vector(time(Ozone_thickness)),xlab = 'Time (in years)', ylab='Resid
uals (Standardized)',type='o', main = " Ozone layer thickness Residuals plot vs Time (in years)"
)
```



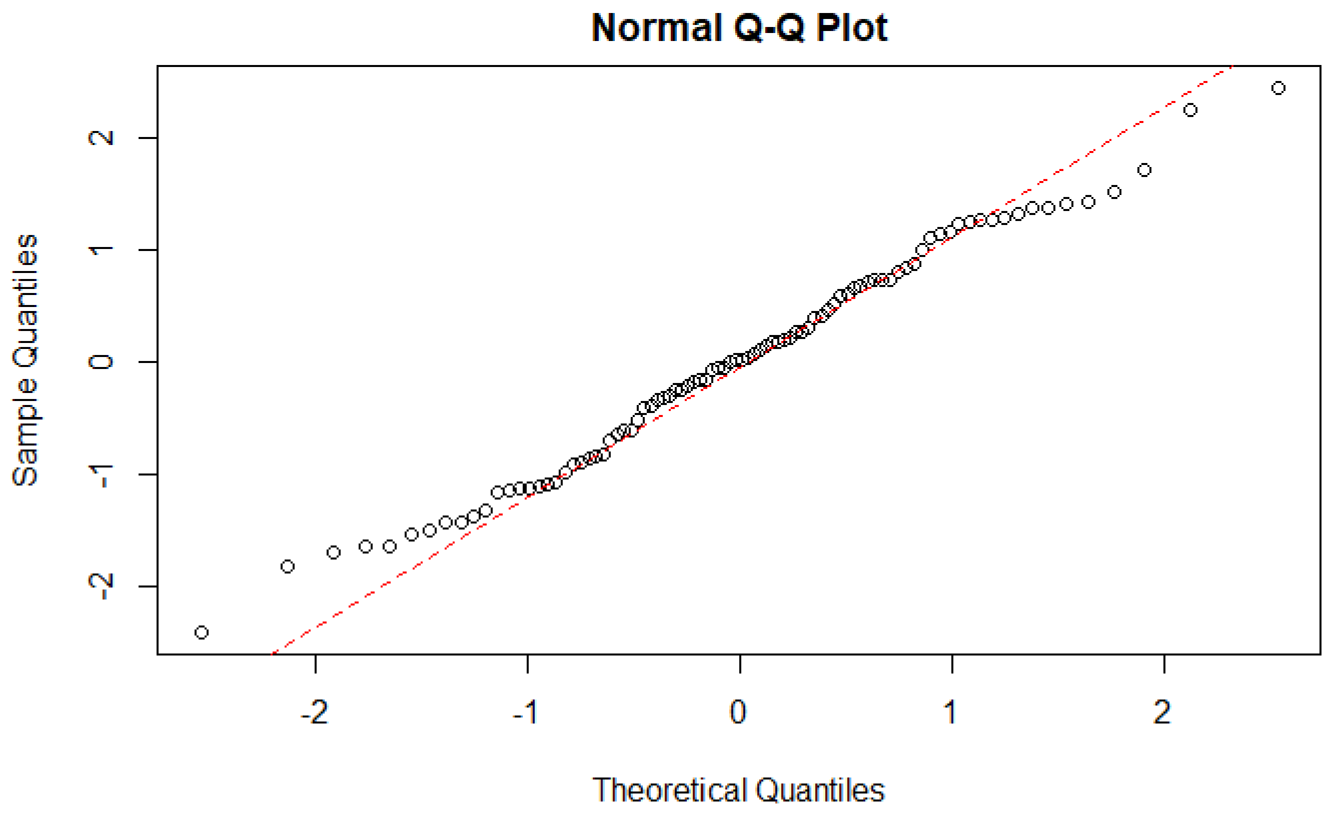
The above graph represents that residuals are not capturing the change in variance.

*NORMALITY CHECK OF LINEAR MODEL RESIDUALS (USING QQ PLOT)-

Quantile-quantile (QQ) plot appears roughly like a straight line for normally distributed values.

Hide

```
qqnorm(residual_lm)
qqline(residual_lm, col = 2, lwd = 1, lty = 2)
```



At the end tails, a deviation from normality can be observed.

*NORMALITY CHECK OF LINEAR MODEL RESIDUALS USING SHAPIRO-WILK TEST-

The Shapiro-Wilk test calculates correlation between residuals and the corresponding normal quantiles. Lower correlation indicates lower normality. Likewise, higher correlation indicates higher evidence of normality.

Hide

```
shapiro.test(residual_lm)
```

```

	Shapiro-Wilk normality test

data:  residual_lm
W = 0.98733, p-value = 0.5372
```

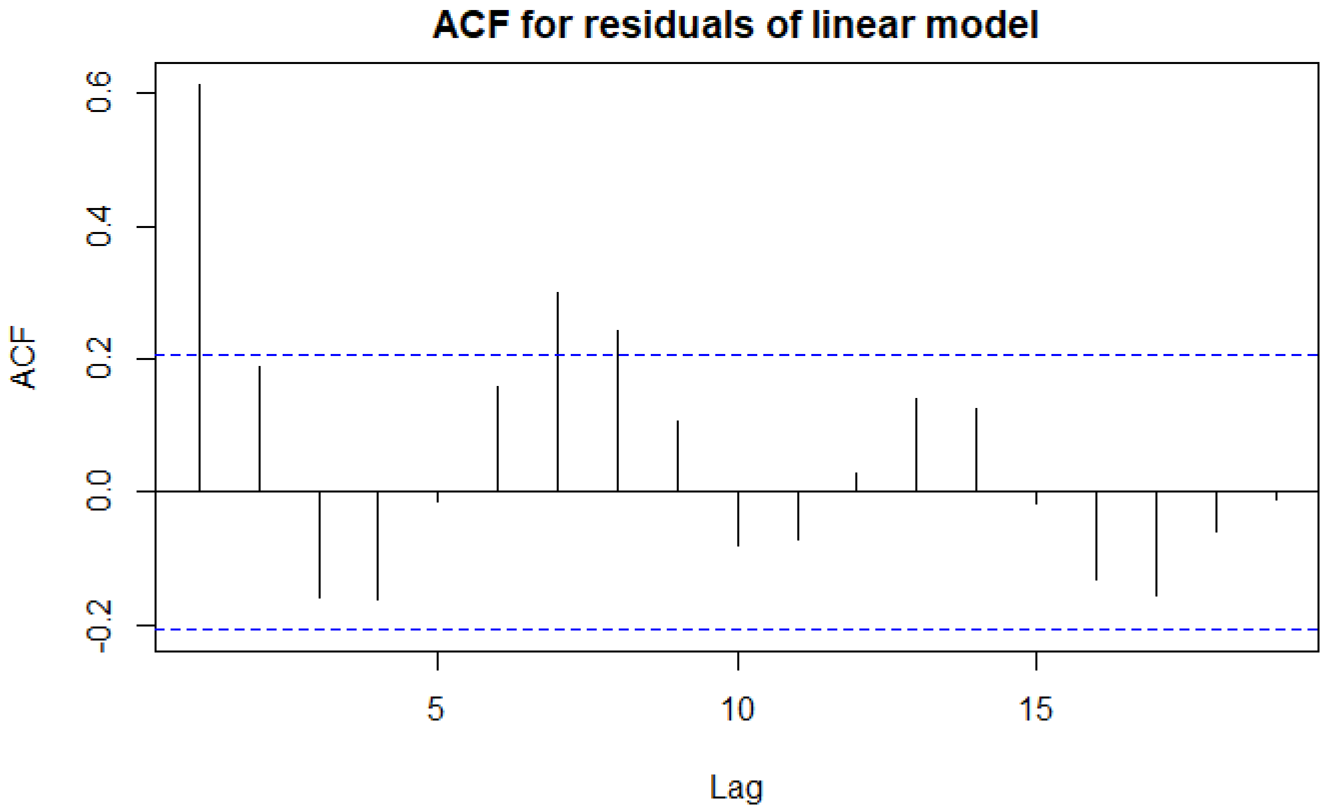
With the p-value, it can be said that we fail to reject the null hypothesis that the stochastic component for this model are normally distributed.

*CHECK FOR POSSIBLE DEPENDENCE FOR THE LINEAR MODEL RESIDUALS-

To conclude about the white noise process, Sample auto-correlation function can be used for standardized residuals.

Hide

```
acf(residual_lm, main="ACF for residuals of linear model")
```



Few of the values are falling outside the horizontal dashed lines, which can not be the case with the white noise process. Hence, stochastic component is not white noise for the series. As a result, significance dependence is present in stochastic component violating assumption of independence.

2. QUADRATIC MODEL-

Hide

```
t = time(Ozone_thickness)
t2 = t^2
model_quadratic = lm(Ozone_thickness~t + t2)
summary(model_quadratic)
```

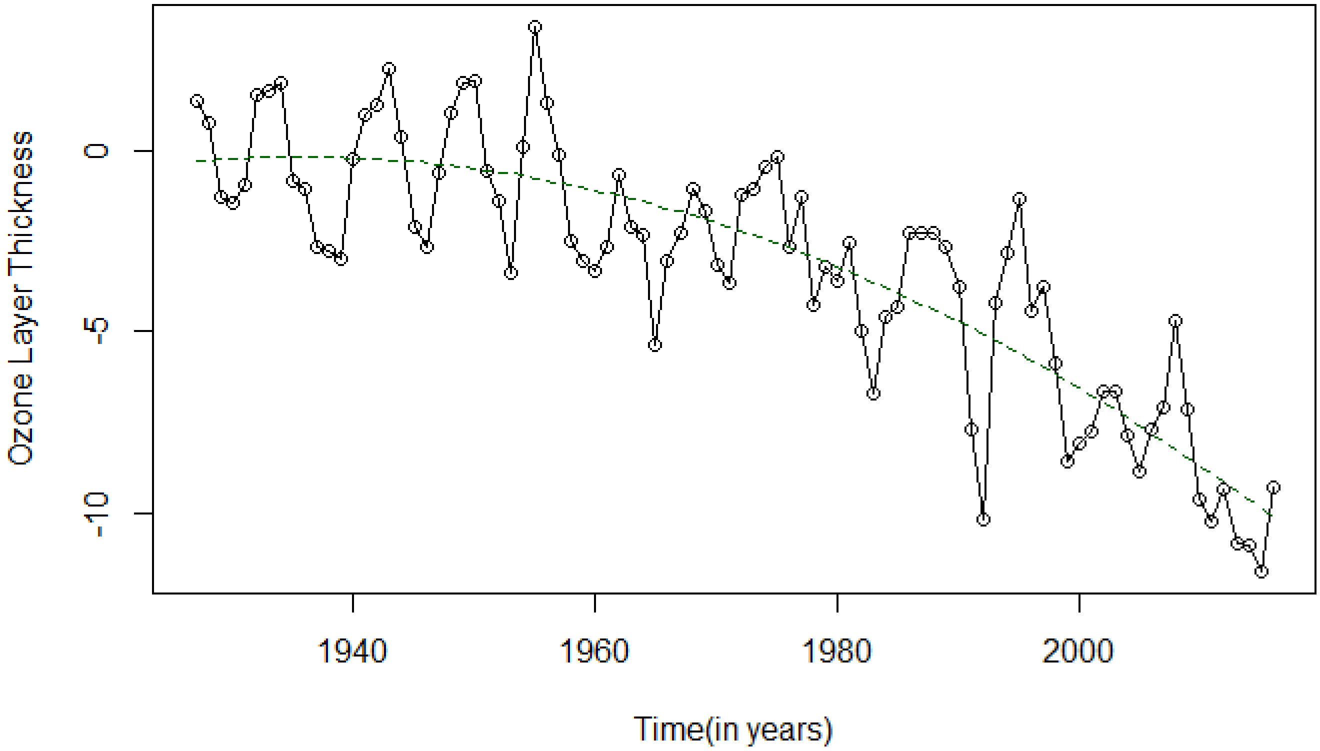
```

Call:
lm(formula = Ozone_thickness ~ 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
```

- The summary shows that the quadratic coefficients are significant. The F-statistic value is also significant.
- The R-squared value indicates that more than 73% of variation in the data can be explained by the quadratic trend.
- The adjusted R-sqaured value is also good.

- SUPERIMPOSING THE LINE OF BEST FIT OF THE QUADARTIC MODEL-

Hide

```
plot(Ozone_thickness,type='o',ylab='Ozone Layer Thickness',xlab='Time(in years)')
points(t,predict.lm(model_quadratic), type="l", lty=2,col="dark green")
```



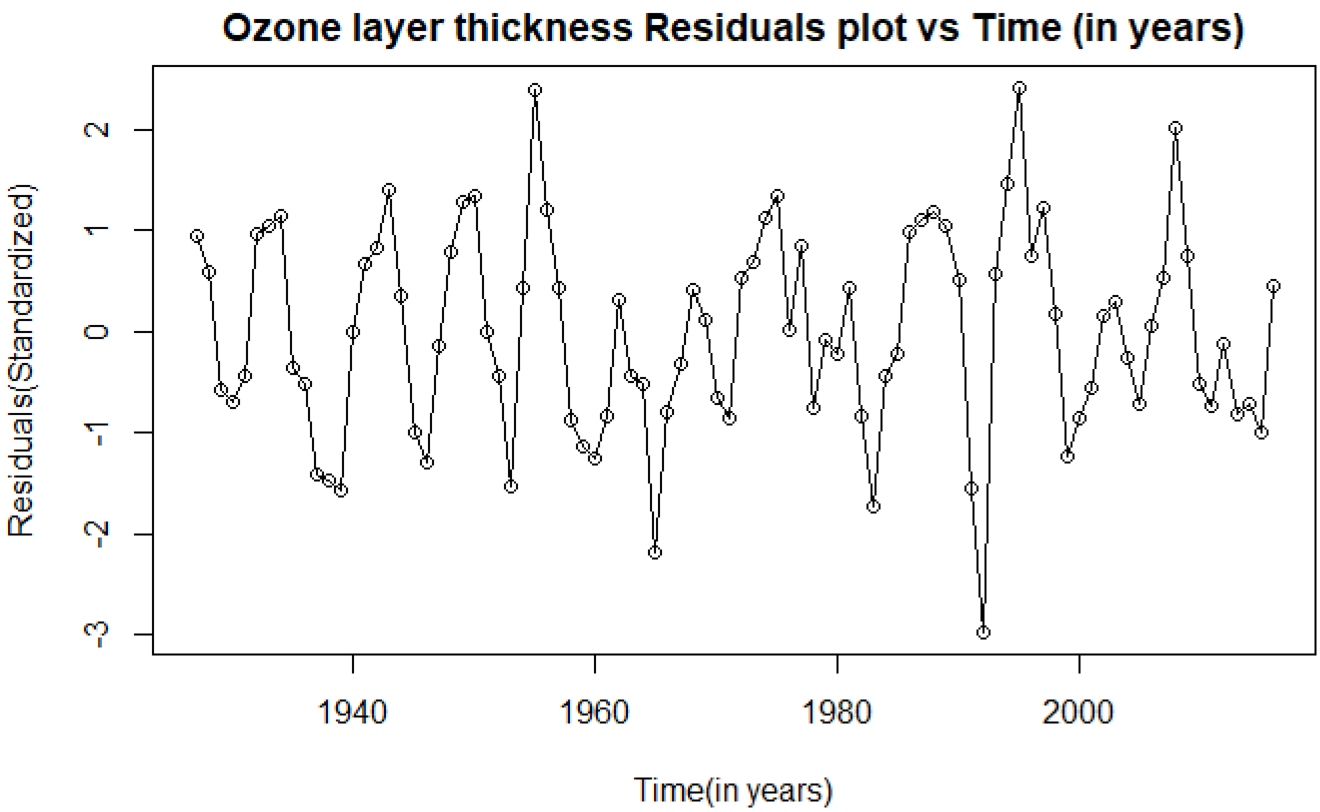
By comparing the quadratic model from linear model it can be said that R-squared value has become better fom 665 to 73% of the variation for quadratic model. Also, the fit is better for quadratic model than linear model, capturing more number of points.

*QUADRATIC MODEL RESIDUALS-

Checking of the stochastic components is white noise or not and if the residuals are behaving like normally distributed independent random variables.

Hide

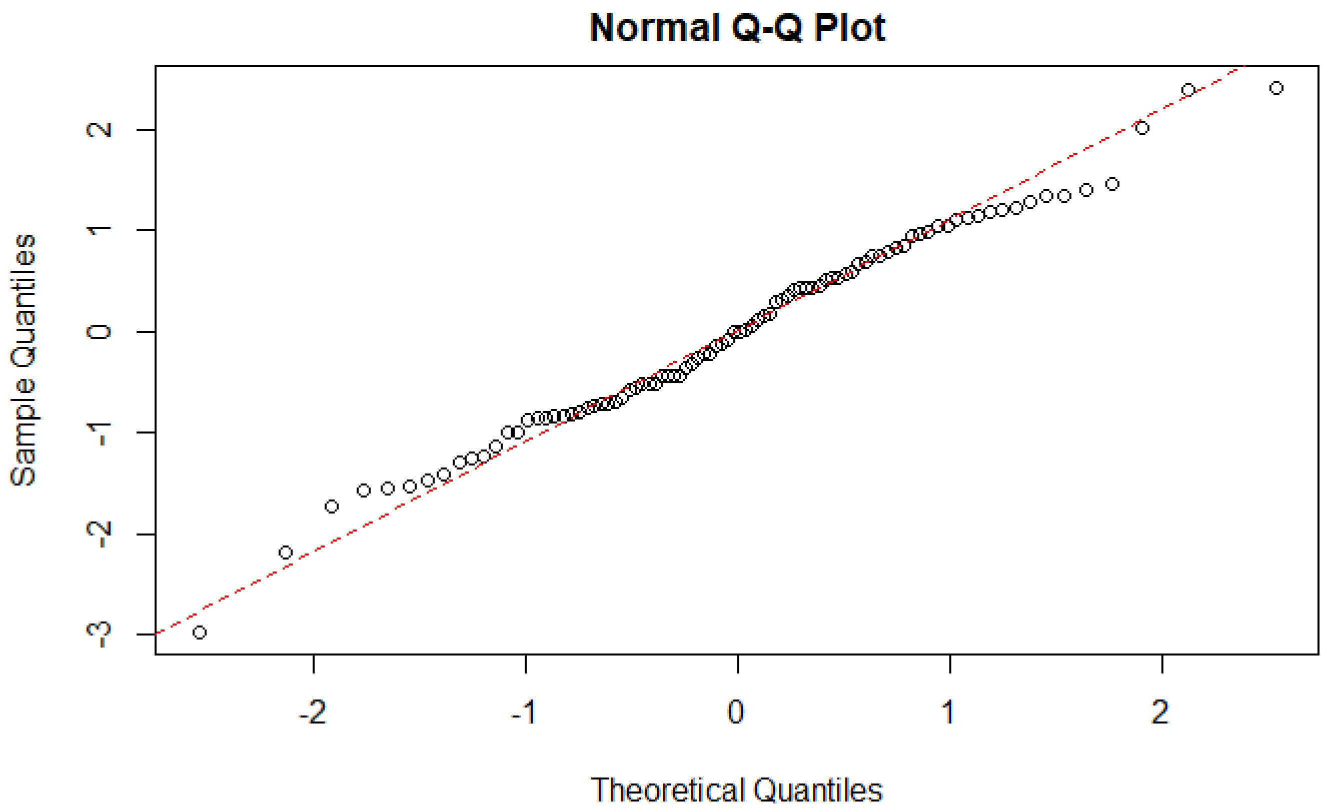
```
residual_quad = rstudent(model_quadratic)
plot(y = residual_quad, x = as.vector(time(Ozone_thickness)),xlab ='Time(in years)', ylab='Resid
uals(Standardized)', main="Ozone layer thickness Residuals plot vs Time (in years)",type='o')
```



*NORMALITY CHECK OF QUADRATIC MODEL RESIDUALS (USING QQ PLOT)–

Hide

```
qqnorm(residual_quad)
qqline(residual_quad, col = 2, lwd = 1, lty = 2)
```



Lesser deviation from normality as compared to quadratic model.

*NORMALITY CHECK OF QUADRATIC MODEL RESIDUALS USING SHAPIRO-WILK TEST-

Hide

```
shapiro.test(residual_quad)
```

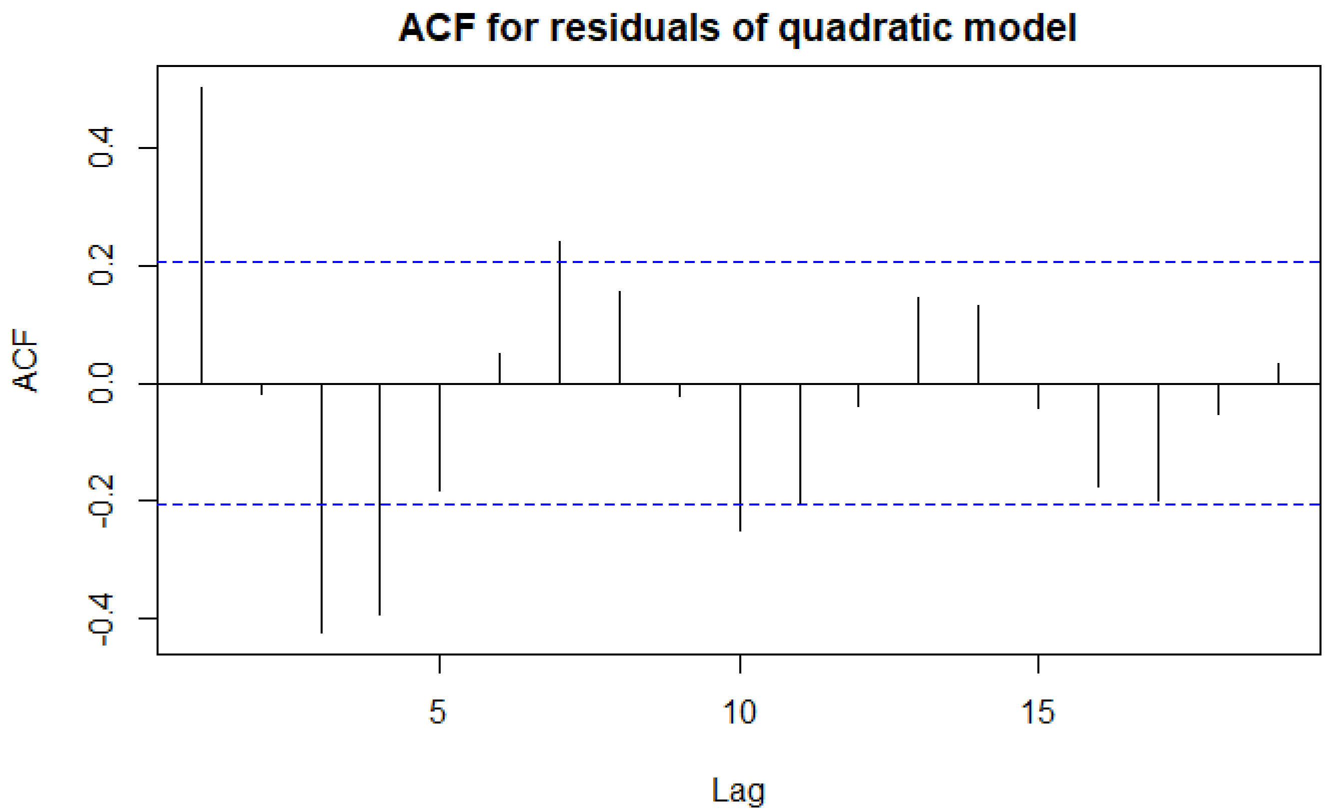
```

	Shapiro-Wilk normality test

data:  residual_quad
W = 0.98889, p-value = 0.6493
```

With the p-value, it can be said that we fail to reject the null hypothesis that the stochastic component for this model are normally distributed. Also, the score for the quadratic model is higher than that of the linear model, suggesting the reason of lesser deviation from the normality.

*CHECK FOR POSSIBLE DEPENDENCE FOR THE QUADRATIC MODEL RESIDUALS-

Hide

```
acf(residual_quad, main="ACF for residuals of quadratic model")
```



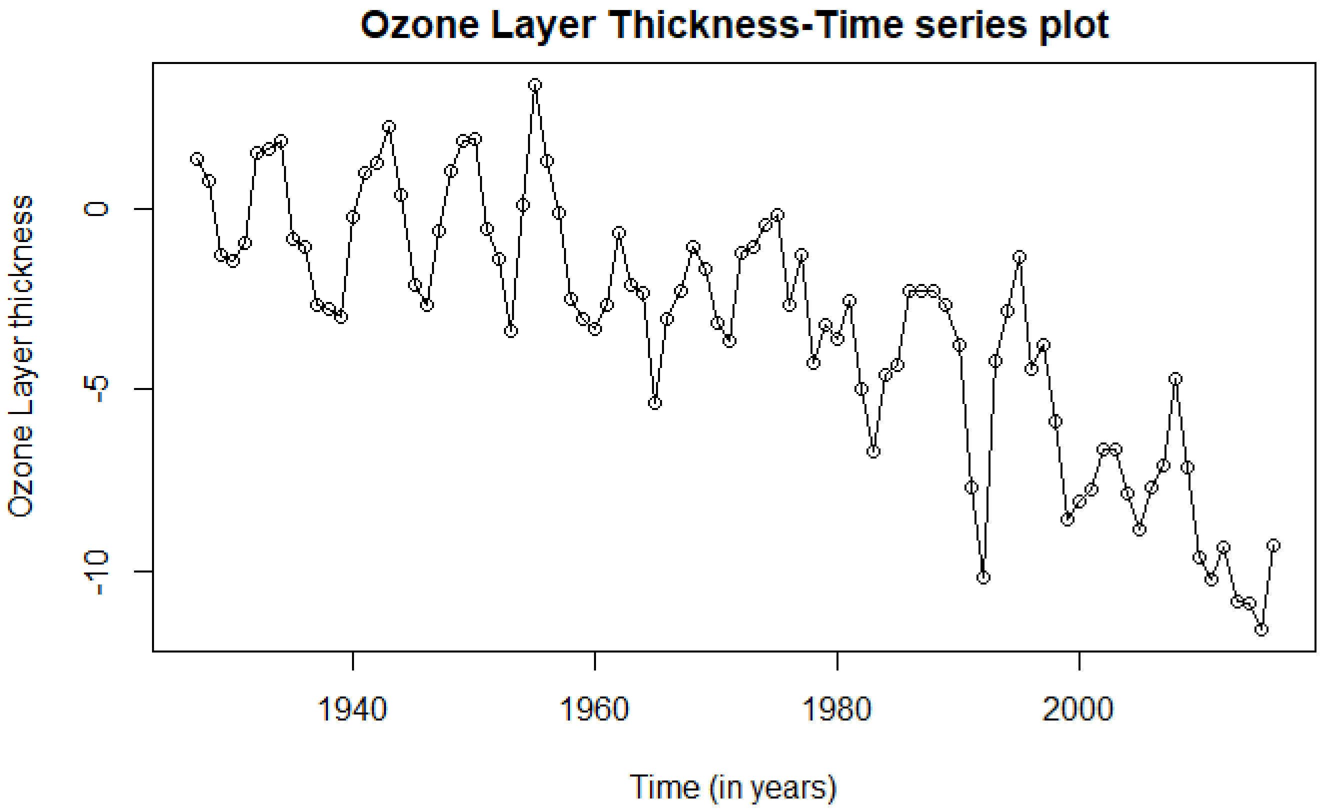
With the values falling outside the horizontal dashed lines, it can be said that stochastic component is not white noise for the series. As a result, significance dependence is present in stochastic component violating assumption of independence.

HARMONIC MODEL-

Since, there was no seasonal trend observed for the time series plot of te given data, hence, a cosine trend can not be fitted for the given data.

Hide

```
plot(Ozone_thickness,type='o',ylab='Ozone Layer thickness', xlab='Time (in years)', main="Ozone
 Layer Thickness-Time series plot")
```



RESULT-

By comparing both linear and quadratic models, quadratic model came out to be a better model as it has given higher Shapiro-Wilk values and R-squared values than those of linear model. Also, quadratic model fitted better for the given dataset. Moreover, the ACF for the residuals were not significant for both the models.

TASK 3- PREDICTIONS OF YEARLY CHANGES FOR THE NEXT 5 YEARS-

Predicting ozone layer thickness for the next five years by using quadratic model.

Hide

```
t_1 = time(Ozone_thickness)
t_2 = t_1^2
model_prediction <- lm(Ozone_thickness~t_1+t_2)
#Reading in a vector for the next 5 years
t_1 = c(2017,2018,2019,2020,2021)
t_2 = t_1^2
model_prediction_new = data.frame(t_1, t_2)
final_prediction = predict(model_prediction, model_prediction_new, interval="prediction")
print(final_prediction)
```

```
        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
```

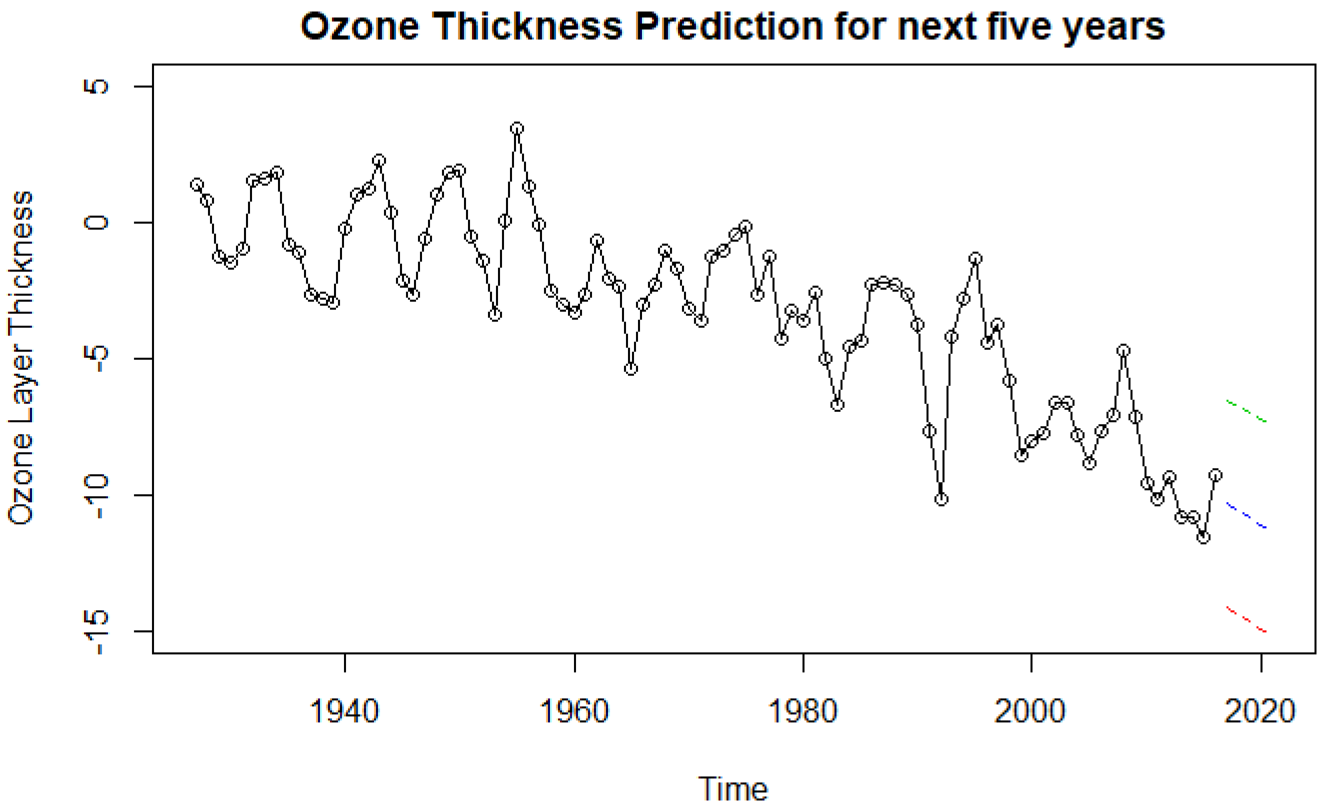
The above shows the prediction for the next five years with the upper and lower limit.

Hide

```
plot(Ozone_thickness, xlim = c(1927,2021), ylim = c(-15,5),type = 'o', ylab = "Ozone Layer Thickness", main = "Ozone Thickness Prediction for next five years")
lines(ts(as.vector(final_prediction[,1]), start = 2017), col = 4, lwd = 1, lty = 2)  #blue- predicted value
```

Hide

```
lines(ts(as.vector(final_prediction[,2]), start = 2017), col = 2, lwd = 1, lty = 2)  #red-lower limt
lines(ts(as.vector(final_prediction[,3]), start = 2017), col = 3, lwd = 1, lty = 2)  #green- upper limit
```



CONCLUSION-

Even though quadratic model came out to be the better fit for the given dataset, it still showed some deficiency with the ACF for residuals, as it violated the assumption of independence, when the ideal condition is no significant correlation for any of the lags. The reason behind this can be explained from the visualization of the time series that showed a negative trend, i.e., non-stationary, or in other words the process behaviour changes over time. Hence, for finding the best suited model, it is suggested to first convert the non-stationary process to a stationary process.