Student Name : HILL JIGISHKUMAR MODI

Student ID : S3827516

I certify that this is all my own original work. If I took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in my submission. I will show I agree to this honour code by typing "Yes": ***Yes*****.**

ASSIGNMENT 1 – TIME SERIES ANALYSIS

Time Series Analysis to find best fitted model on the Yearly Time Series data of the changes in Ozone Layer Thickness (Dobson units) and find set of possible ARIMA(p,d,q) model of the same data

Hill Jigishkumar Modi
s3827516@student.rmit.edu.au
RMIT University, City Campus
Melbourne, Australia

Date: 19/04/2021

CONTENTS

INTRODUCTION

Data is taken from Assignment Section of Time Series Analysis.

Here, we have yearly dataset of Changes in Ozone Layer Thickness in Dobson Units from year 1927 to 2016. As dataset is about change, value will be respective to the previous year. If value is positive, then it represents increase in the Ozone Layer Thickness, if negative, it represents decrease in the Ozone Layer Thickness.

AIM

The Main objective/aim of report is to determine which Model is best suitable for the data, if we observe deterministic trend. For, stochastic trend, our main goal will be to identify set of Possible ARIMA(p,d,q) Models for the annual time series data of Ozone Layer Thickness.

IMPORTANT TERMINOLOGY

MA

MA is moving average. If time-series is fluctuates between time, then we can say, behaviour of time-series is MA.

AR

AR is Auto-Regressive. If time-series has successive time-points, then it's behaviour is AR.

ACF

ACF is AutoCorrelation Function. We can determine q(order of MA) at the lag from the ACF plot. But, if there is pattern in ACF plot, then we can say q(order of MA) = 0. If there is wave like pattern, we can say there is seasonality in the data, if slowly decaying pattern, there is trend.

PACF

PACF is Partial AutoCorrelaton Function. We can determine p(order of AR) at the lag from PACF plot.

Both, ACF and PACF tells us autocorrelation at lag.

DETERMINISTIC TREND

Deterministic Trend is the trend in which we can obtain trend by straight-forwardly looking at the equation. It only depends on the time.

$y_t = \beta_0 + \beta_1 t$

Where y_t = current time-point value

β_0 = intercept,

β_1 = slope

t = time point

Deterministic trend (DT) : yt = βt + ▯t Stochastic trend (ST) : yt = β + yt−1 + ▯t ,

STOCHASTIC TREND

Stochastic Trend is the trend in which current time-point value is also dependent on the previous time-point value.

$y_t = \beta_0 + \beta_1 t + y_{t-1}$

Where y_t = current time-point value

y_{t-1} = previous time-point value

β_0 = intercept,

β_1 = slope

t = time point

STATIONARITY / NON- STATIONARITY

If mean of Time-series is same through the time points, then it is stationary.

If mean of time-series plot is changes through the time points, then series is non-stationary.

SHAPIRO-WILK TEST

Shapiro-Wilk test will check the normality.

Hypothesis

Null Hypothesis – data is normally distributed.

Alternate Hypothesis – data is not normally distributed.

Conclusion

if p-value is greater than 0.05, then we fail to reject null hypothesis.

Data is normally distributed.

AIC

AIC is Akaike's Information Criterion. AIC calculates the relative quality of each model and give as output as model with lowest AIC value among them.

BIC

BIC is Bayesian Information Criterion. It gives us model with lowest BIC value.

ARIMA(P,D,Q) MODEL

ARIMA is AutoRegressive Integrated Moving Average model.

Where p = order of AR (AutoRegressive)

d = number of difference

q = order of MA (Moving Average)

ADF TEST

ADF Test is Augmented Dicky-Fuller Test. It is used to check the stationarity of time-series.

Hypothesis

Null Hypothesis – Series is Non-stationary

Alternate Hypothesis – Series is stationary.

Conclusion

if p-value is greater than 0.05, then we fail to reject null hypothesis.

Series is non-stationary.

EACF

EACF is Extended AutoCorrelation Function.

We will look for top-most 0s with vertex. Then we will include that order of p and q to ARIMA(p,d,q) and we will also include its neighbour orders.

BIC TABLE

BIC Table is Bayesian Information Criterion Table.

It compares the BIC value with significant possible ARIMA order. If respective box is darker, then it is significant.

METHODOLOGY

First, I'll read the data, if class of data is not time-series, I'll convert the data into time-series with help of ts(). Then, I'll do descriptive analysis of data with 5 valid points and look through ACF and PACF.

TASK 1

In task 1, I will try to fit linear, quadratic, seasonal and cosine model. Then, I will do residual analysis/diagnostic checking of residuals of respective model.

After that, I'll choose the best fitting model on the basis of lowest AIC and BIC value and model with best residual analysis.

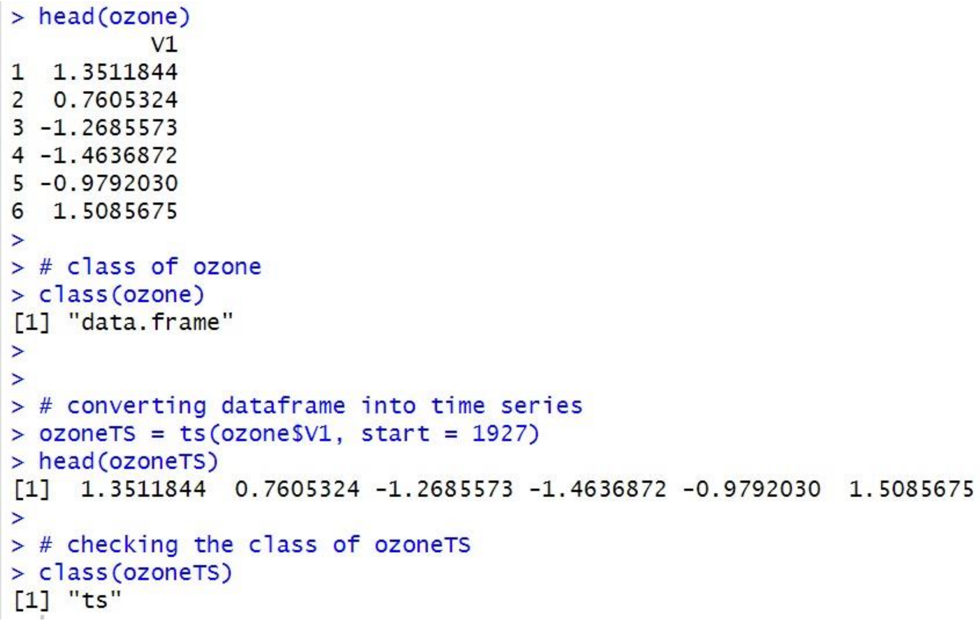
Then, I'll forecast the change of Ozone Layer Thickness for next 5 years with best fitted model.

TASK 2

In task 2, I will check if time-series data has changing variance and stationarity. If it is there, then I'll deal with it with help of log or Box-Cox transformation and with taking the differences till I found stationary time-series.

Then, I’ll find set of possible ARIMA(p,d,q) model with help of ACF, PACF, EACF and BIC table.

DESCRIPTIVE STATISTICS



```
> head(ozone)
          V1
1  1.3511844
2  0.7605324
3 -1.2685573
4 -1.4636872
5 -0.9792030
6  1.5085675
>
> # class of ozone
> class(ozone)
[1] "data.frame"
>
>
> # converting dataframe into time series
> ozoneTS = ts(ozone$V1, start = 1927)
> head(ozoneTS)
[1]  1.3511844  0.7605324 -1.2685573 -1.4636872 -0.9792030  1.5085675
>
> # checking the class of ozoneTS
> class(ozoneTS)
[1] "ts"
```

Figure 1

I have read the data, but we can’t use it for Time Series Analysis Directly. For that, class of data must by time-series. As our original data is data-frame. I converted into time-series object called ozoneTS.

Descriptive Analysis is one of the most important factor to look for. It gives us basic understanding, behaviour about data. In this section, I will discuss 5 valid points of time-series.

1.Trend

2.Changing Variance

3.Seasonality

4.Behaviour

5.Intervention Point

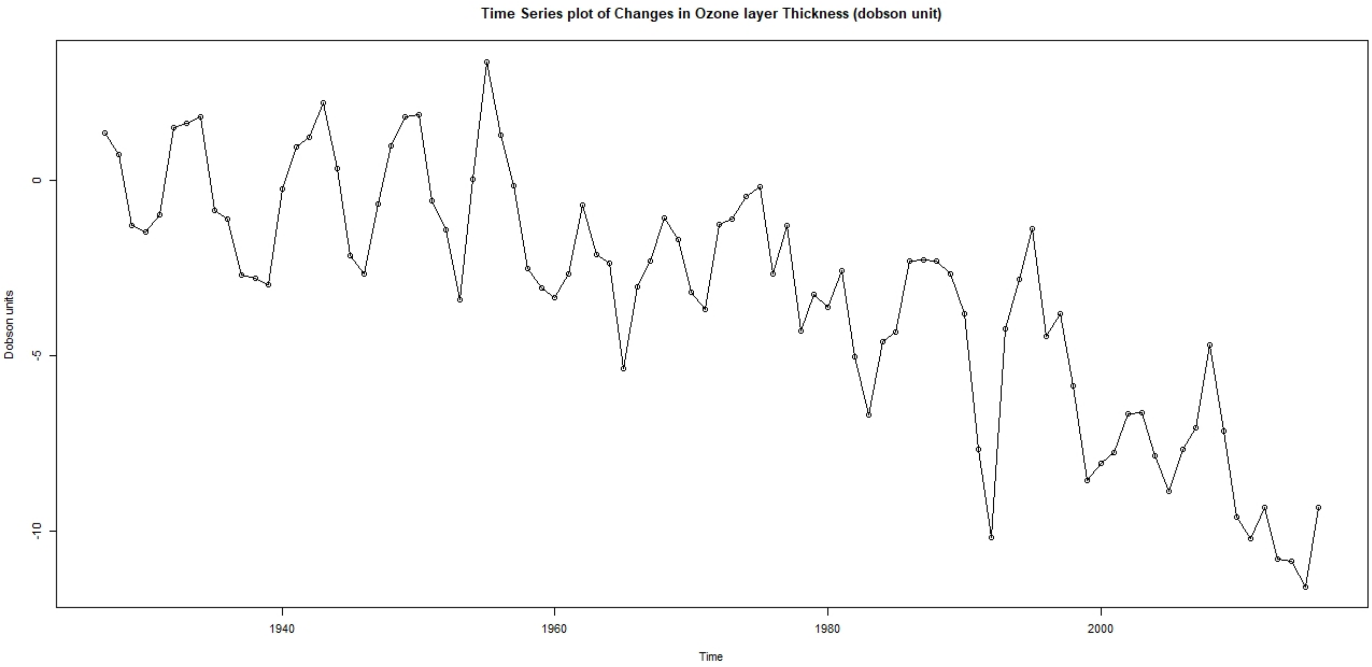


Figure 2

Figure 2 is the Time Series Plot of changes in Ozone Layer Thickness in Dobson units. From this, graph, we will notice 5 important characteristics of time series, which is described below.

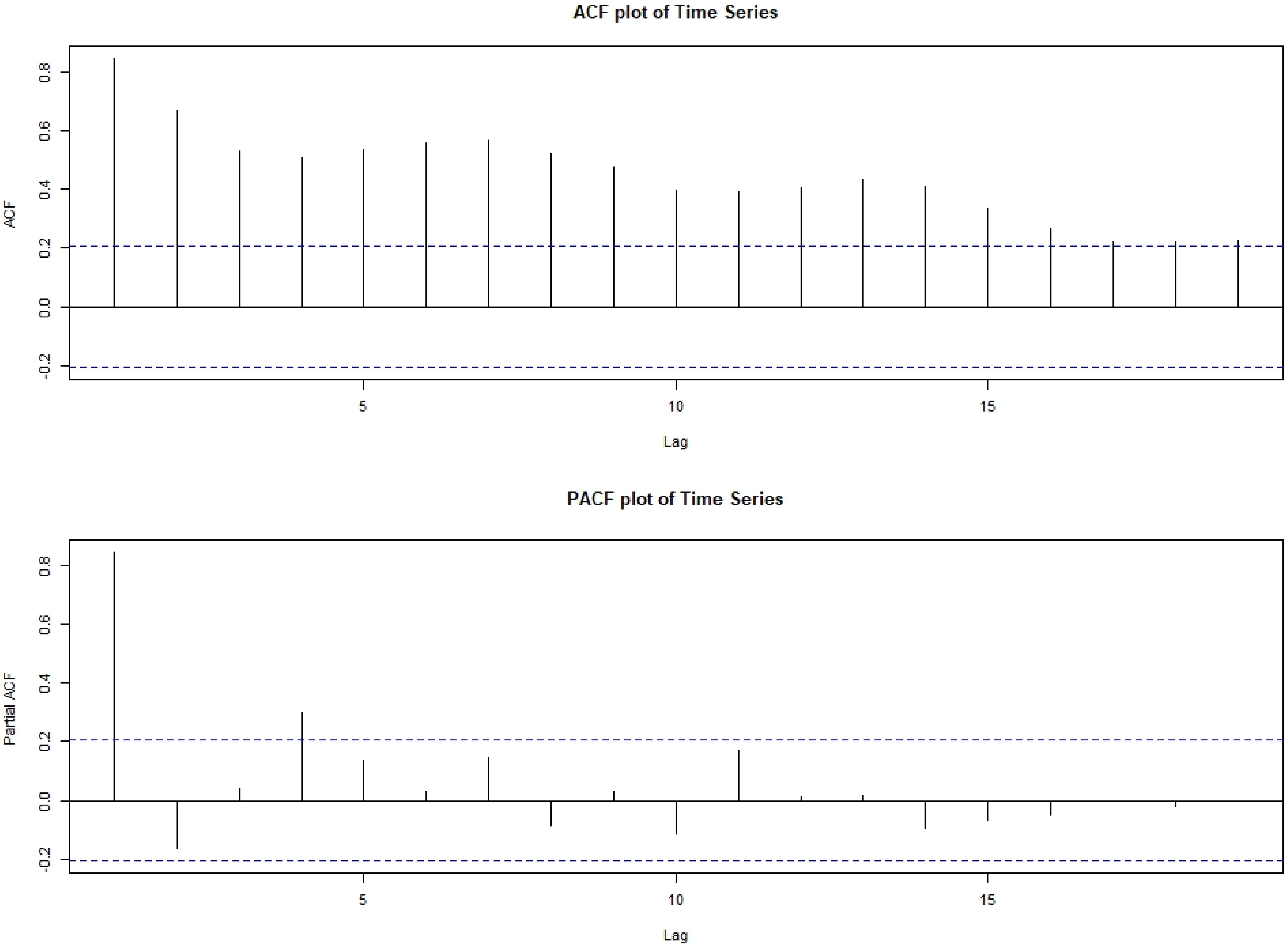


Figure 3

As we see, in the ACF plot, there is slowly decaying pattern with waves. Slowly decaying pattern indicates us there is a trend in time series. While noticeable waves represent seasonality in the time-series. We can find p from PACF plot. 1 significant bar above dashed blue line (confidence Interval) indicates that p =1. Which give us hint, series is may be Stochastic. We notice 1 slightly insignificant autocorrelation at lag 2. So, we can include p = 2 also.

TREND

From Time-Series Plot, we can say there was no trend from 1927 to 1962, as till that data was behaving live stationary data and mean was 0. However, after 1962, there is absolute strong negative trend. Slowly decaying pattern confirms our assumption. Overall, we can say that, there is an absolute trend in the time-series.

CHANGING VARIANCE

Variance was stable till 1962. After that, we can notice there are some low variance from 1970 to 1980. While, after that, from 1980 to 2000, there are huge variance indicates that variance have been changed though period.

SEASONALITY

There is obvious and strong seasonality from 1927 to 1962. However, there is also seasonality after 1962, but not strong. We can see ACF plot for Seasonality behaviour. In ACF plot, we can see wave like patterns which indicates there is Seasonality.

BEHAVIOUR

Behaviour of this series is both Moving Average and Auto Regressive. As we can see, plot is fluctuating between years and also have auto-regressive points.

INTERVENTION POINT

If after some point, any characteristic of time-series have noticeable change whether it is trend, behaviour, variance or seasonality. It is considers there is intervention point. In this time-series plot, we can say year 1962 is intervention point. As after that, we notice negative trend while there was no trend in till 1962. We also notice changing variance as discussed above.

RELATION BETWEEN CONSECUTIVE YEARS

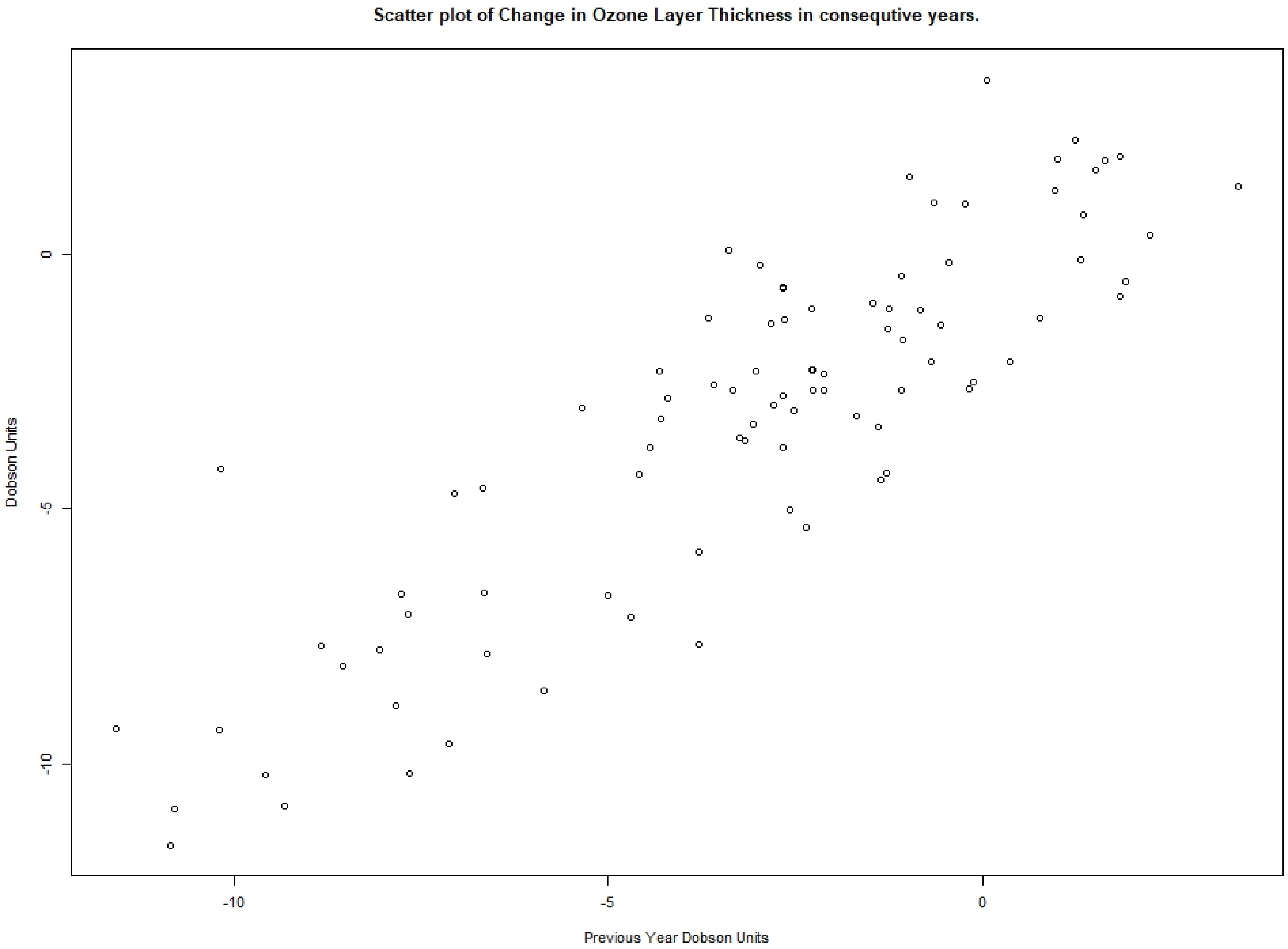
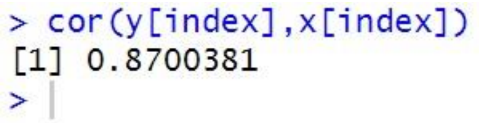


Figure 4

Figure 4 tells us changes of Ozone layer Thickness correlates to previous year or not. Linear line indicates there is strong positive correlation between previous year Dobson Unit change and Current year Dobson unit change. Hence, it is strong indication, our series has stochastic trend, not deterministic trend.



```
> cor(y[index],x[index])
[1] 0.8700381
>
```

Figure 5

Figure 5 shows correlation between consecutive years (87%). It just confirms our finding that changes in Ozone Layer Thickness is depend to the previous year.

TASK 1

The aim of Task 1 is to find the model which is best fitted to the Time Series Data.

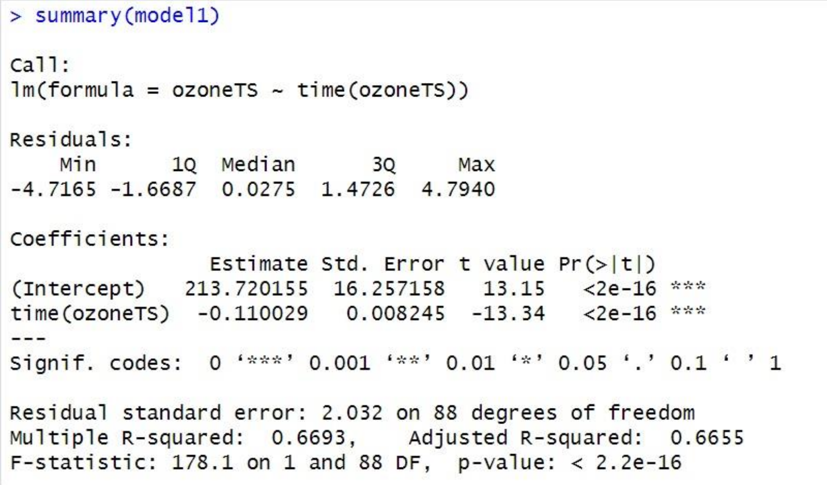
LINEAR MODEL

Linear Model is expressed as µt = β0 + β1t,

Where β0 = intercept,

β1 = slope,

t = corresponding time.



```
> summary(model1)

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

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(ozoneTS)  -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
```

Figure 6

Pr(>|t|) value of both intercept and time (slope) is < 0.05. Hence, both values are significant. While p-value of linear model is also < 0.05. Tells us, linear model is significant.

Multiple R^2 = 0.6693 – means nearly 67% of variation of Ozone Layer Thickness is explained by our Linear Model. Our model does not explained 100% if the data. Hence, we have to look at diagnostic checking of residuals.

While adjusted R^2 gives us unbiased estimation of R^2

Hence, our Linear Model is,

Change is Ozone Layer Thickness = 213.72 + (-0.11)*year

Suppose we need to find thickness on 2000 year.

= 213.72 + (-0.11)*2000

= -6.28.

We can confirm this value, by looking to the following Linear Model fitted Time-Series plot.

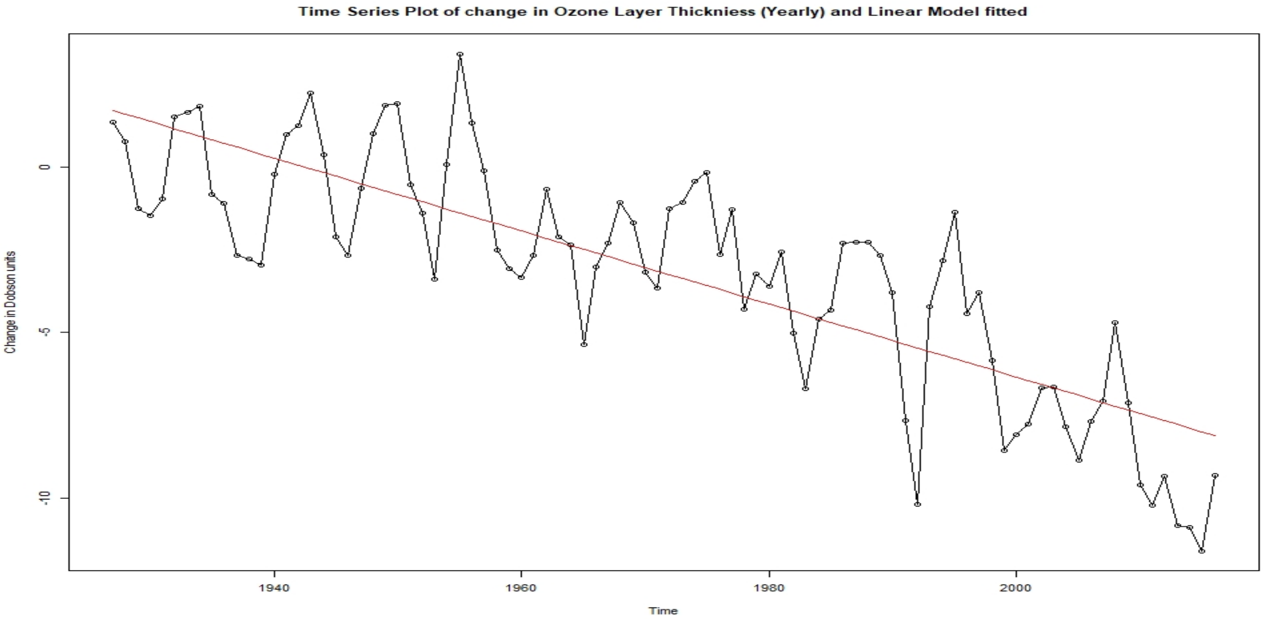


Figure 7

RESIDUAL ANALYSIS / DIAGNOSTIC CHECKING

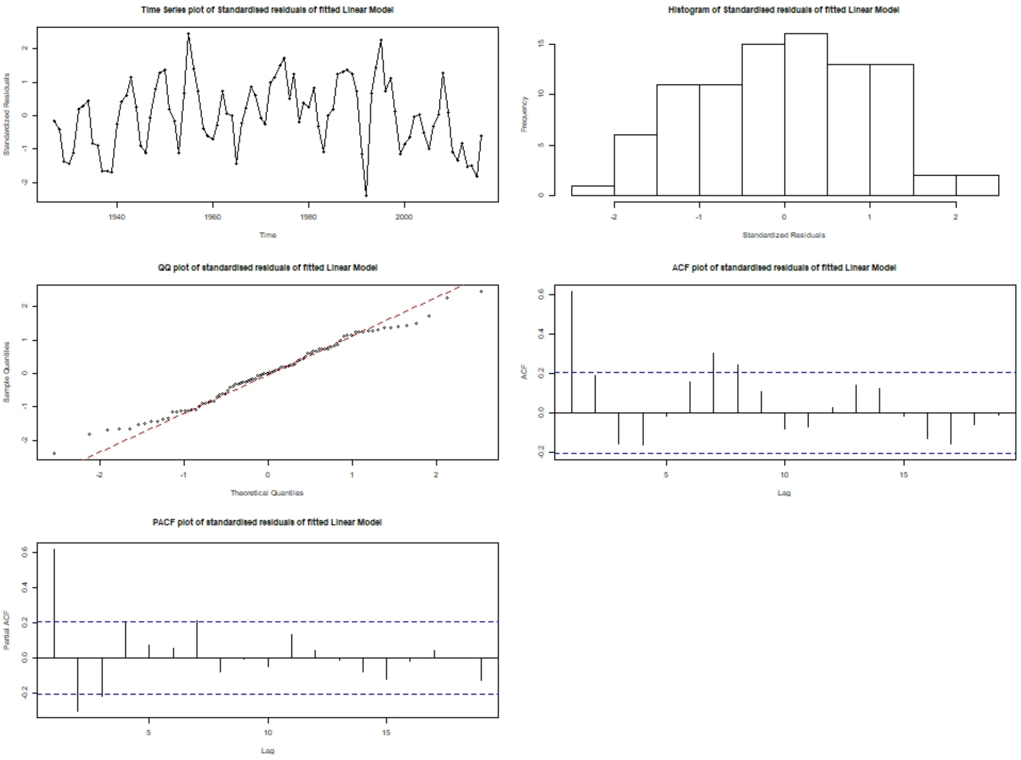


Figure 8

TIME-SERIES PLOT

We have standardised the residuals and now when we look at the time-series plot, we can say that, there is no trend and pattern is completely random. All the residuals are between 3 to -3. Hence, we observe acceptable time-series plot of residuals and no trend left in the residual analysis.

HISTOGRAM

We should expect symmetric histogram plots, as half of residuals are > 0, and rest are < 0. Here, it is satisfied. We should look at the qq-plot for the normality.

QQ-PLOT

All the residual points must be stick to the reference line. Which is true for -1 to 1 quantiles. But, after that, distance between points and line is gradually increasing.

ACF

There is significant autocorrelation at lag 1, which indicates that, there is autocorrelation in the series and our linear model is not able to capture that. It is indication that, possibly we have stochastic trend here.

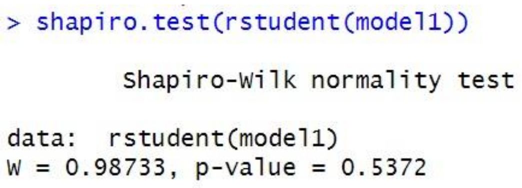
PACF

There is significant auto correlation at lag 1,2 and 3. Which are not good diagnostics.

SHAPIRO-WILK TEST

We can check normality of residuals with the Shapiro-Wilk test.

Here, p-value is >0.05. Which indicates there is normality in the residuals.



```
> shapiro.test(rstudent(model1))

	Shapiro-Wilk normality test

data:  rstudent(model1)
W = 0.98733, p-value = 0.5372
```

Figure 9

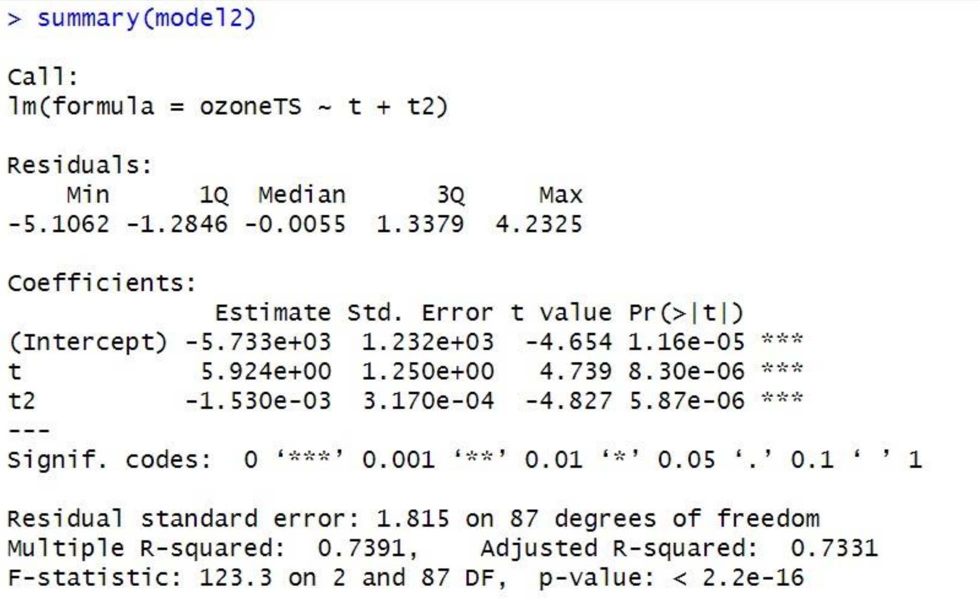
QUADRATIC MODEL

Quadratic Model is expressed as µt = β0 + β1t + β2t^2

Where β0 = intercept,

β1 = linear slope,

β2 = quadratic slope



```
> summary(model2)

Call:
lm(formula = ozoneTS ~ 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
```

Figure 10

Pr(>|t|) value of intercept and t(linear slope), t2(quadratic slope) are < 0.05. Hence, values are significant. While p-value of quadratic model is also < 0.05. Tells us, quadratic model is significant.

Multiple R^2 = 0.7391 – means nearly 74% of variation of Ozone Layer Thickness is explained by our Quadratic Model. Our model does not explained 100% if the data. Hence, we have to look at diagnostic checking of residuals.

While adjusted R^2 gives us unbiased estimation of R^2.

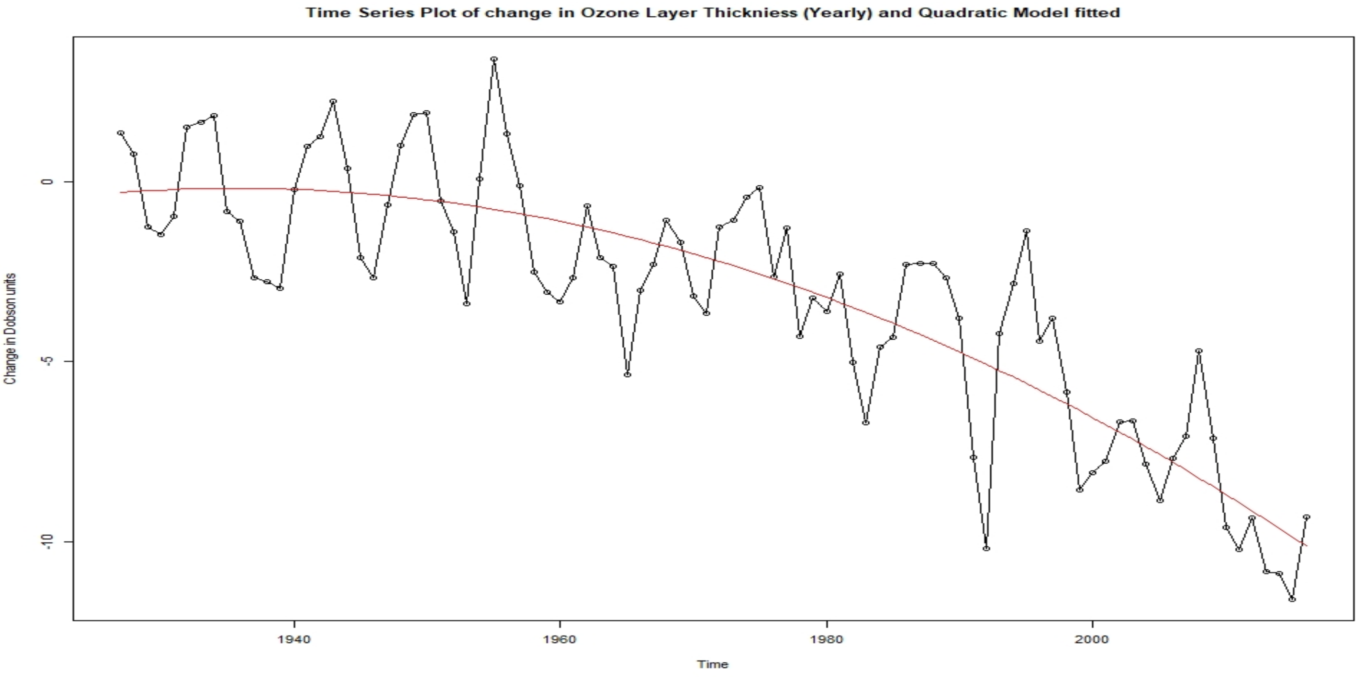
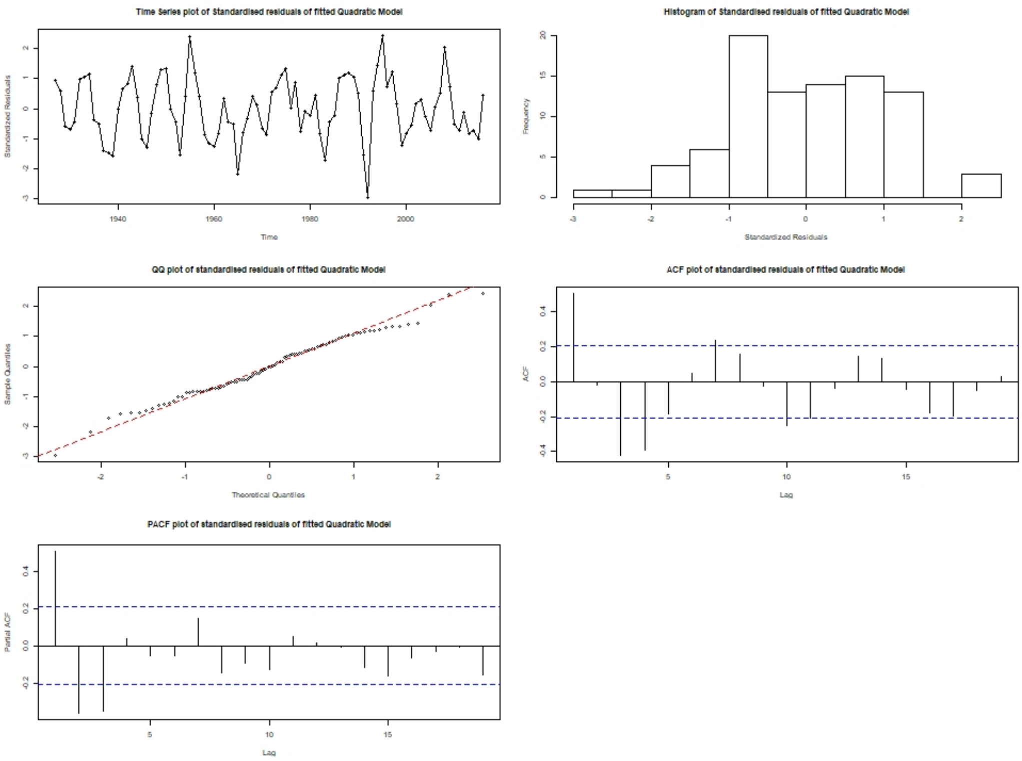


Figure 11

Figure 11 is time series plot with fitted quadratic model.

RESIDUAL ANALYSIS / DIAGNOSTIC CHECKING


Time Series plot of Standardised residuals of fitted Quadratic Model

Histogram of Standardised residuals of fitted Quadratic Model

QQ plot of standardised residuals of fitted Quadratic Model

ACF plot of standardised residuals of fitted Quadratic Model

PACF plot of standardised residuals of fitted Quadratic Model

Figure 12

TIME-SERIES PLOT

All standardised residuals are between -3 and 3. There is no trend in the residuals. Although, it looks it there is seasonality left in the residuals.

HISTOGRAM

Histogram of residuals is not excellent, as there it is unbalanced. However, it is not very bad at all. Histogram does not follow great normality. We should expect symmetric histogram plots, as half of residuals are > 0, and rest are < 0.

QQ-PLOT

All the residual points must be stick to the reference line. Which is true for -1 to 1 quantiles. But, after that, distance between points and line is gradually increasing.

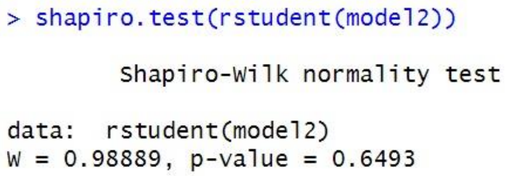
ACF

Here, we have significant autocorrelation, which indicates our model does not capture autocorrelation very well.

PACF

There is significant autocorrelation at lag 1,2,3.

SHAPIRO-WILK TEST


```
> shapiro.test(rstudent(model2))

	Shapiro-Wilk normality test

data:  rstudent(model2)
W = 0.98889, p-value = 0.6493
```

Figure 13

We can check normality of residuals with the Shapiro-Wilk test.

Here, p-value is >0.05. Which indicates there is normality in the residuals.

SEASONAL MODEL

Now, I am considering data is following seasonal trend.

But, we have annual data. So from, original time-series plot (Figure 2) and from ACF plot of Time Series (Figure 3), I am assuming that we have seasonal trend in every 7 years, as I noticed repeating wave pattern in every 7th lag in ACF indicating strong autocorrelation between current and 7th year. Same I noticed in time-series plot.

So, our seasonal model can be expressed as $Yt = \mu t + Xt$,

where $E(Xt) = 0$,

$\mu t =$ $\beta 1$ for t = 1927, 1934, 1941, …

$\beta 2$ for t = 1928, 1935, 1942, …

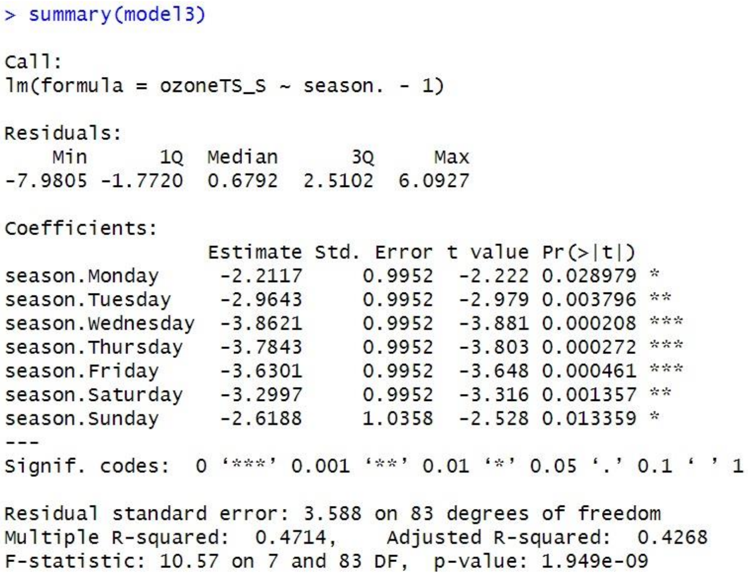
$\beta 3$ for t = 1929, 1936, 1943, …

$\beta 4$ for t =1930, 1937, 1944,…

$\beta 5$ for t =1931, 1938, 1945,..

$\beta 6$ for t = 1932, 1939, 1946,…

$\beta 7$ for t =1933, 1940, 1947



```
> summary(model3)

Call:
lm(formula = ozoneTS_S ~ season. - 1)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.9805 -1.7720  0.6792  2.5102  6.0927 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
season.Monday     -2.2117     0.9952  -2.222 0.028979 *  
season.Tuesday    -2.9643     0.9952  -2.979 0.003796 ** 
season.Wednesday  -3.8621     0.9952  -3.881 0.000208 ***
season.Thursday   -3.7843     0.9952  -3.803 0.000272 ***
season.Friday     -3.6301     0.9952  -3.648 0.000461 ***
season.Saturday   -3.2997     0.9952  -3.316 0.001357 ** 
season.Sunday     -2.6188     1.0358  -2.528 0.013359 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.588 on 83 degrees of freedom
Multiple R-squared:  0.4714,	Adjusted R-squared:  0.4268 
F-statistic: 10.57 on 7 and 83 DF,  p-value: 1.949e-09
```

Figure 14

All of the parameters corresponding to seasons are statistically significant at 5% level. Here, Monday-Saturday is displaying because R identify frequency = 7 belongs to 7 days a week. While p-value of seasonal model is also < 0.05. Tells us, seasonal model is significant.

Multiple R^2 = 0.4714 – means nearly 47% of variation of Ozone Layer Thickness is explained by our Seasonal Model. Our model does not explained 100% if the data. Hence, we have to look at diagnostic checking of residuals.

While adjusted R^2 gives us unbiased estimation of R^2.

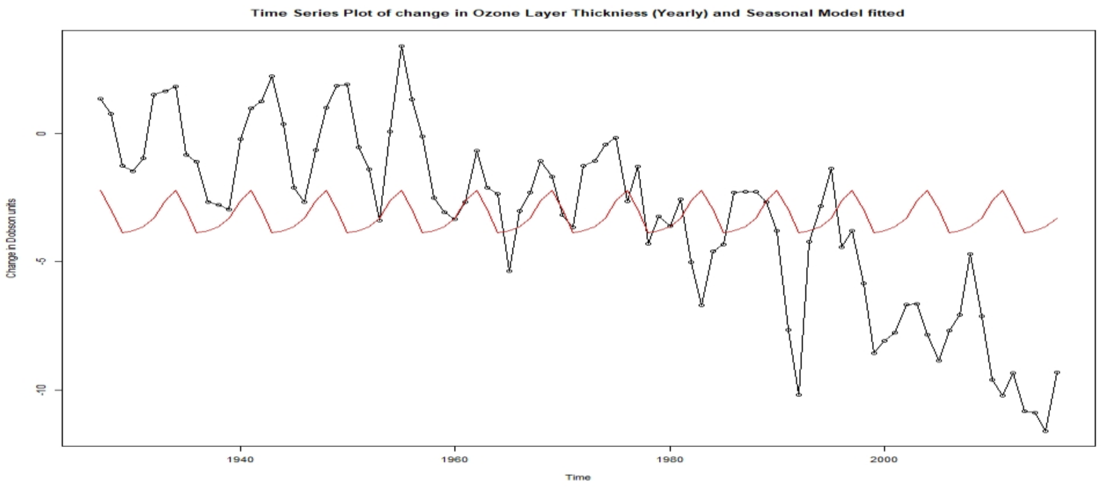


Figure 15

Figure 11 is time series plot with fitted quadratic model.

RESIDUAL ANALYSIS / DIAGNOSTIC CHECKING

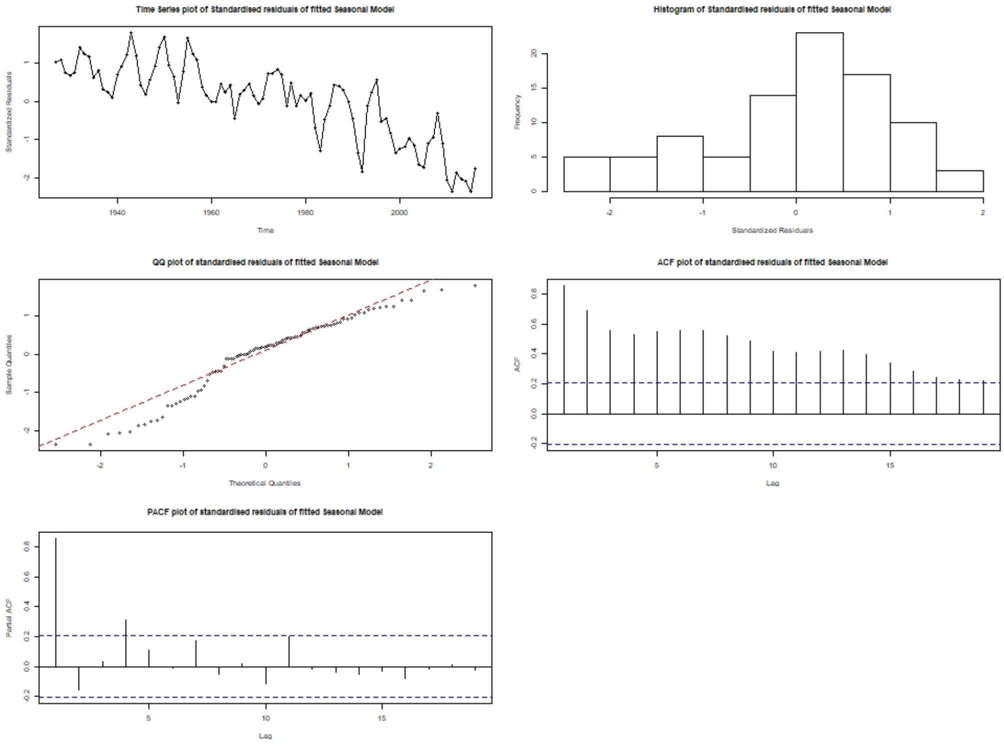


Figure 16

TIME-SERIES PLOT

We have standardised the residuals and now when we look at the time-series plot, we can say that, there is strong negative trend left in the residuals. Most of the residuals are between 2 to -2. Hence, time series plot of residual is poor.

HISTOGRAM

We should expect symmetric histogram plots, as half of residuals are > 0, and rest are < 0. Here, it is not satisfied. We have left-skewed histogram for residuals of seasonal model.

QQ-PLOT

All the residual points must be stick to the reference line. Here, points are hardly stick to the line. Poor diagnostics of qq-plot.

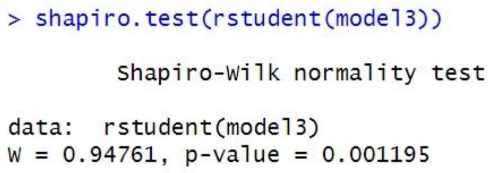
ACF

There are significant autocorrelation which indicates that, there is autocorrelation in the series and our model is not able to capture that. Apart from that, there is still seasonality left in the residuals which does not capture by our model.

PACF

There is significant auto correlation at lag 1 and 4. Which are not good diagnostics.

SHAPIRO-WILK TEST



```
> shapiro.test(rstudent(model3))

	Shapiro-Wilk normality test

data:  rstudent(model3)
W = 0.94761, p-value = 0.001195
```

Figure 17

We can check normality of residuals with the Shapiro-Wilk test.

Here, p-value is < 0.05. Which indicates residuals are not normally distributed.

Overall residual diagnostic of seasonal model is not up to the mark. It is poor.

COSINE MODEL

In the seasonal model, we separate the effect of repeated 7 years pattern.
However, it is just shape of seasonal model, but we can assign cosine curve as information to the shape of seasonal trend.
We can do it with mean function µt.
Where , µt = βcos(2πft+Φ),

Where, β(>0) = amplitude,

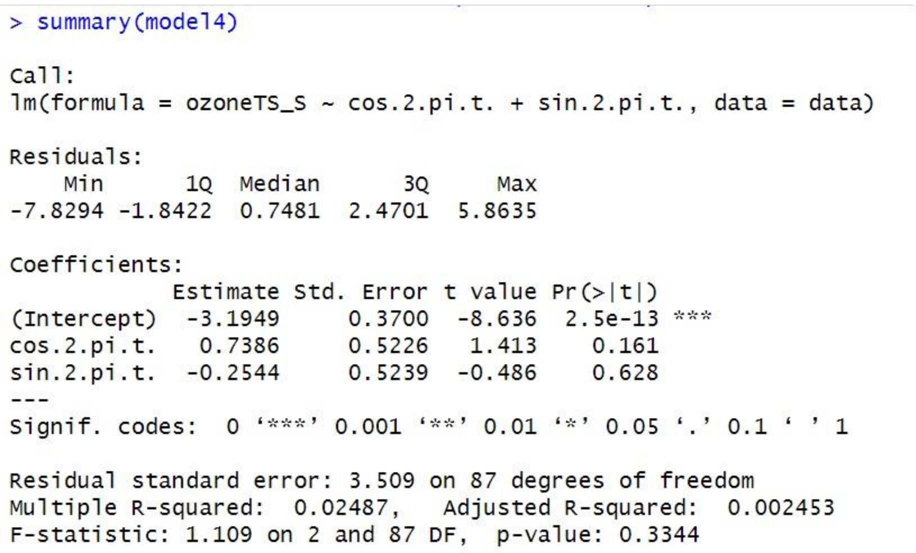
f = frequency

Φ = phase of the curve

Consequently, we will use cos(2πft) and sin(2πft) to estimate β1 and β2, respectively.

Cosine model can be expressed as µt = β0 + β1cos(2πft) + β2sin(2πft),

Where β0 = cosine with f = 0



```
> summary(model4)

Call:
lm(formula = ozoneTS_S ~ cos.2.pi.t. + sin.2.pi.t., data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.8294 -1.8422  0.7481  2.4701  5.8635 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -3.1949     0.3700  -8.636  2.5e-13 ***
cos.2.pi.t.   0.7386     0.5226   1.413    0.161    
sin.2.pi.t.  -0.2544     0.5239  -0.486    0.628    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.509 on 87 degrees of freedom
Multiple R-squared:  0.02487,	Adjusted R-squared:  0.002453 
F-statistic: 1.109 on 2 and 87 DF,  p-value: 0.3344
```

Figure 18

Pr(>|t|) value of intercept is significant. While cos and sin terms are insignificant. P value of cosine model is also insignificant, represents our model is not good for the prediction and multiple R^2 =0.02487 – means only 2% of variation of Ozone Layer Thickness is explained by our cosine Model. Hence, cosine model is appeared to be worst model for the data. Diagnostic check of the residuals would also poor.

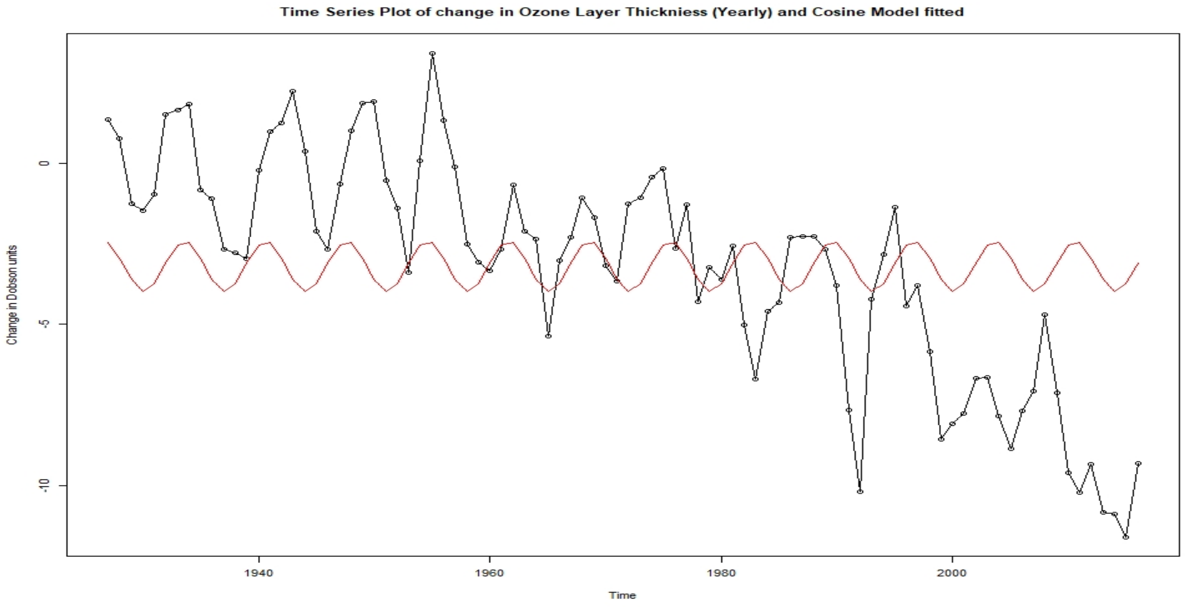
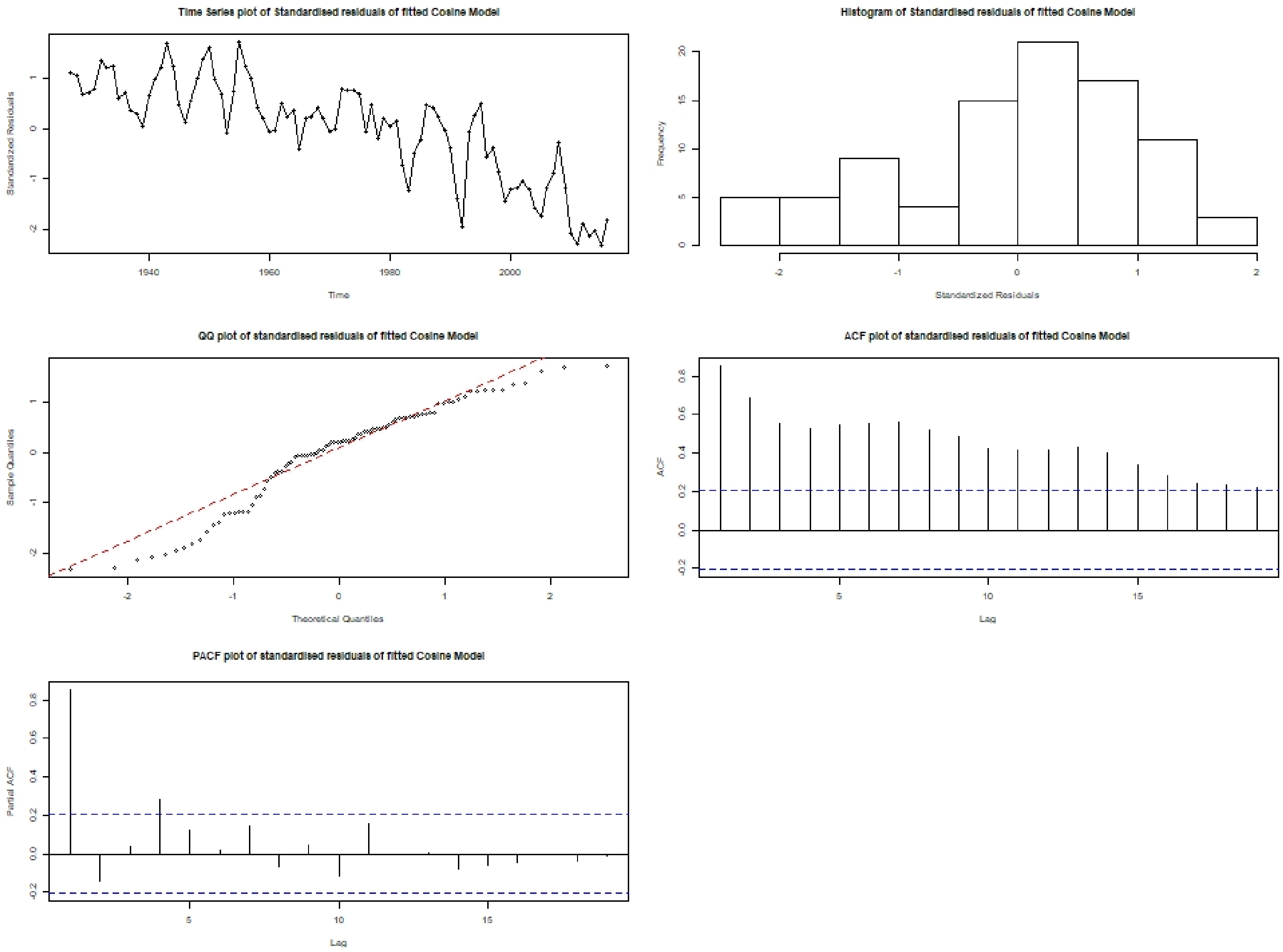


Figure 19

Figure 11 is time series plot with fitted quadratic model.

RESIDUAL ANALYSIS / DIAGNOSTIC CHECKING


Time Series plot of Standardised residuals of fitted Cosine Model

Histogram of Standardised residuals of fitted Cosine Model

QQ plot of standardised residuals of fitted Cosine Model

ACF plot of standardised residuals of fitted Cosine Model

PACF plot of standardised residuals of fitted Cosine Model

Figure 20

TIME-SERIES PLOT

We have standardised the residuals and now when we look at the time-series plot, we can say that, there is strong negative trend left in the residuals. Most of the residuals are between 2 to -2. Hence, time series plot of residual is poor.

HISTOGRAM

We should expect symmetric histogram plots, as half of residuals are > 0, and rest are < 0. Here, it is not satisfied. We have left-skewed histogram for residuals of cosine model.

QQ-PLOT

All the residual points must be stick to the reference line. Here, points are hardly stick to the line. Most of point are away fom the reference line. Poor diagnostics of qq-plot.

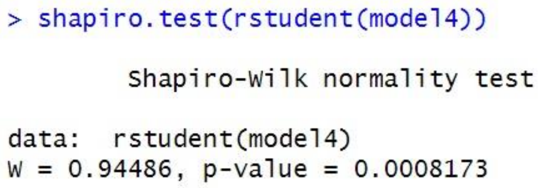
ACF

There are significant autocorrelation which indicates that, there is autocorrelation in the series and our model is not able to capture that. Apart from that, there is still seasonality left in the residuals which does not capture by our model.

PACF

There is significant auto correlation at lag 1 and 4. Which are not good diagnostics.

SHAPIRO-WILK TEST



```
> shapiro.test(rstudent(model4))

        Shapiro-Wilk normality test

data:  rstudent(model4)
W = 0.94486, p-value = 0.0008173
```

Figure 21

We can check normality of residuals with the Shapiro-Wilk test.

Here, p-value is < 0.05. Which indicates residuals are not normally distributed.

Overall residual diagnostic of seasonal model is not up to the mark. It is extremely poor.

FINDING BEST MODEL

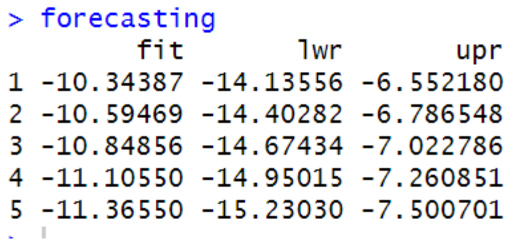
FORECASTING MODEL

I have used Quadratic model to predict the next 5 year change in Ozone Layer Thickness. I generated 5 ahead years from the last observe year with h = 5. Hence ahead years will be 2017 – 2021.

Then I created dataframe with time (t) and time^2(t2),

Where t = 2017,..,2021.

Then I predict the change in Ozone Layer Thickness for (2017- 2021) with help of predict() function and combined original data with the predicted data to plot the forecasting.



```
> forecasting
        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
```

Figure 23

Here, we have prediction of change in Ozone Layer Thickness (Dobson Unit) with lower and upper confidence Interval for Quadratic model.

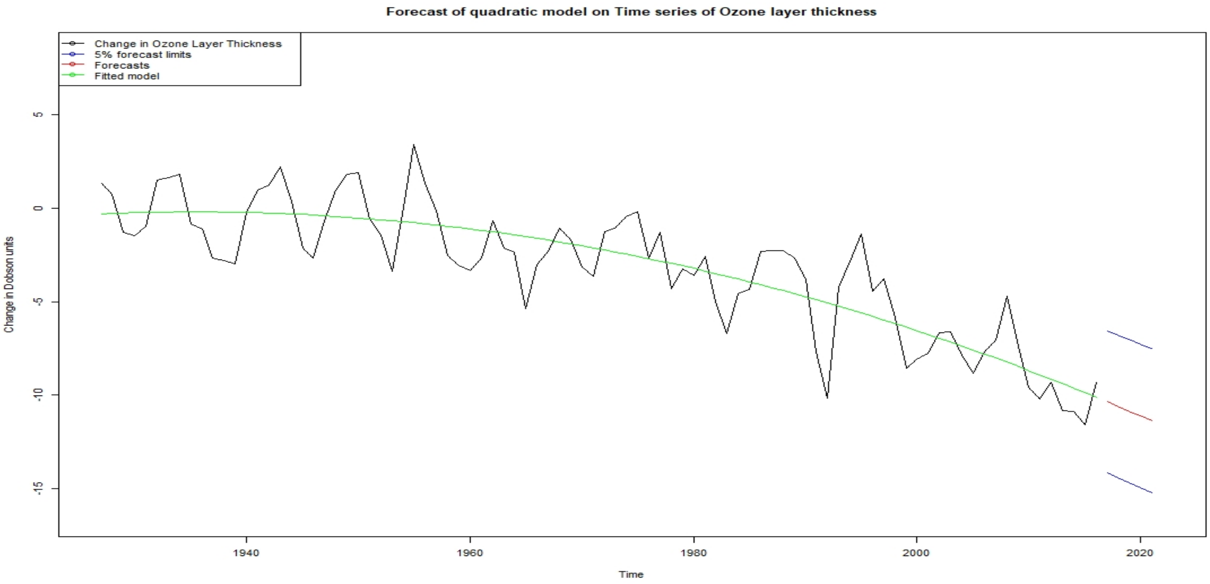


Figure 24

In figure 23, red line indicates forecast for next 5 years, while blue lines indicate upper and lower confidence Interval. From our prediction, we can say that, Thickness of Ozone Layer in upcoming 5 years will be decrease.

CONCLUSION

In a nutshell, I tried four models to explain the time-series data.

1. Linear Model

2. Quadratic Model

3. Seasonal Model

4. Cosine Model

From which, quadratic model best explained the data with best residual diagnostics among all four models. AIC and BIC also suggests to go with quadratic model.

Next best model is linear model with 66% variation explanation of data and with good residual diagnostics.

While seasonal and cosine models were not good enough to explain the change in Ozone Layer Thickness.

I noticed auto-correlation in all the ACF and PACF plot of residual analysis, which indicate that it is not the deterministic trend, but it is stochastic trend. Hence, we should use ARIMA(p,d,q) model and if we notice seasonality, we should use SARIMA model.

From forecasting, we can conclude that, Thickness of Ozone Layer will be decreased.

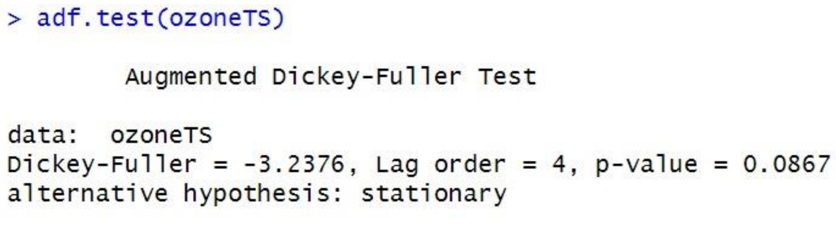
For example, -10.35 Dobson Units in 2017, with lower confidence Interval of -14.14 and upper confidence Interval of -6.55.

TASK 2

Here, our aim is to determine the set of ARIMA(p,d,q) model.

CHECKING STATIONARITY

First, we need to check the stationarity of data. For that, I will use Augmented Dicky-Fuller Test.



```
> adf.test(ozoneTS)

        Augmented Dickey-Fuller Test

data:  ozoneTS
Dickey-Fuller = -3.2376, Lag order = 4, p-value = 0.0867
alternative hypothesis: stationary
```

Figure 25

Here, p-value is 0.0867, which is > 0.05.

Hence, we fail to reject null Hypothesis. Which means data has Unit root and series is non-stationary.

It represents there is a trend in the series which is correct (by looking at Figure 1). Hence, we need to convert the non-stationary time – series into Stationary time-series. But, before that, we need to transform the series with log transformation or with the BoxCox transformation to deal with the changing variance.

LOG TRANSFORMATION

Here, we have negative values in the data points, so I can't directly apply log transformation. First I need to move the series, so there won't be any negative or zero value. For this, I added the (absolute of minimum + 1) to the entire data. Then I transformed series into log transformation.

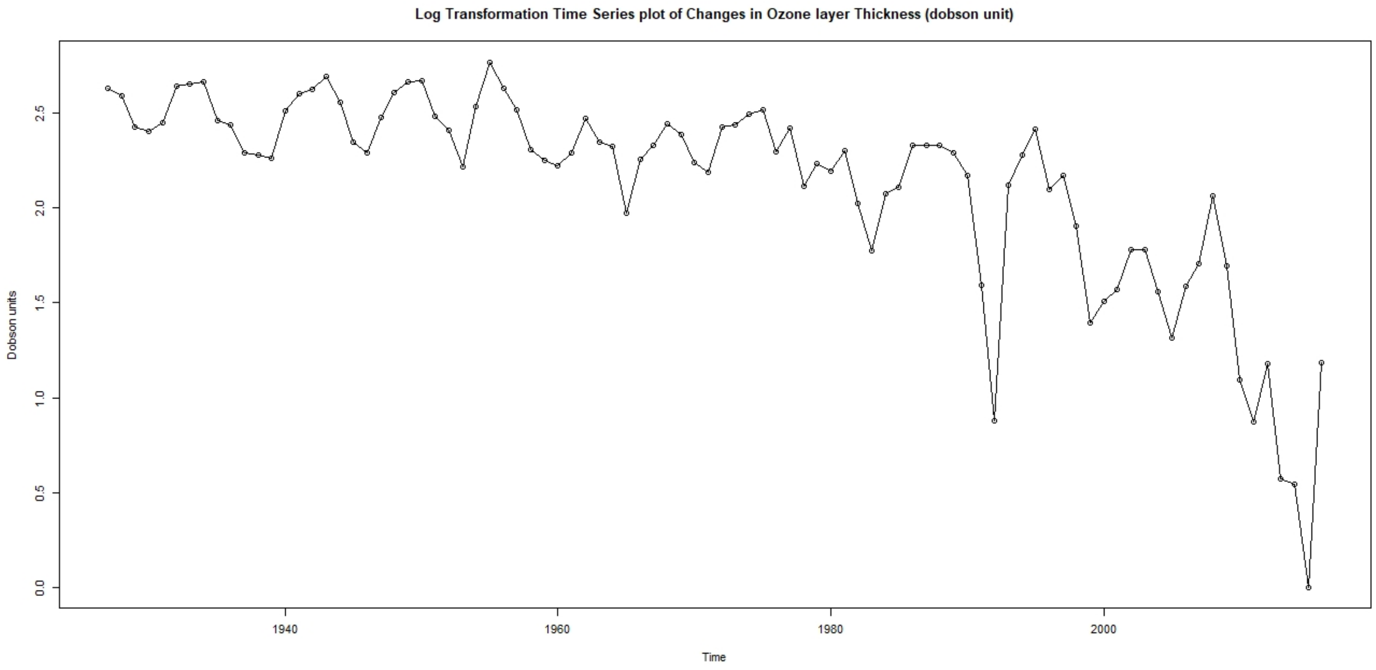
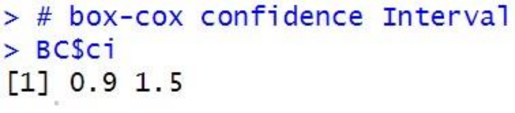


Figure 26

Figure 26 shows the log transformed time Series. But, there is still changing variance in later year compared to early year. Now, I'll transformed the series with box-cox transformation.

BOX-COX TRANSFORMATION

Confidence Interval of maximum likelihood of Box-Cox transformation is 0.9 (lower) and 1.5(upper), which is also plotted on Figure 28.



```
> # box-cox confidence Interval
> BC$ci
[1] 0.9 1.5
```

Figure 27

Now, I have find maximum likelihood (lambda) of Box-Cox transformation. It is middle vertical dashed line.

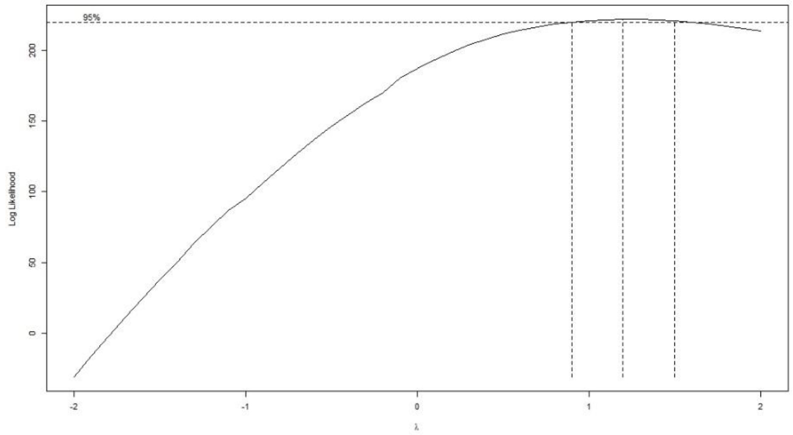
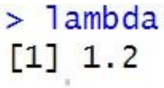


Figure 28



```
> lambda
[1] 1.2
```

Figure 29

Here, maximum likelihood(lambda) = 1.2, so I did Box-Cox transformation with lambda = 1.2.

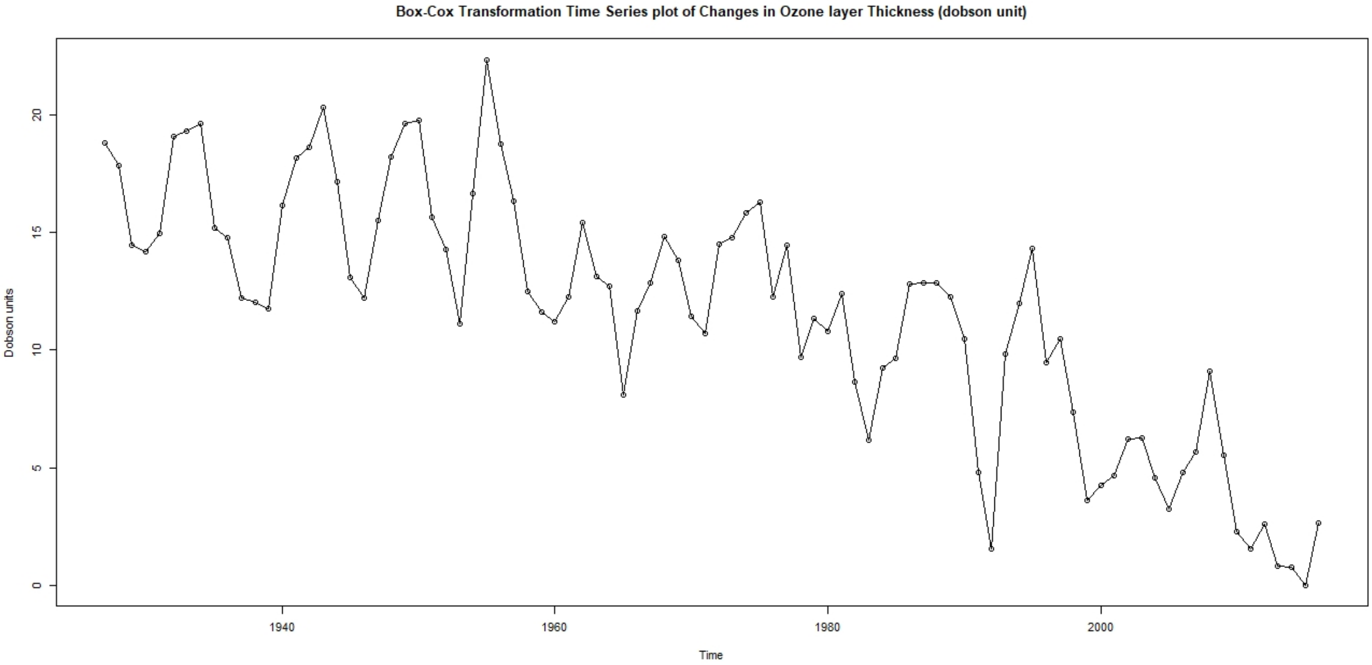


Figure 30

From Figure 30, I can say that, we have better time-series plot than log transformed time-series plot in terms of changing variance. Because, log transformation is done with lambda = 0, while value maximum likelihood (lambda) = 1.2.

Time-series needed to be transformed with the lambda = 1.2. Box-Cox transformation would be same to original time-series data for lambda = 1. Perhaps, that is the reason, we don't have noticeable changing variance because lambda = 1.2 is near to lambda = 1.

FIRST DIFFERENCE AFTER BOX-COX TRANSFORMATION

Now, as we have dealt with changing variance, we have to deal with negative trend presence in the time-series. We need to convert our non-stationary time-series with stationary time-series. We will do this by taking the 1st difference of Box-Cox transformed time-series.

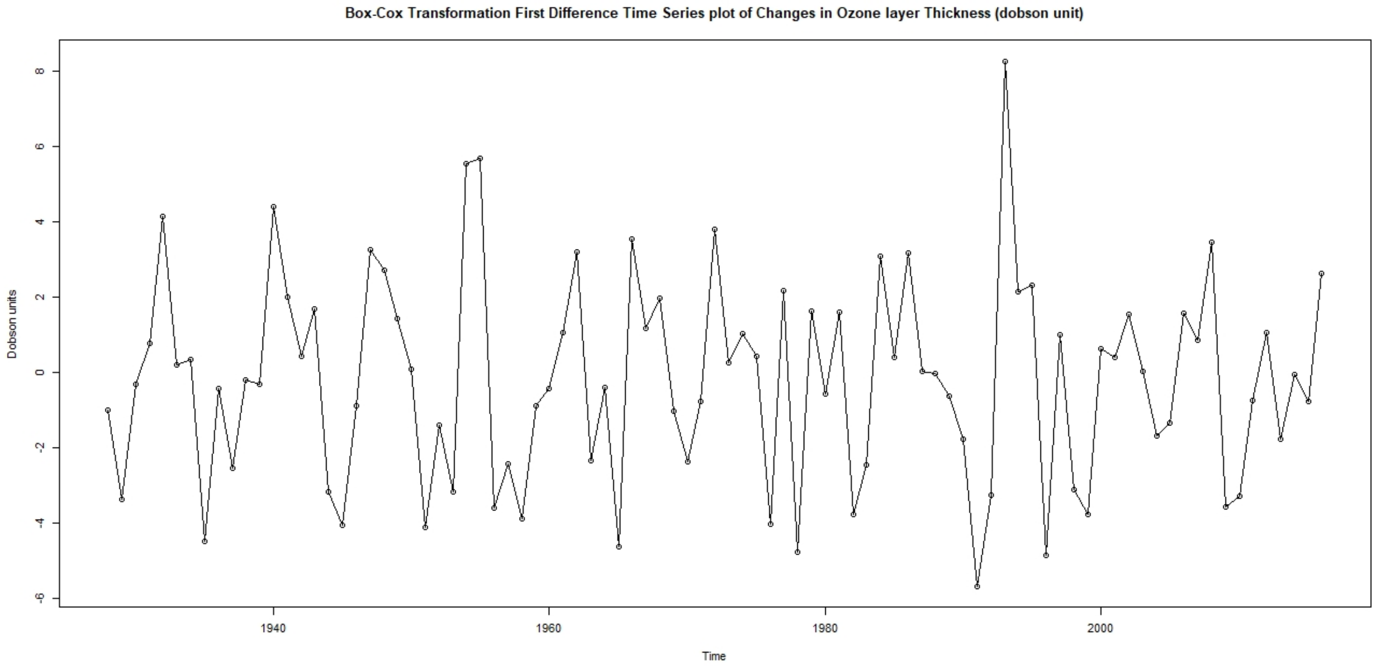


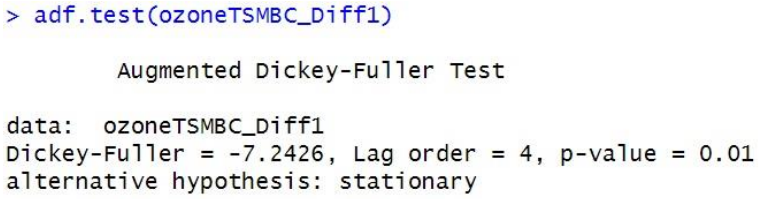
Figure 31

Figure 31 shows the 1st Difference Box-Cox Transformed Time Series. It depicts that, there is no trend in the series now and series is converted into stationary series as mean value is same across the plot.

Hence, we don’t need to do more differences as it will increase the complexity.

CHECKING STATIONARITY

Now, as we have detrend the series, we will again check for stationarity with help of adf test.



```
> adf.test(ozoneTSMBC_Diff1)

        Augmented Dickey-Fuller Test

data:  ozoneTSMBC_Diff1
Dickey-Fuller = -7.2426, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

Figure 32

P-value of adf test is 0.01 which is < 0.05. Means – we will reject null Hypothesis. Which means data does not have unit root and data follows stationarity.

DETERMINING SET OF POSSIBLE ARIMA(P,D,Q) MODELS

Now, we can determine the set of possible ARIMA(p,d,q) model. With help of different tools.

Here, p = order of AR,

d = number of difference

q = order of MA

ACF & PACF

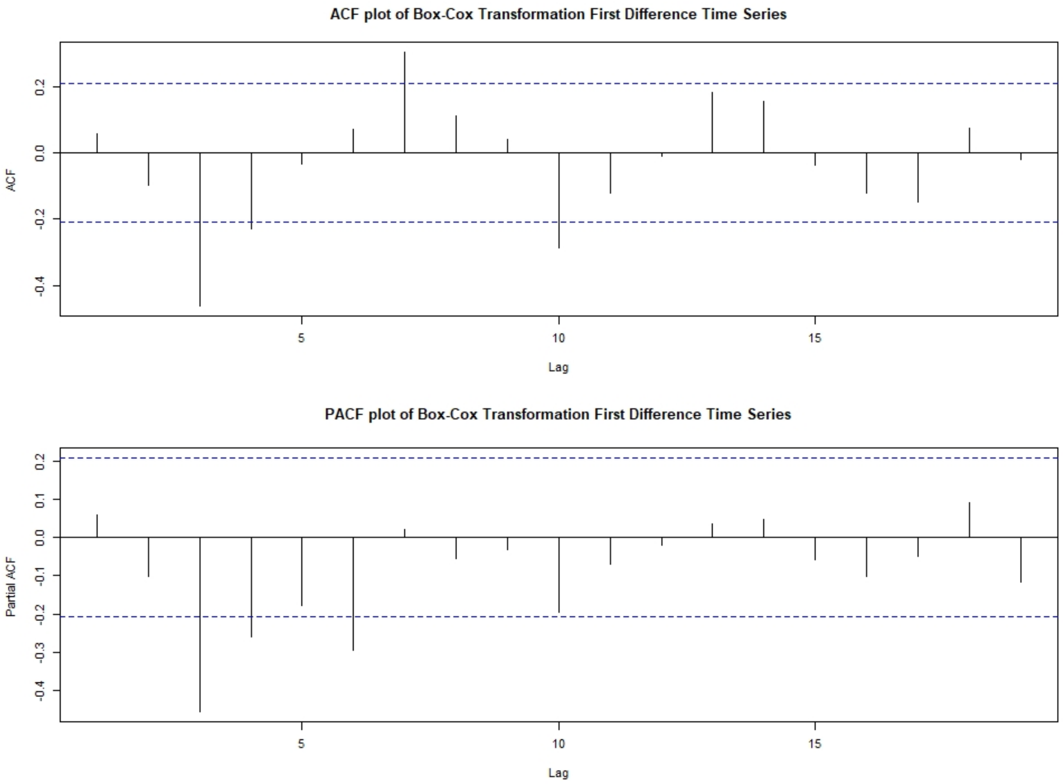


Figure 33

Figure 33 shows us ACF and PACF of 1st Difference BC transformed time series data.

From ACF, we can say, there is significant autocorrelation at lag 3, while there is slightly significant autocorrelation at lag 4. There is also significant AutoCorrelatoin at lag7. But, it is quite away from the origin. So, I will discard it.

So, q = 1, 2

From PACF, there is significant AutoCorrelaton at lag 3 and slightly significant AutoCorrelation at lag 4. There is also one significant autocorrelation at lag 6 and it is near to Autocorrelation at lag 4. It might be in the set of possible AR order, there is no harm to include it. Hence, I will include it.

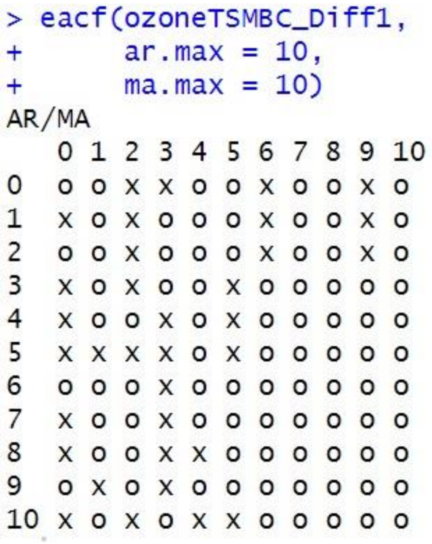
So, p = 1, 2, 3.

So, our set of ARIMA model form ACF and PACF are,

{ ARIMA(1,1,1), ARIMA(2,1,1), ARIMA(3,1,1),

ARIMA(1,1,2), ARIMA(2,1,2), ARIMA(3,1,2) }

EACF



```
> eacf(ozoneTSMBC_Diff1,
+      ar.max = 10,
+      ma.max = 10)
AR/MA
   0 1 2 3 4 5 6 7 8 9 10
0  o o x x o o x o o x o
1  x o x o o o x o o x o
2  o o x o o o x o o x o
3  x o x o o x o o o o o
4  x o o x o x o o o o o
5  x x x x o x o o o o o
6  o o o x o o o o o o o
7  x o o x o o o o o o o
8  x o o x x o o o o o o
9  o x o x o o o o o o o
10 x o x o x x o o o o o
```

Figure 34

Figure 34 shows us EACF of 1st Difference BC transformed time series data.

Here, we will look at the vertex of top-left 0 which is not disturbed by the x.

From Figure 34, we found vertex of top-left 0 at AR = 1 and MA = 3.

Hence, our ARIMA model will be ARIMA(1,1,3) and it's neighbour.

So, our Set of possible ARIMA models from EACF are,

{ ARIMA(1,1,3), ARIMA(1,1,4),
ARIMA(2,1,3), ARIMA(2,1,4) }

BIC TABLE

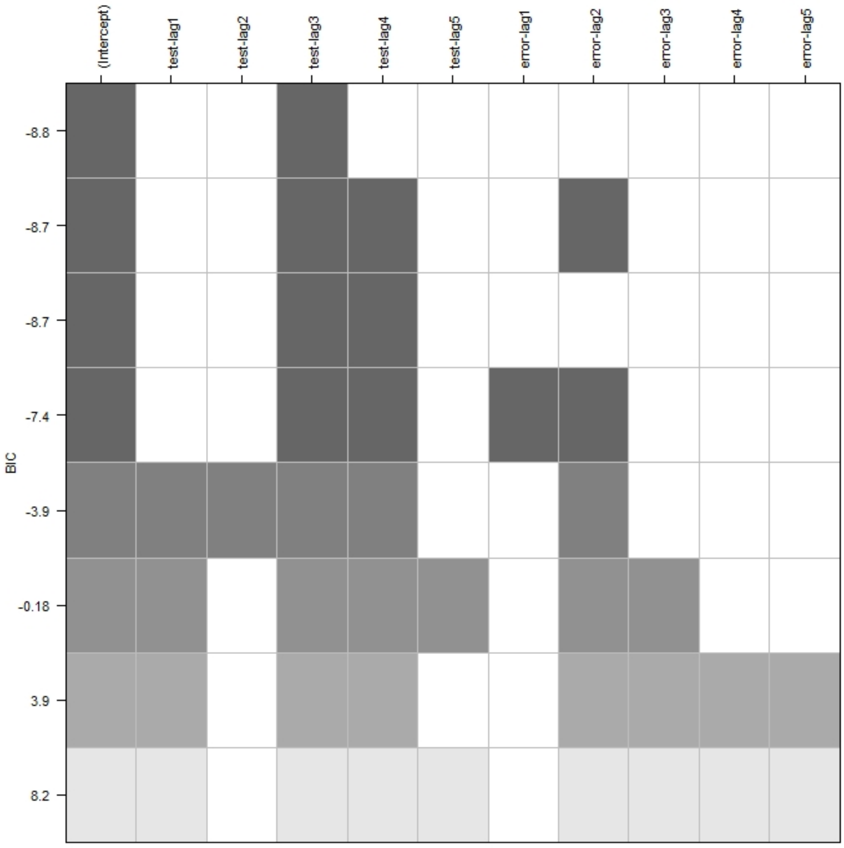


Figure 35

Figure 35 shows us BIC Table of 1st Difference BC transformed time series data.

From table, significant at testlag represents p and significant at errorlag represents q.

From table, testlag3 is obvious choice as it is significant for all the BIC model and we can also include testlag4 as it is significant in 2nd most model and other lower models.

So, p = 3, 4

While, from errorlag section, errorlag2 is the only choice as it is significant in every model apart from 1st and 3rd best model.

So, q = 2.

So out set of Possible model form BIC table are

{ ARIMA(3,1,2), ARIMA(4,1,2) }

CONCLUSION

In nutshell, time-series data has trend and changing variance. It is non-stationary time-series. So, I converted it into Stationary Time series by applying Box-Cox transformation to the time series and then doing 1st difference on the Box-Cox transformed time-series. As order is important.

Then I find set of possible ARIMA(p,d,q) model with help of ACF, PACF, EACF and BIC Table tool.

So, my final set of possible ARIMA(p,d,q) models are,
{ ARIMA(1,1,1), ARIMA(2,1,1), ARIMA(3,1,1),
ARIMA(1,1,2), ARIMA(2,1,2), ARIMA(3,1,2),
ARIMA(1,1,3), ARIMA(1,1,4), ARIMA(2,1,3),
ARIMA(2,1,4), ARIMA(4,1,2) }

If there is seasonality, SARIMA model is more useful.

REFERENCES

Class Presentations (1-5)

Class Lecture - Notes (1-5)

Class Recordings (1-5)

APPENDIX

ABBREVIATION

ARIMA = Auto-Regressive Integrated Moving Average

AR = Auto-Regressive,

MA = Moving Average

BC = Box-Cox

adf = Augmented Dicky-Fuller Test

pp = Phillops-Perron Test

CI = Condifence Interval

AIC = Akaike Information Criterion

BIC = Bayesian Information Criterion

ACF = AutoCorrelation Function

PACF = Partial AutoCorrelation Function

EACF = Extended AutoCorrelation Function

BIC Table = Bayesian Information Criterion

TS = Time-Series

CODE

```
# Clearing the environment
rm(list=ls())

# importing the necessary libraries
library(readr)
library(TSA)
library(tseries)

# reading the data
ozone <- read.csv("D:\\Study_Material\\SEM_3\\Time Series Analysis\\Assignment
1\\data1.csv",header=FALSE)
head(ozone)

# class of ozone
class(ozone)


# converting dataframe into time series
ozoneTS = ts(ozone$V1, start = 1927)
head(ozoneTS)

# checking the class of ozoneTS
class(ozoneTS)

#-----------------------------------------------------------------------
#------------------------------ Functions ------------------------------
#-----------------------------------------------------------------------
```

```
# Functions for Time Series
plot_ts <- function(ts, transformation)
{
  win.graph(width = 20,
            height = 10,
            pointsize = 8)
  plot(ts,
       ylab = "Dobson units",
       main = c(paste0(toString(transformation),
                       " plot of Changes in Ozone layer Thickness (dobson unit)")),
       type="o")
}


# Function for ACF and PACF
autoCorrelation <- function(ts, time_series)
{
  win.graph(width = 20,
            height = 15,
            pointsize = 8)

  par(mfrow=c(2,1))
  acf(ts,
      main = c(paste0("ACF plot of ",toString(time_series))))

  pacf(ts,
       main = c(paste0("PACF plot of ",toString(time_series))))
  par(mfrow=c(1,1))
}


# Functions for Model and its Residual Analysis
model <- function(ts, model, modelname)
{
  win.graph(width = 20,
            height = 15,
            pointsize = 8)
  par(mfrow=c(1,1))
  plot(ts,
       ylim = c(min(c(fitted(model), as.vector(ts))),
                max(c(fitted(model),as.vector(ts)))),
       ylab='Change in Dobson units',
       main = c(paste0("Time Series Plot of change in Ozone Layer Thickniess (Yearly)
and ",
                       toString(modelname)," fitted")),
       type='o')

  lines(ts(fitted(model),
           start = 1927,
           end = 2016,
           frequency = 1),
        col="red",
        type="l")
  res.model = rstudent(model)
  win.graph(width = 20,
            height = 15,
            pointsize=8)

  par(mfrow=c(3,2))

  plot(y = res.model,
       x = as.vector(time(ozoneTS)),
       main = c(paste0("Time Series plot of Standardised residuals of fitted ",
                       toString(modelname))),
       xlab = 'Time',
       ylab = 'Standardized Residuals',
       type = 'o')

  hist(res.model,
       main = c(paste0("Histogram of Standardised residuals of fitted ",
                       toString(modelname))),
       xlab='Standardized Residuals')

  qqnorm(y=res.model,
         main = c(paste0("QQ plot of standardised residuals of fitted ",
                         toString(modelname))))

  qqline(y=res.model,
         col = 2,
         lwd = 1,
         lty = 2)

  shapiro.test(res.model)

  acf(res.model,
      main = c(paste0("ACF plot of standardised residuals of fitted ",
                      toString(modelname))))

  pacf(res.model,
       main = c(paste0("PACF plot of standardised residuals of fitted ",
                       toString(modelname))))
  par(mfrow=c(1,1))
}

#----------------------------------------------------------------------
#------------------------ Descriptive Analysis ------------------------
#----------------------------------------------------------------------


# Original Time Series
plot_ts(ozoneTS, "Time Series")

# ACF and PACF of original Time Series
autoCorrelation(ozoneTS, "Time Series")

win.graph(width = 20,
          height = 15,
          pointsize=8)
```

```
# Scatter Plot of Consecutive Years
plot(y=ozoneTS,
     x=zlag(ozoneTS),
     ylab='Dobson Units',
     xlab='Previous Year Dobson Units',
     main = "Scatter plot of Change in Ozone Layer Thickness in consequtive years.")

# Correlation of Consecutive Years
y = ozoneTS
x = zlag(ozoneTS)
index = 2:length(x)
cor(y[index],x[index])

#----------------------------------------------------------------------
#------------------------------- Task 1 -------------------------------
#----------------------------------------------------------------------



#----------------------------
#-----Linear trend model-----
#----------------------------

model1 = lm(ozoneTS~time(ozoneTS))
summary(model1)

model(ozoneTS, model1, "Linear Model")

#-------------------------------
#-----Quadratic trend model-----
#-------------------------------

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

model(ozoneTS, model2, "Quadratic Model")

#-------------------------------
#-----Seasonal trend model------
#-------------------------------

ozoneTS_S <- ts(ozone, start=1927, frequency = 7)

season.=season(ozoneTS_S)
model3=lm(ozoneTS_S~season. -1)
summary(model3)

model(ozoneTS, model3, "Seasonal Model")

#-------------------------------
#-----harmonic trend model------
#-------------------------------
```

```
har. <- harmonic(ozoneTS_S, 0.4)
data <- data.frame(ozoneTS_S, har.)
model4 <- lm(ozoneTS_S ~ cos.2.pi.t. + sin.2.pi.t. , data = data)
summary(model4)

model(ozoneTS, model4, "Cosine Model")

#--------------------------------
#------ Finding Best Model ------
#--------------------------------

AIC(model1,model2,model3,model4)
BIC(model1,model2,model3,model4)

#-------------------------------
#-------- Prediction -----------
#-------------------------------

h <- 5
t <- time(ozoneTS)
t2 <- t^2

firstYear <- t[1]
lastYear <- t[length(t)]

newYear <- seq(lastYear+1, lastYear+h, 1)

newYearData <- data.frame(t = newYear,
                         t2 = newYear^2)

forecasting <- predict(model2, newdata = newYearData, interval = "prediction")

combine <- c(ozoneTS,forecasting[,1])

win.graph(width = 15,
          height = 10,
          pointsize=8)

plot(ozoneTS,
     xlim= c(1927, 2016+1+h),
     ylim = c(min(combine)-5, max(combine)+5),
     ylab = "Change in Dobson units",
     main = "Forecast of quadratic model on Time series of Ozone layer thickness")

lines(ts(as.vector(forecasting[,3]),
         start = c(2017), frequency = 1),
      col="blue")

lines(ts(as.vector(forecasting[,1]),
         start = c(2017),
         frequency = 1),
      col="red")

lines(ts(fitted(model2),
         start = c(1927),
         frequency = 1),
      col="green")

lines(ts(as.vector(forecasting[,2]),
         start = c(2017),
         frequency = 1),
      col="blue")

legend("topleft",
       lty=1,
       pch=1,
       col=c("black","blue","red","green"),
       text.width = 18,
       c("Change in Ozone Layer Thickness","5% forecast limits", "Forecasts","Fitted
model"))

#-----------------------------------------------------------------------
#-------------------------------- Task 2 -------------------------------
#-----------------------------------------------------------------------

# Check the stationarity
adf.test(ozoneTS)

# Moving the Time-Series for positive values
ozoneTSM <- ozoneTS + abs(min(ozoneTS))+1

#-------------------------------------
#---------Log Transformation---------
#-------------------------------------

ozoneTSMLog <- log(ozoneTSM)
plot_ts(ozoneTSMLog, "Log Transformation Time Series")

#-------------------------------------
#------Box-cox Transformation--------
#-------------------------------------
BC <- BoxCox.ar(ozoneTSM)

# box-cox confidence Interval
BC$ci

# maximum likelihood
lambda <- BC$lambda[which(max(BC$loglike) == BC$loglike)]
lambda

# box-cox transformation
ozoneTSMBC = ((ozoneTSM^lambda)-1)/lambda

plot_ts(ozoneTSMBC, "Box-Cox Transformation Time Series")

#-------------------------------------
#-------first differencing-----------
#-------------------------------------
```

```
ozoneTSMBC_Diff1 <- diff(ozoneTSMBC,
                         differences = 1)

plot_ts(ozoneTSMBC_Diff1, "Box-Cox Transformation First Difference Time Series")

# Check the stationarity
adf.test(ozoneTSMBC_Diff1)

#-------------------------------------
#-----------Model Parameters---------
#-------------------------------------

# ACF and PACF
autoCorrelation(ozoneTSMBC_Diff1, "Box-Cox Transformation First Difference Time 
Series")
# ARMIA(1,1,1), ARMIA(2,1,1), ARMIA(3,1,1)
# ARMIA(1,1,2), ARMIA(2,1,2), ARMIA(3,1,2)

# EACF
eacf(ozoneTSMBC_Diff1,
     ar.max = 10,
     ma.max = 10)
# ARIMA(1,1,3), ARIMA(2,1,3),
# ARIMA(1,1,4), ARIMA(2,1,4)

# BIC
win.graph(width = 10,
          height = 10,
          pointsize = 8)

par(mfrow=c(1,1))
res = armasubsets(y=ozoneTSMBC_Diff1,
                  nar=5,
                  nma=5,
                  y.name='test',
                  ar.method='ols')
plot(res)
# ARIMA(3,1,2), ARIMA(4,1,2)

# Possible set of Models
# ARMIA(1,1,1), ARMIA(2,1,1), ARMIA(3,1,1)
# ARMIA(1,1,2), ARMIA(2,1,2), ARMIA(3,1,2)
# ARIMA(1,1,3), ARIMA(2,1,3),
# ARIMA(1,1,4), ARIMA(2,1,4)
# ARIMA(4,1,2)
```