MATH1318 - Time Series Analysis - Assignment 1

Samuel Holt

March 15, 2019

Contents

Introduction

The dataset provided is of the yearly changes in thickness of the Ozone layer from 1927 to 2016 in Dobson units. A single data point per year is the percentage change in Dobson units compared to the previous year.

In this report I will be investigating if any deterministic trends exist in the time series data. This will include a linear model, a quadratic model, and a hybrid quadratic harmonic model. Each model will be evaluated in it's p-values for the variables involved, the R squared value, and also its residuals. For the residuals of each respective model, normality will be investigated using histograms, quantile-quantile plots and validated using the Shaprio-Wilks test. The autocorrelations of each lag of the residuals will also visualised using Auto Correlation Function plots, to investigate significant autocorrelative trends in the residuals of the model. Finally, I will forecast the following 5 years with each model, but propose the most statistically valid model to finalise forecast predictions for this report.

Load appropriate packages.

```
library(TSA)
library(dplyr)
```

Import and Convert Data

Load data, no headers present in csv file, hence set header argument to FALSE.

```
dataset <- read.csv('data1.csv', header = FALSE)
```

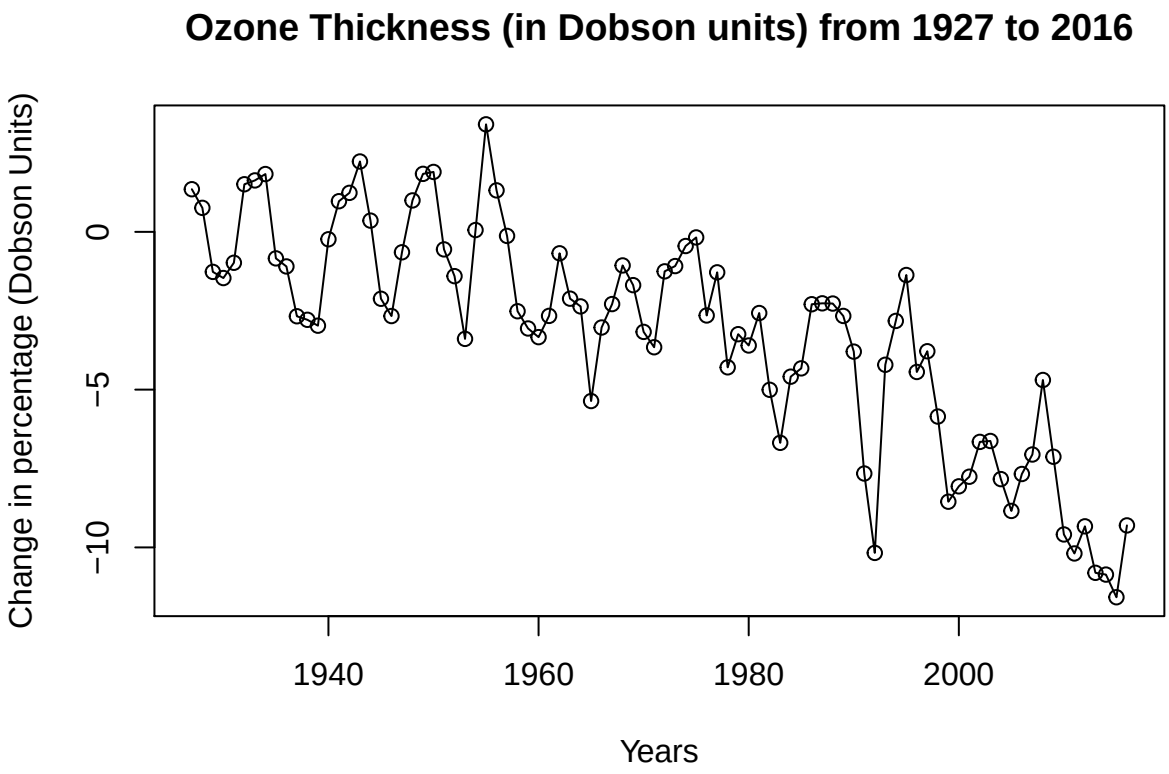
Convert dataframe to a time series object.

```
data_ts <- ts(dataset,
              start = 1927,
              end = 2016,
              frequency = 1)
```

Visualise Data

Visualise the time series data in order to gauge what models could be tested.

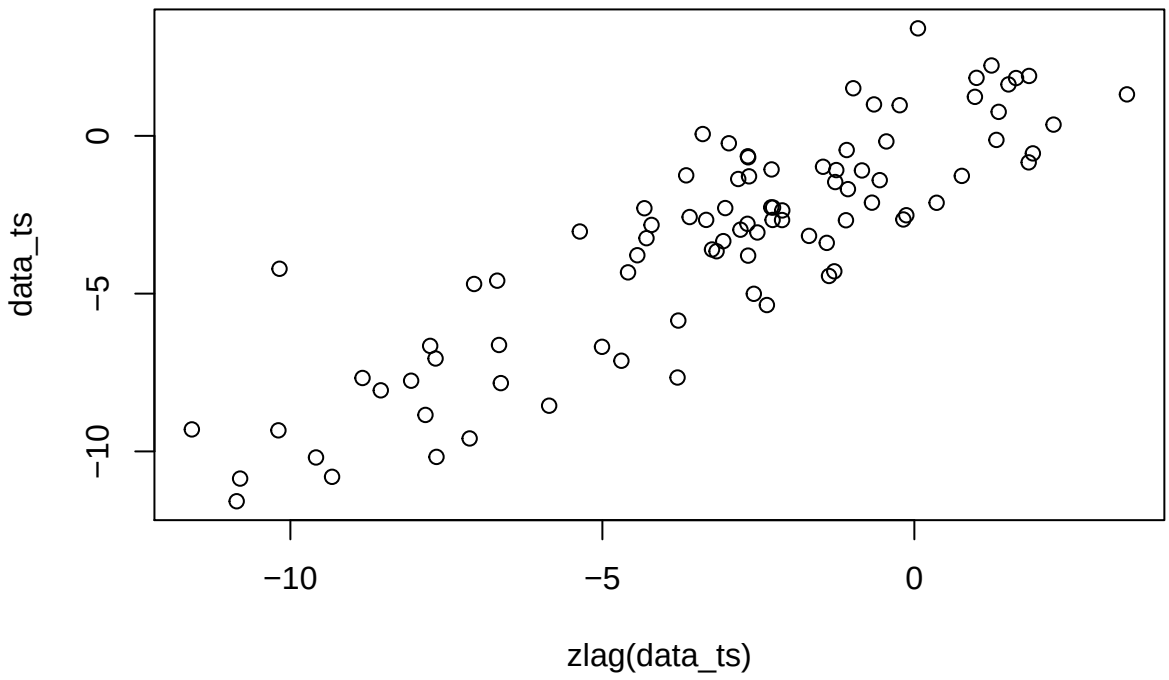
```
plot(
  data_ts,
  xlab = 'Years',
  ylab = 'Change in percentage (Dobson Units)',
  type = 'o',
  main = 'Ozone Thickness (in Dobson units) from 1927 to 2016'
)
```



Demonstrated by the scatterplot, there is clearly a a downward trend.

Let's investigate if the first lag holds any visual clues to a valid correlation.

```
plot(y = data_ts,
     x = zlag(data_ts))
```



A notable linear correlation is present with the given scatterplot, a single obvious outlier also present, but this shouldn't cause any deviations.

Here I will determine the autocorrelation coefficient to see if there is any significant correlations with the original time series data.

```
x <- data_ts
y <- zlag(data_ts)
index <- 2:length(x)
cor(x[index],
    y[index])
```

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

An autocorrelation coefficient of 0.87 shows a significant correlation between the time series data and its first lag.

Linear Model

Fitting the Linear Model

Given the evident trend in the original time series scatterplot, a linear model will be fit and investigated for statistical significance.

```r
# fit a linear regression model
t <- time(data_ts)
lin_model <- lm(data_ts ~ t)
summary(lin_model)
```

```
##
## Call:
## lm(formula = data_ts ~ t)
##
## 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 ***
## t            -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
```

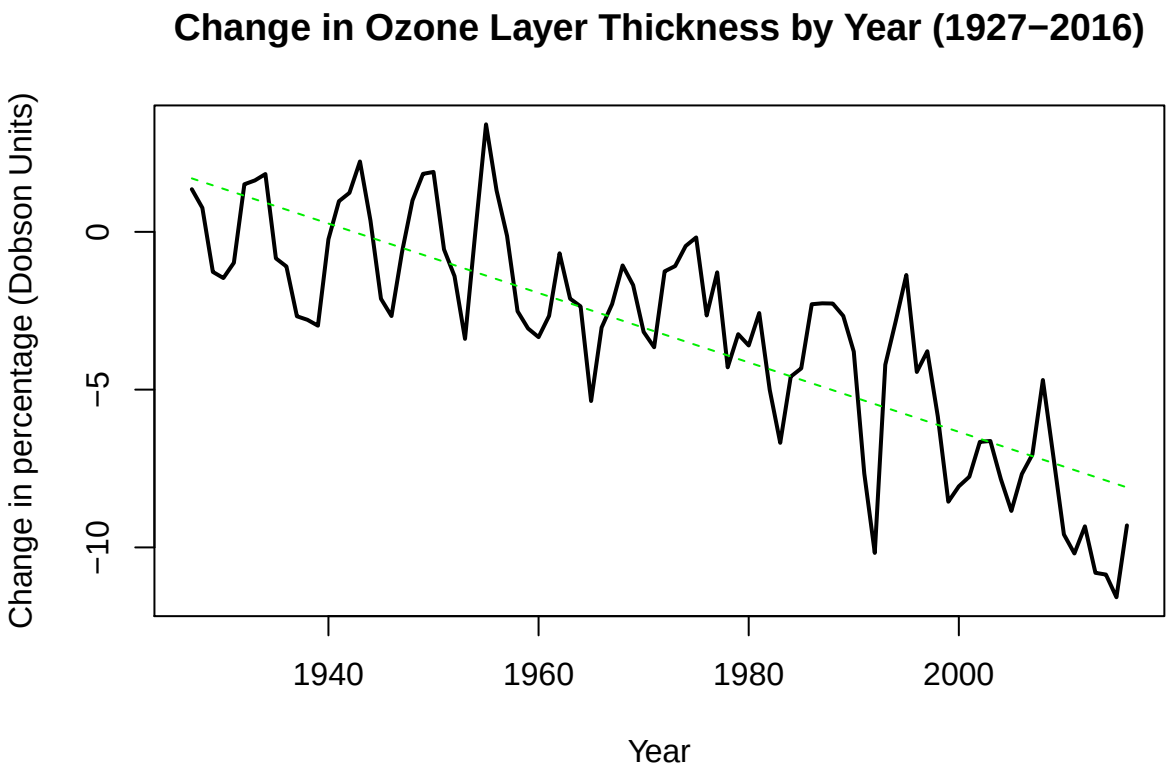
```r
# R^2: 0.67
```

Both coefficients, the intercept and time (t), have been shown to have statistical significance in regard to the dependent variable, Dobson Units by year. Also note an R squared value of 0.6693, which isn't deemed as high, but certainly not insignificant.

Visualise Linear Model

Here we will plot the linear model against the time series and then overlay a prediction of the following 5 years with the model.

```r
# Visualise data with the linear model overlay
plot(
  data_ts,
  lwd = 2,
  xlab = 'Year',
  ylab = 'Change in percentage (Dobson Units)',
  main = 'Change in Ozone Layer Thickness by Year (1927-2016)'
)
lines(1927:2016, lin_model$fitted.values, col = 'green2', lty = 2)
```

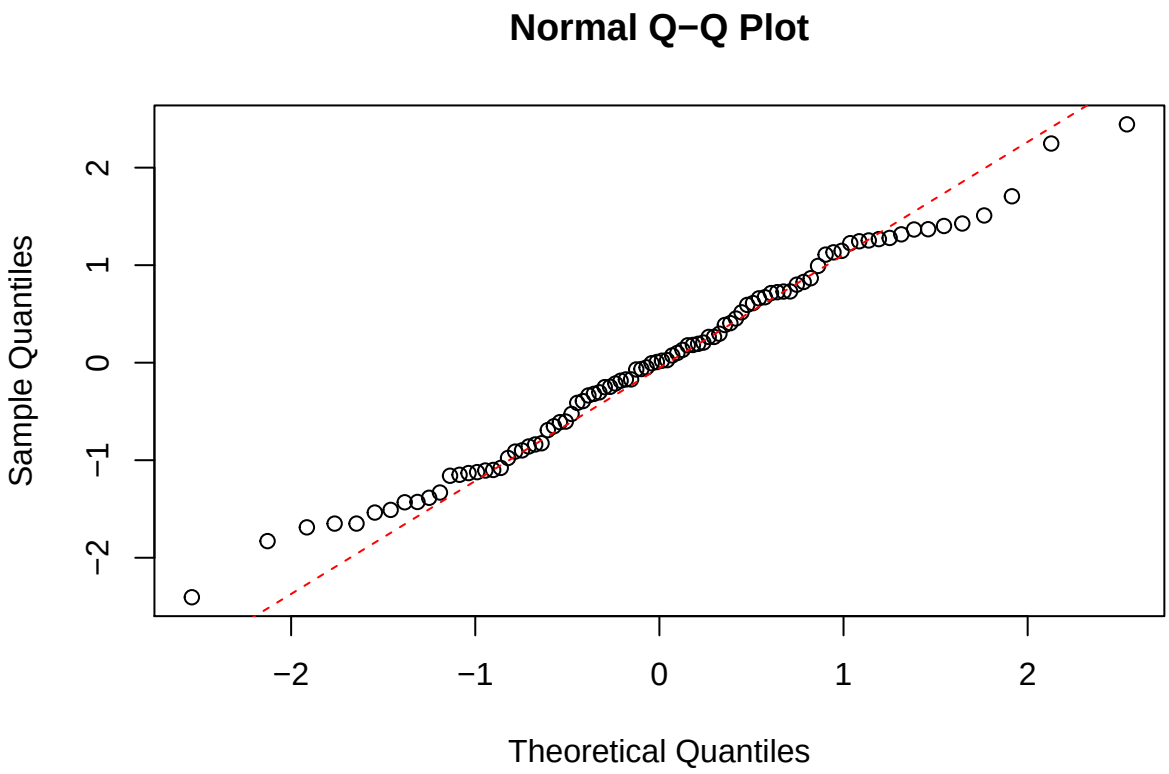


Although the linear model seems to fit visually, it would be unwise to take it for face value.

Normality of Standardised Residuals for Linear Model

Here I will investigate the residuals of the linear model. First with the normality of residuals with a Q-Q plot.

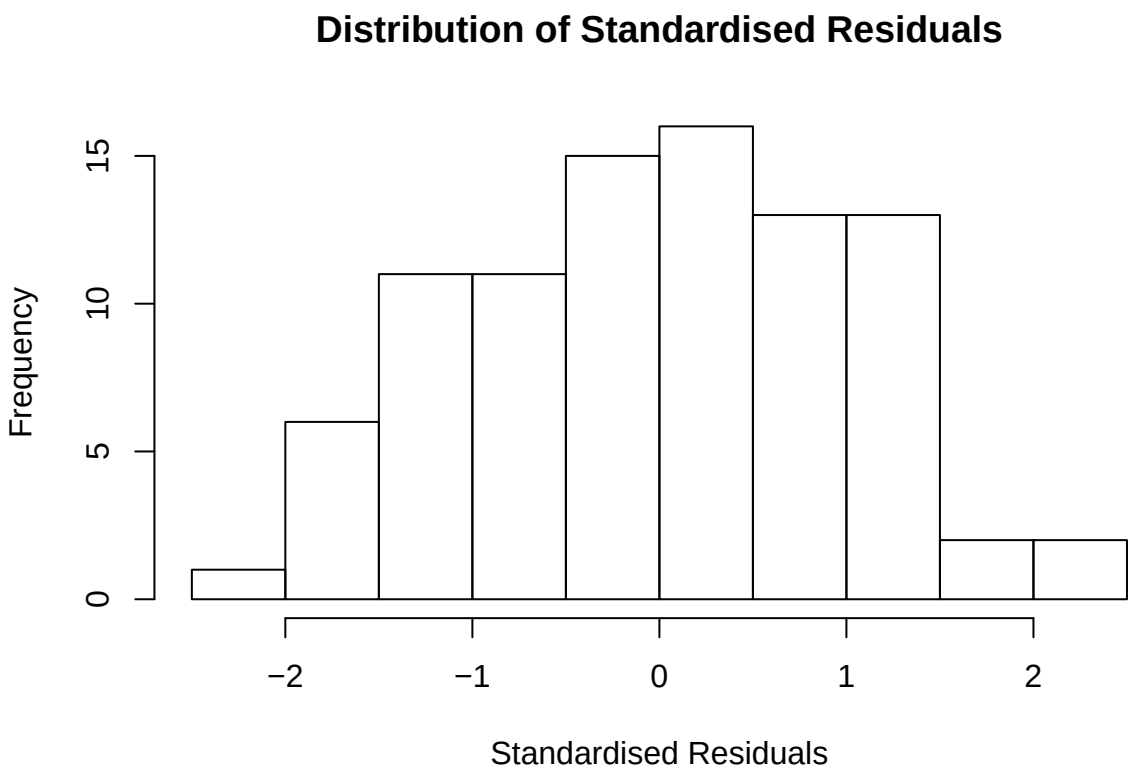
```
rlin <- rstudent(lin_model)

qqnorm(rlin)
qqline(rlin, col = 'red', lty = 2)
```



There is evidence of some deviation from the quantile-quantile line at the ends of the quantiles. This interferes with any confidence in stating the standardised residuals are completely normal. So we shall refine our investigation.

A histogram will further demonstrate possible normality of residuals.

```
hist(rlin,
     breaks = 10,
     xlab = 'Standardised Residuals',
     main = 'Distribution of Standardised Residuals')
```



The histogram here shows further evidence of normality of residuals.

A Shapiro-Wilks test will finalise the significance of normality present in the standardised residuals of the fitted linear model.

```
shapiro.test(rlin)
```

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

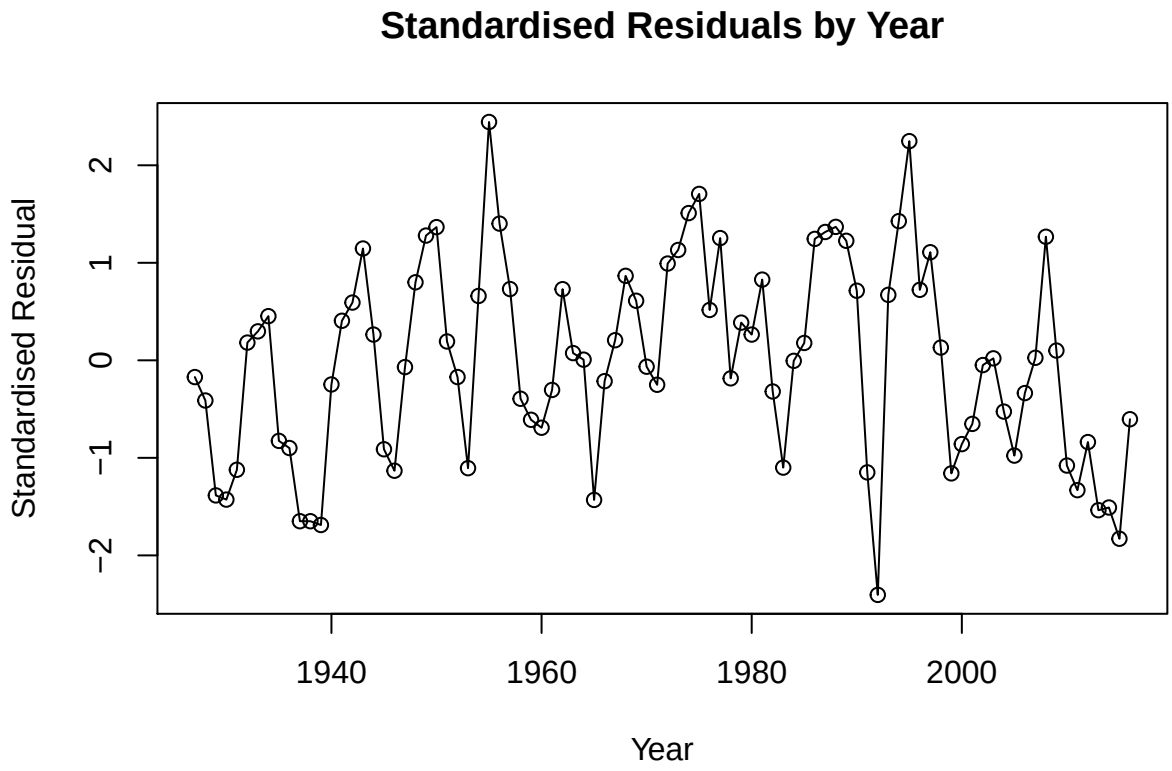
With a p-value > 0.05, the presumed significance level, we reject the null hypothesis that is standardised residuals are not normally distributed and accept that the residuals are normally distributed.

Visualise Standardised Residuals over Time

Now that normality is deemed to be significant in the residuals, we can look at their behavior over time.

```
# residuals
plot(y = rlin,
     x = as.vector(time(data_ts)),
     type = 'o',
     xlab = 'Year',
```

```
     ylab = 'Standardised Residual',
     main = 'Standardised Residuals by Year')
```

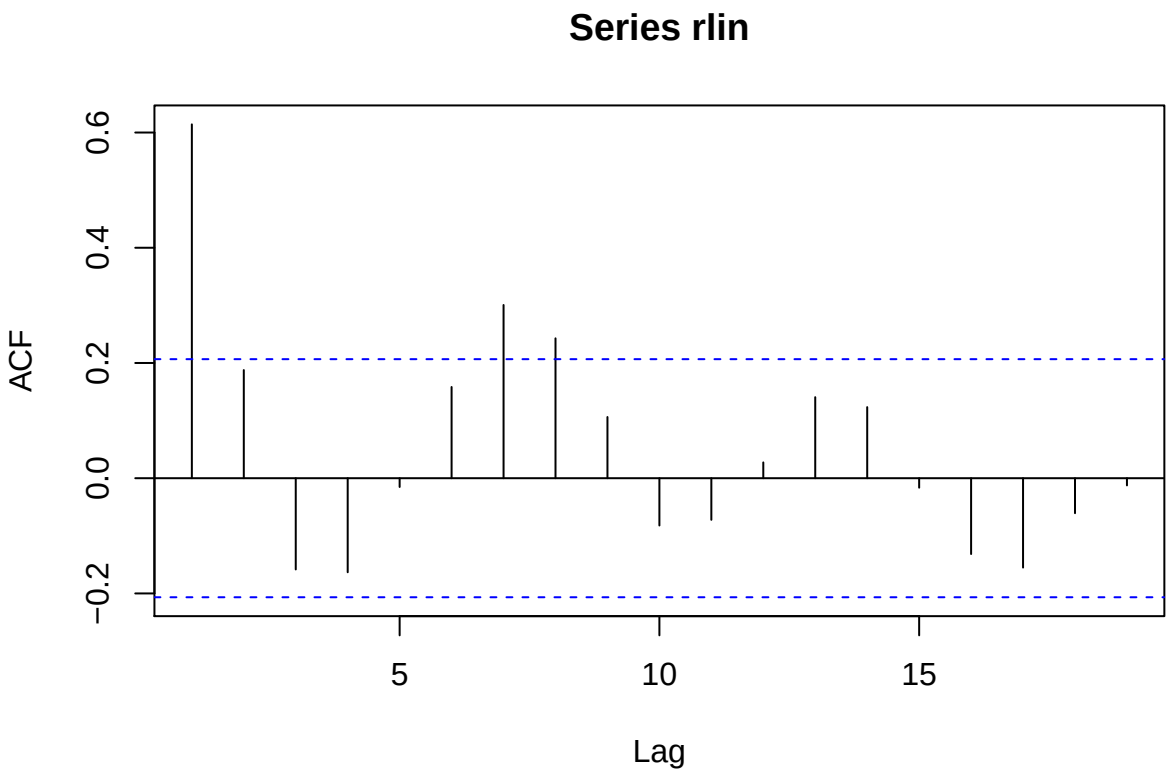


The graph shows some major deviations, but not completely adherent to white noise. There also seems to be a slight polynomial trend present in the residuals, something not being captured by the linear model.

Auto Correlation Function of Linear Model Residuals

The autocorrelation function will show any significant correlations at sequential lags.

```
acf(rlin)
```



There are three (3) intercepts showing statistically significant autocorrelations at lags 1, 6, and 7. This means that the model is definitely capturing all aspects of the time series process with a slight harmonic trend being present in the ACF. The first lag is the most significant at ~0.6, we shall take note of that value for further models.

We conclude that a linear model is not capturing multiple facets of the time series, and thus is not a suitable model for forecasting predictions.

Quadratic Model

Fitting Quadratic Model

Due to the residual plot over time demonstrating a polynomial characteristic in the residuals, I will attempt to fit and evaluate a quadratic model.

```
# fit a quadratic regression model
t <- time(data_ts)
t2 <- t ^ 2
quad_model <- lm(data_ts ~ t + t2)
summary(quad_model) # all stat sig
```

```
##
## Call:
## lm(formula = data_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
```

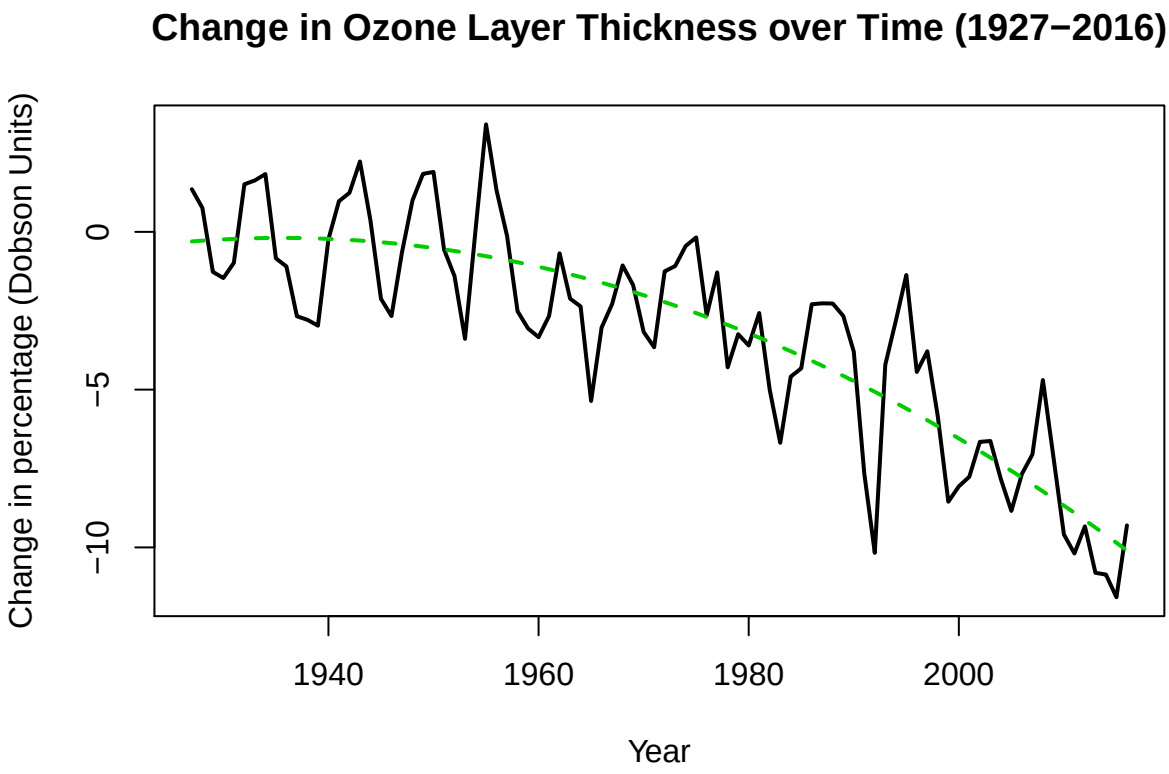
```
# R^2: 0.74 higher than a linear model
```

Here we can see an R squared value of 0.7391, meaning 73.91% of variation in the data is explained by this model. This is an improvement upon the linear model which had an R squared value of 0.6693. All deemed coefficients statistically significant with p values < 0.001.

Visualise Quadratic Model

Given the improvement upon the R squared valued, we will visualise the fit of the quadratic model over the original time series data.

```
# Visualise time series and quadratic model
plot(
  data_ts,
  lwd = 2,
  xlab = 'Year',
  ylab = 'Change in percentage (Dobson Units)',
  main = 'Change in Ozone Layer Thickness over Time (1927-2016)'
)
# Overlay quadratic model
years <- seq(1927, 2016, 1)
lines(years, quad_model$fitted.values, col = 'green3', lty = 2, lwd = 2)
```

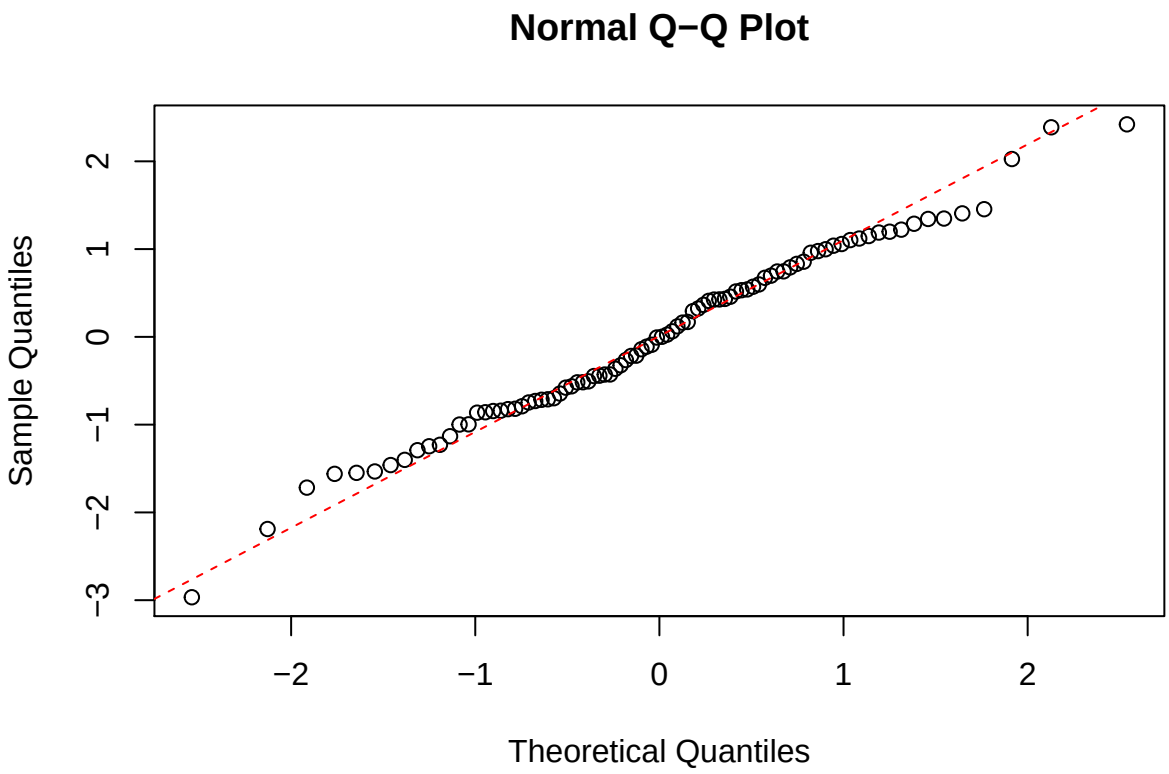


The plot seems to capture the quadratic trend, but of course we will further investigate the residuals of the model to dictate the significance of the fitted model.

Normality of Standardised Residuals of Quadratic Model

First we shall inspect the normality of the standardised residuals with a Q-Q plot.

```
# validate model (all p-values below significance level of 0.05, but check assumptions)
rquad <- rstudent(quad_model) # grab standardised residuals
qqnorm(rquad) # create quantile quantile plot
qqline(rquad, col = 'red', lty = 2)
```

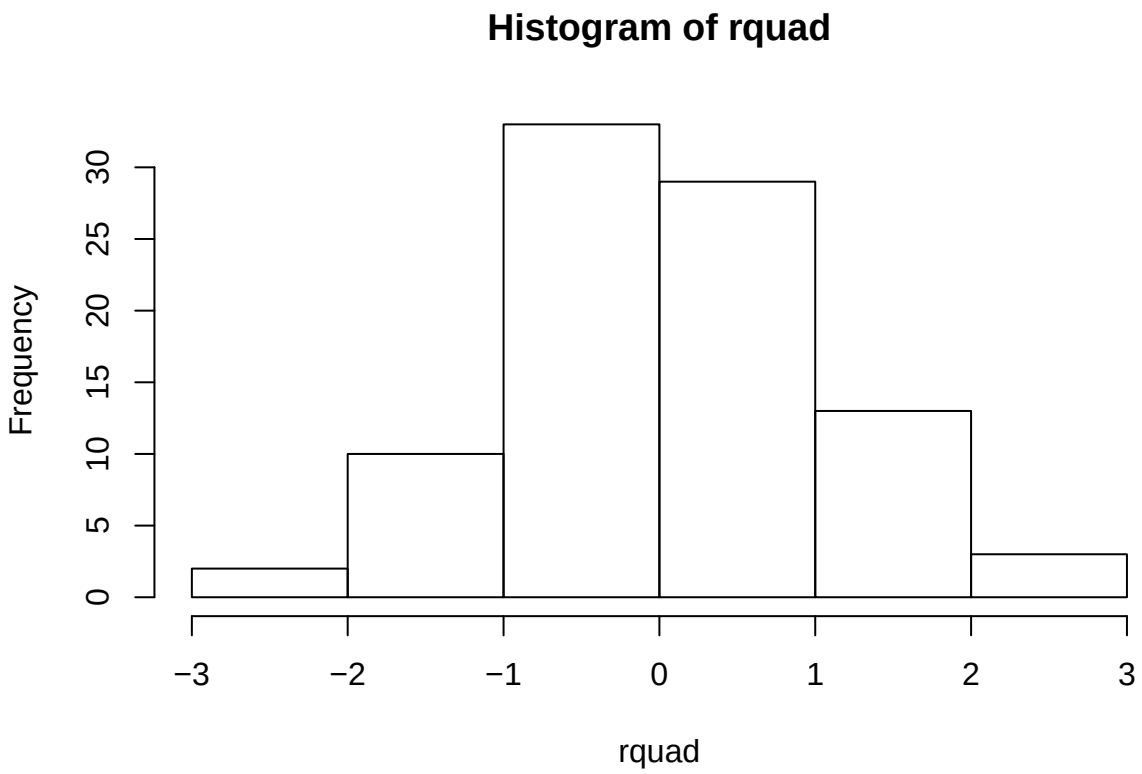



```
# some deviation from the quantile-quantile line at the ends of the distrubition
```

The standardised residuals of the quadratic model tend to deviate at the end quantiles, but to a lesser degree than the linear model.

We can further investigate the distribution of the residuals with a histogram.

```
hist(rquad, breaks = 6) # plot histogram
```



The distribution of the standardised residuals of the quadratic model seems to follow a normal distribution. A Shapiro-Wilks test will statistically validate that claim.

```
shapiro.test(rquad) # fail to reject, assume normality, closer to 1 than lin model
```

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

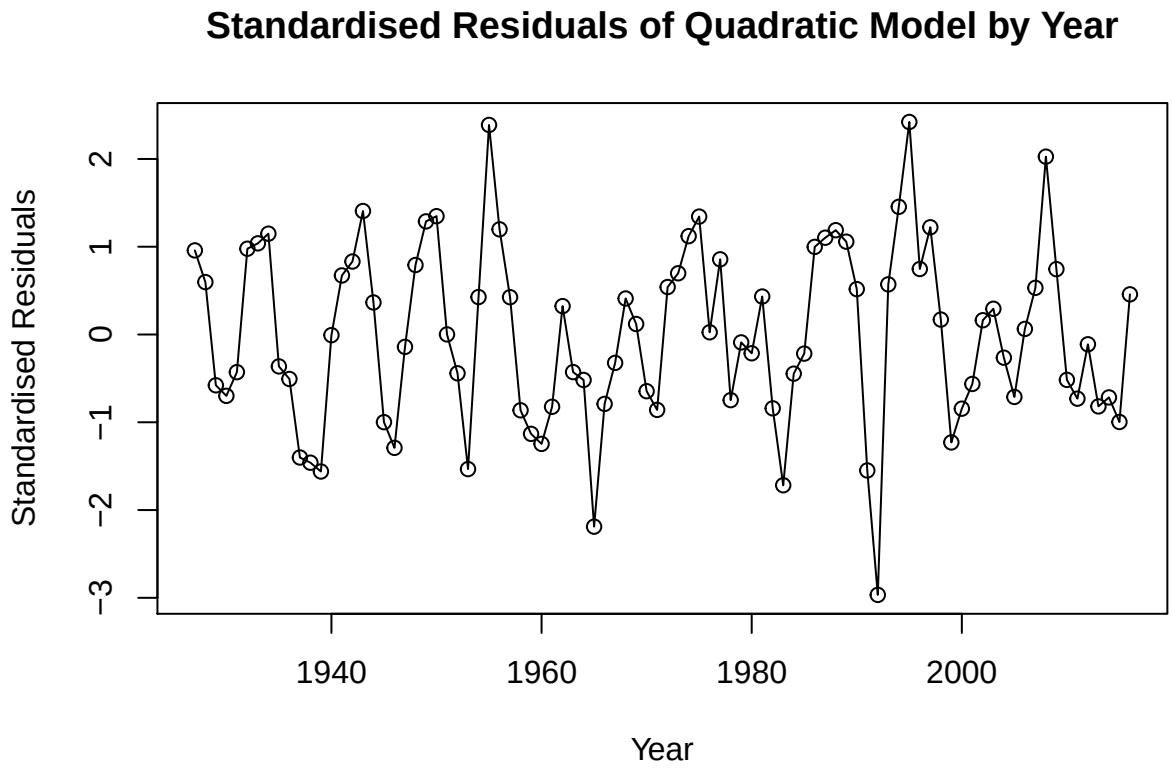
The Shapiro-Wilks test shows a p-value of 0.6493, this is larger than the linear model's Shapiro-Wilks test p-value, and also exceeds the significance value of 0.05, hence we reject the null hypothesis and accept normality of the standardised residuals.

Visualise Standardised Residuals over Time

Now that normality has been demonstrated in the standardised residuals, we move on to inspecting for trends in the standardised residuals with a standard plot.

```
# residuals
plot(y = rquad,
     x = as.vector(time(data_ts)),
     type = 'o',
     xlab = 'Year',
```

```
     ylab = 'Standardised Residuals',
     main = 'Standardised Residuals of Quadratic Model by Year')
```



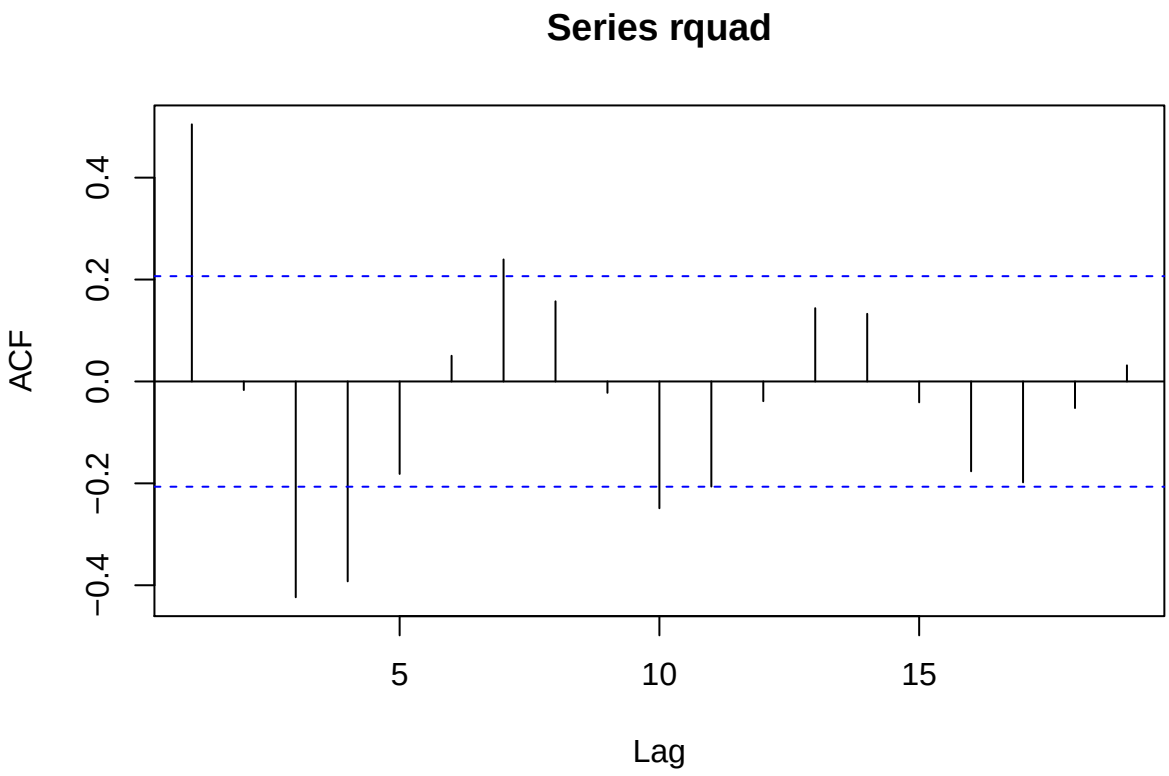

```
# some major deviations, not white noise
```

The plot demonstrates some slight changes in variance, circa 1960 to 1985 has a lesser variance compared to the decades either side, although the average seems to be stationary we cannot state this has stationary behavior due to the changes in variance. A noteworthy issue with this plot is that a cyclical/seasonal trend is present in the residuals, suggesting that a harmonic parameter is not being captured by the quadratic model.

Auto Correlation Function of Quadratic Model Residuals

We can further investigate the standardised residuals with an autocorrelation function to show significant autocorrelations at sequential lags.

```
acf(rquad)
```



The ACF clearly demonstrates intercepts of the thresholds and also shows a pattern indicating the quadratic model does not capture all behavioral aspects of the time series data. The first lag shows an autocorrelation value of ~0.5, which is an again an improvement upon the linear model which held an autocorrelation at the first lag of ~0.6.

Due to the patterns in the residual plots and ACF, I will create a harmonic and quadratic model and test its parameters against the quadratic model, if improves upon the new model can be shown then I will forecast the following 5 years with the new harmonic quadratic model.

Harmonic Quadratic Model

Finding Optimal Frequency for Time Series Object

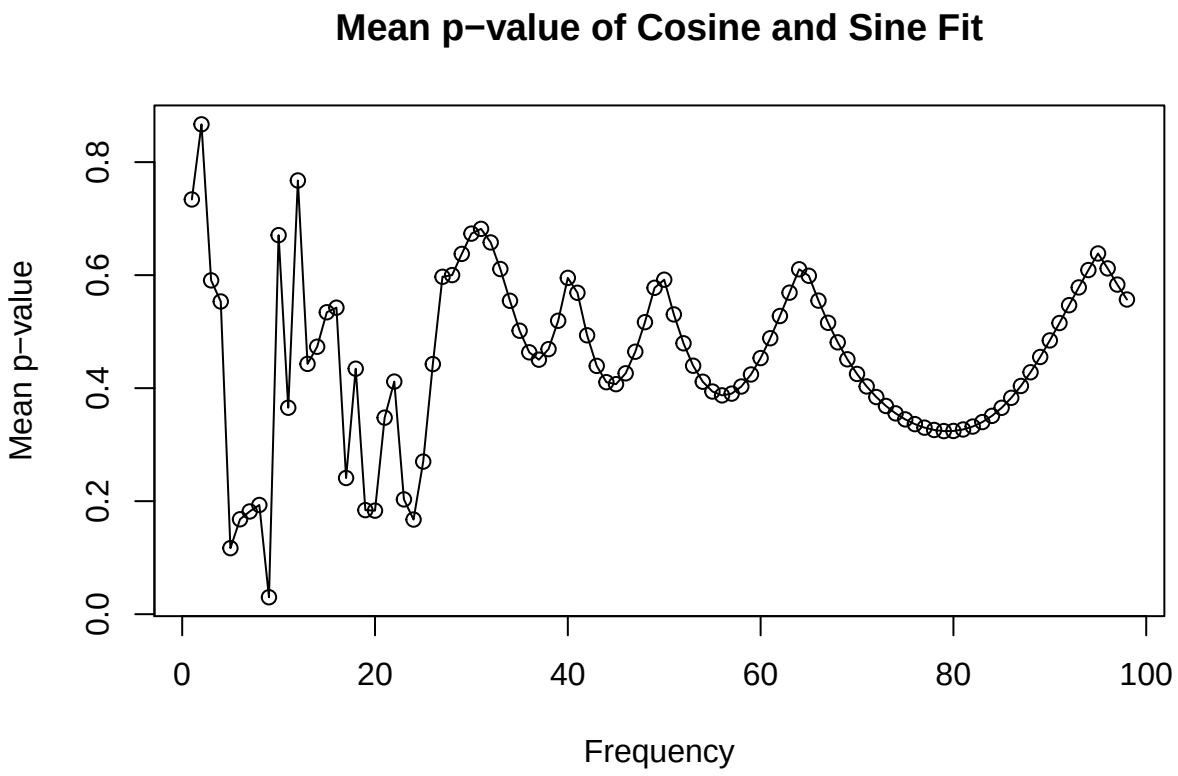
In order to create a harmonic based model, I need to look for the most suitable frequency value of the time series data. Here I wrote a loop to determine the most statistically significant frequency values, as we need to reconvert the original data to a time series with the frequency argument.

```
# Create empty vector to fill with loop results for Cosine waves
mod_sum_vec <- vector()
# Create loop
for (i in 3:100) {
  data_freq <- ts(dataset, frequency = i)
  har <- harmonic(data_freq, 1)
  t <- time(data_freq)
  t2 <- t^2
```

```
  # Fit quadratic harmonic model with only the sine parameter
  quad_har_model <- lm(data_freq ~ t + t2 + har[,1])
  # Grab the pvalue for each iteration
  mod_sum_vec[i] <- summary(quad_har_model)$coefficients[,4][4]
}

# Create dataframe with frequency column matched to p-values
library(dplyr)
df <- data_frame(freq = 1:length(mod_sum_vec),
                 pval_cos = mod_sum_vec)
```

```
# Create empty vector to fill with loop results for Sine waves
mod_sum_vec <- vector()
# Create loop
for (i in 3:100) {
  data_freq <- ts(dataset, frequency = i)
  har <- harmonic(data_freq, 1)
  t <- time(data_freq)
  t2 <- t^2
  # Fit quadratic harmonic model with only the sine parameter
  quad_har_model <- lm(data_freq ~ t + t2 + har[,2])
  # Grab the pvalue for each iteration
  mod_sum_vec[i] <- summary(quad_har_model)$coefficients[,4][4]
}
```

```
# Create dataframe with frequency column matched to p-values
library(dplyr)
df <- cbind(df, mod_sum_vec) # Combine p-values of cos and sin
colnames(df) <- c('Freq','Pval_Cos','Pval_Sin')
df <- df[complete.cases(df),]
df <- df %>%  mutate(Mean_Pval = (Pval_Cos + Pval_Sin)/2)

# Visualse average of p-values
plot(df$Mean_Pval, type = 'o',
     main = 'Mean p-value of Cosine and Sine Fit',
     xlab = 'Frequency',ylab = 'Mean p-value')
```




```
# Find minimum pair of p-values
df[min(df$Mean_Pval) == df$Mean_Pval,]
```

```
##   Freq   Pval_Cos   Pval_Sin  Mean_Pval
## 9   11 0.01663375 0.04333215 0.02998295
```

```
# Freq = 11
```

As shown up the plot and the minimum value found, frequency should be set to 11 to determine the most statistically significant coefficients in regard to a harmonic fit.

Fit Harmonic Quadratic Model

Below I create the harmonic quadratic model with frequency of time series data set to 11.

```
#### CREATE HARMONIC/QUADRATIC MODEL
data_ts_11 <- ts(dataset, frequency = 11)
t <- time(data_ts_11) # determine quadratic variables
t2 <- t^2
t3 <- cos(2*pi*t) # determine harmonic variables
t4 <- sin(2*pi*t)

mod_11 <- lm(data_ts_11 ~ t + t2 + t3 + t4)
summary(mod_11)
```

```
##
## Call:
## lm(formula = data_ts_11 ~ t + t2 + t3 + t4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5918 -1.2351  0.0611  1.2529  3.9525
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.74955    0.84734  -0.885   0.3789
## t            0.62291    0.37805   1.648   0.1031
## t2          -0.17999    0.03667  -4.908 4.38e-06 ***
## t3          -0.64550    0.25678  -2.514   0.0138 *
## t4           0.55759    0.26078   2.138   0.0354 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.73 on 85 degrees of freedom
## Multiple R-squared:  0.7685, Adjusted R-squared:  0.7576
## F-statistic: 70.55 on 4 and 85 DF,  p-value: < 2.2e-16
```

The new model shows an R squared value of 0.7611, this is a further improvement upon the previous R squared value from the quadratic model, 0.7391. One problematic issue with this model however is that the linear coefficient t is no longer statistically significant.

Fit 2nd Harmonic Quadratic Model

I will look into remodeling without the t variable and evaluate the model against the first harmonic quadratic model.

```
# Check without t variable
mod_11v2 <- lm(data_ts_11 ~ t2 + t3 + t4)
summary(mod_11v2) # lower R squared, go with original Harmonic Quadratic model
```

```
##
## Call:
## lm(formula = data_ts_11 ~ t2 + t3 + t4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3222 -1.2116  0.0192  1.3870  4.0941
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.558682   0.298914   1.869   0.0650 .
## t2          -0.120846   0.007588 -15.927   <2e-16 ***
## t3          -0.674602   0.258714  -2.608   0.0108 *
## t4           0.541686   0.263190   2.058   0.0426 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

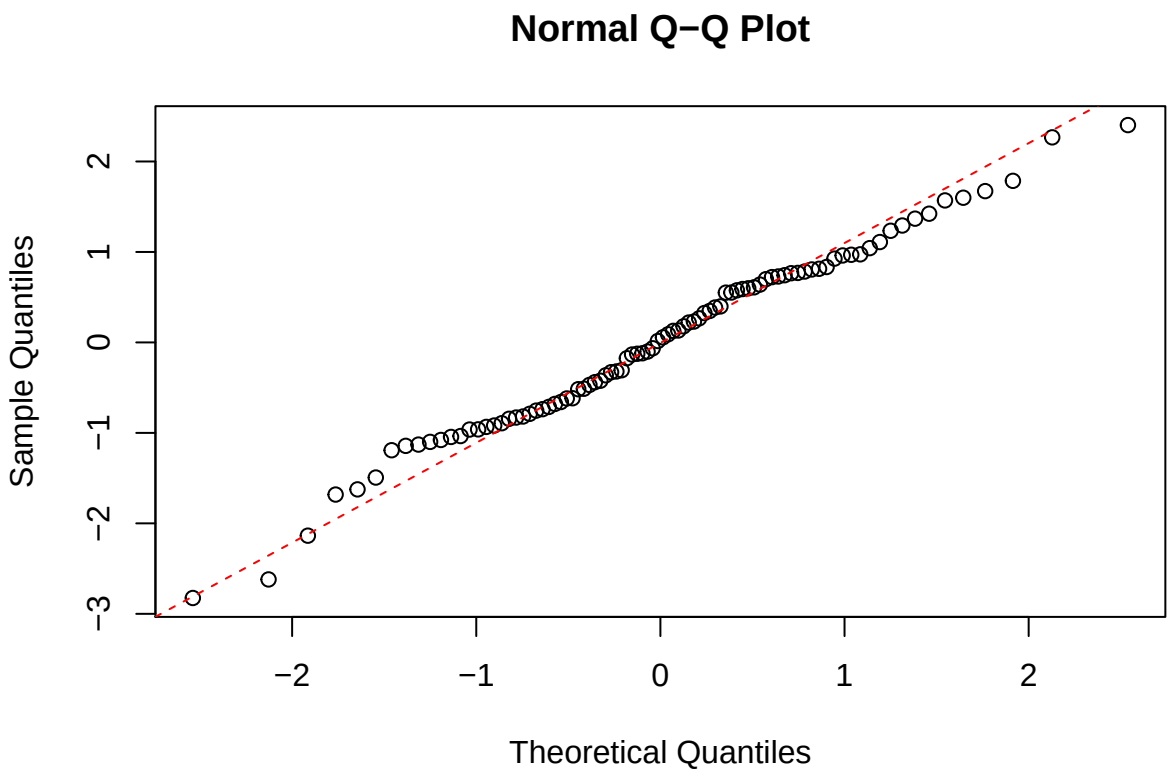
```
## Residual standard error: 1.747 on 86 degrees of freedom
## Multiple R-squared:  0.7611, Adjusted R-squared:  0.7528
## F-statistic: 91.33 on 3 and 86 DF,  p-value: < 2.2e-16
```

The new model shows a decrease in the R squared value from 0.7611 to 0.7321, a further detriment to the models significance is the increase in p-value of the sine parameter deeming it also no longer significant. Thus, the previous harmonic quadratic model will be used for further investigation of residuals and forecasting.

Normality of Standardised Residuals of Harmonic Quadratic Model

Here we will inspect the normality of the standardised residuals of the harmonic quadratic model, first off with a quantile quantile plot and look into any significant deviations from the given line.

```
rquad_har <- rstudent(mod_11) # grab standardised residuals
qqnorm(rquad_har) # plot QQ plot
qqline(rquad_har, col = 'red', lty = 2) # plot normal distribution line
```



The standardised residuals do deviate slightly from the line on the end quantiles, similar to previous linear and quadratic models, indicating all models are not capturing a certain aspect of the time series.

A Shapiro-Wilks test is run to statistically validate any claims of normality found in the distribution of the standardised residuals of the harmonic quadratic model.

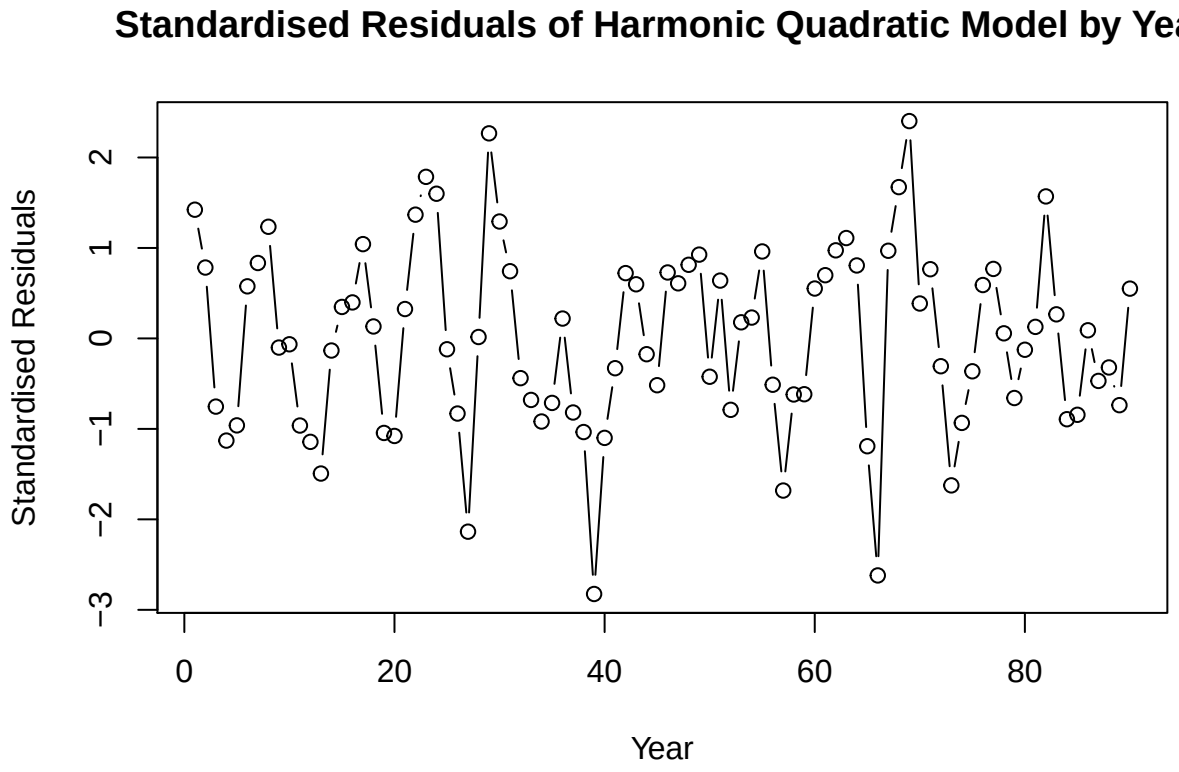
```
shapiro.test(rquad_har)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rquad_har
## W = 0.98801, p-value = 0.5855
```

The Shapiro-Wilks test shows a p-value exceeding the significance level of 0.05. Therefore we reject the null hypothesis and presume normality of the distribution of standardised residuals.

Visualise Standardised Residuals of Harmonic Quadratic Model

```
plot(rquad_har, type = 'b',
     xlab = 'Year',
     ylab = 'Standardised Residuals',
     main = 'Standardised Residuals of Harmonic Quadratic Model by Year')
```

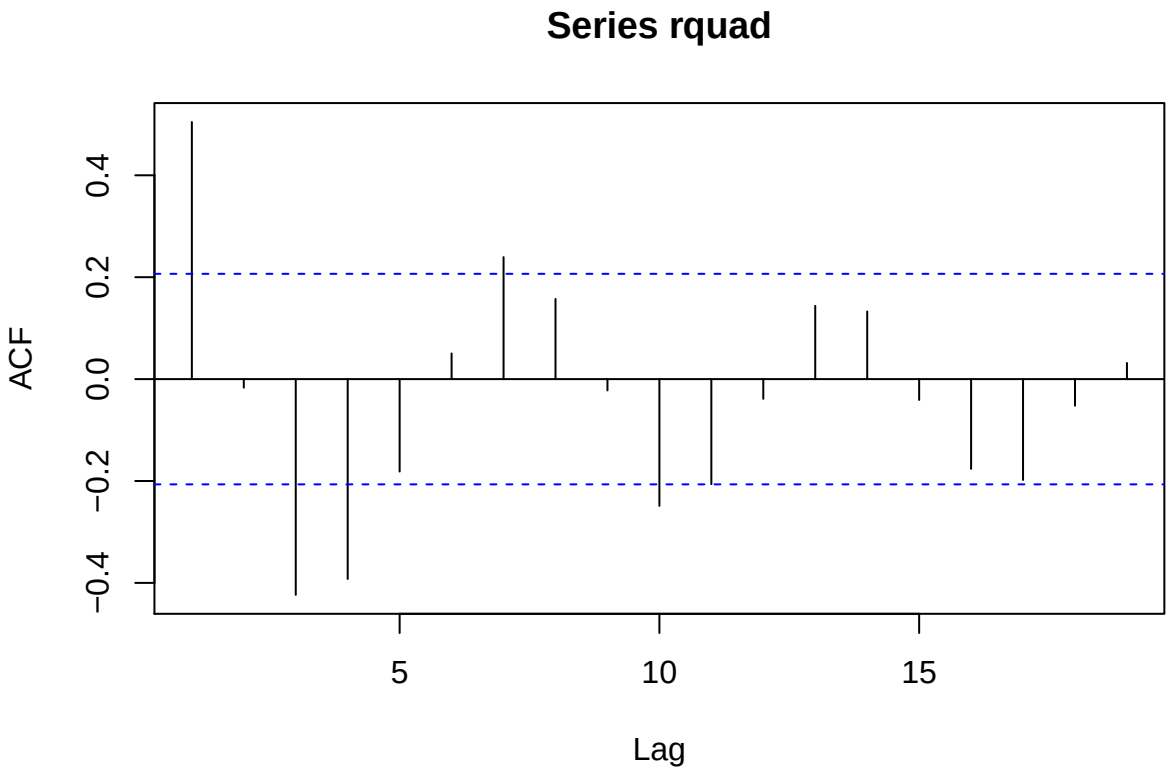


The given plot of the standardised residuals of the harmonic quadratic model show a relatively constant mean but some similar changes in variation compared to the previous quadratic model. There is still a cyclical trend present in the residuals, which provides us with the notion that this model does not capture some cyclical or seasonal trend in the time series data.

Auto Correlation Function of Harmonic Quadratic Residuals

Here we run an autocorrelation function over the harmonic quadratic residuals at sequential lags to look for any significant autocorrelations. Again, if any significant autocorrelations or patterns are found, the model is unable to capture something within the time series.

```
acf(rquad) # still a ~0.5 signif at lag 1, a clearly a trend in the pattern
```



There is clearly a pattern in the ACF, indicating a trend is not being captured by the harmonic quadratic model. Further to this, there are multiple lags with significant autocorrelations intercepting the given threshold. This again deems this model as inadequate for capturing all trends and behaviors in the time series model. However, the lag 1 autocorrelation is similar to the quadratic model. The only significant difference in performance measure, despite all models being proven inadequate from ACF threshold intercepts and trends found in residuals, was the higher R squared in the harmonic quadratic. Hence, I will forecast the next 5 years using the harmonic quadratic model.

Forecasting with Chosen Model

Predict Values

Below I will find the interval for the given time series object, I will then use this interval value to determine the sequence for the new prediction data. I will then visualise this new data along side the time series and chosen model, being the harmonic quadratic model.

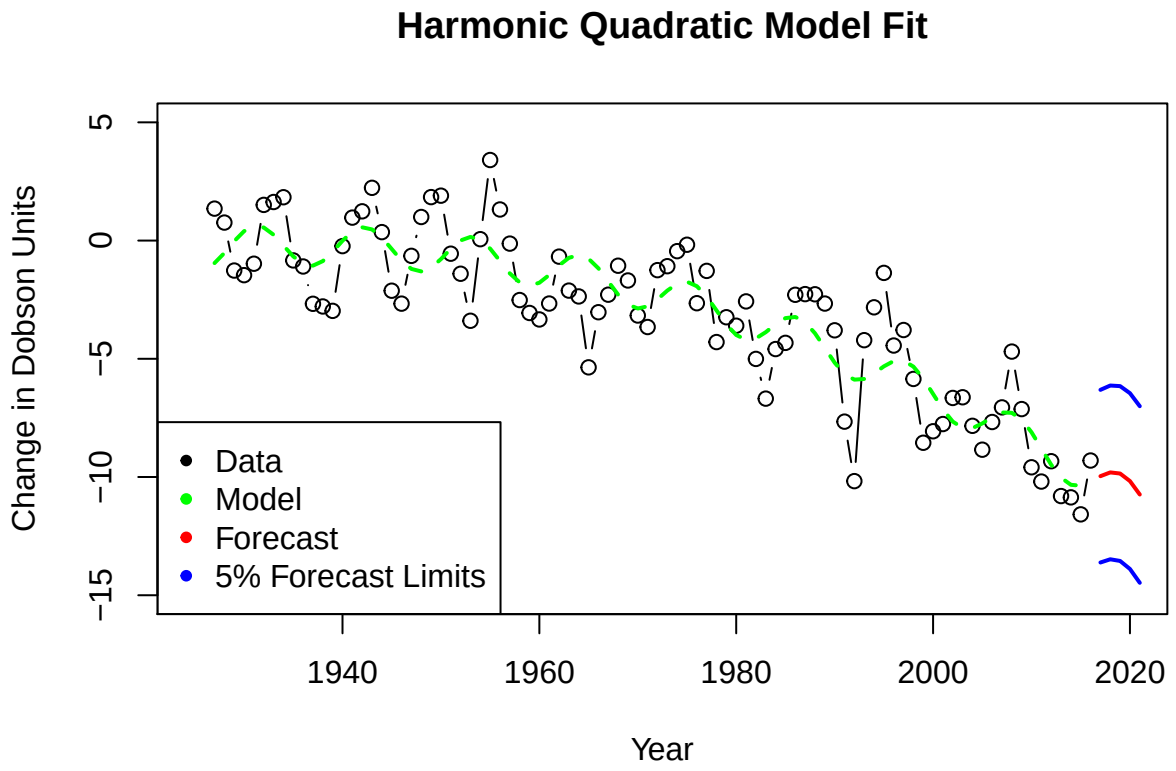
```
### PREDICT HARMONIC
freq <- 11 # set freq of ts

interval <- 1/freq # find the interval values

h <- 5 # set prediction

t_last <- t[length(t)] # find last value

t_new <- seq((t_last + interval), (t_last + interval * h), by = interval) #

t <- t_new
t2 <- t^2
t3 <- cos(2*pi*t)
t4 <- sin(2*pi*t)

new_t_df <- data.frame(t, t2, t3, t4)

predt <- predict(mod_11, new_t_df, interval = 'prediction')
```

Visualise and Forecast

Below I visualise the original time series data with the overlay of the harmonic quadratic model. I then forecast the next 5 years using the previously generated prediction data. A legend is supplied for visual aid.

```
### VISUALISING MODEL AND FORECAST
plot(
  ts(data_ts_11, start = 1927),
  type = 'b',
  xlim = c(1925, 2020),
  ylim = c(-15, 5),
  lwd = 1,
  xlab = 'Year',
  ylab = 'Change in Dobson Units',
  main = 'Harmonic Quadratic Model Fit'
  )  # plot time series
  lines(
  ts(mod_11$fitted.values, start = 1927),
  col = 'green',
  lwd = 2,
  lty = 2
  ) # overly model
  lines(
  ts(as.vector(predt[, 1]), start = 2017),
  col = 'red',
  lwd = 2,
  lty = 1
  ) # forecast model
  lines(
  ts(as.vector(predt[, 2]), start = 2017),
  col = 'blue',
  lwd = 2,
```

```
lty = 1
) # add lower limit
lines(
ts(as.vector(predt[, 3]), start = 2017),
col = 'blue',
lwd = 2,
lty = 1
) # add upper limit
legend(
'bottomleft',
col = c('black', 'green', 'red', 'blue'),
legend = c('Data', 'Model', 'Forecast', '5% Forecast Limits'),
pch = 20
)
```



Summary

For this report, I investigated the fit and performance measures of a linear, quadratic and harmonic model for the given time series data.

The time series data given showed an autocorrelation value of 0.87. This value indicates there is a clear trend in the time series. A visual inspection of the plotted graph showed a clear trend over time. The three aforementioned models were tested to try and capture the behavior and trends of the time series data.

The performance measures for each model were checking the p-values of coefficients, the R squared value, the normality of standardised residuals using Q-Q plots and Shapiro-Wilks tests, visually inspect a plot of the residuals and running auto correlation functions. All models created and observed had significant auto correlations at multiple lags, with ~0.6 for the linear and ~0.5 for both quadratic and harmonic quadratic, rendering them as inadequate models which are unable to capture all trends and behavioral aspects of the timer series.

Despite the statistically insignificant models created, the one with the greatest R squared would be used for forecasting the next 5 years. The harmonic quadratic model had the highest R squared value at 0.7685 (meaning 76.85% of variation in the data is explained by the model) therefore I used this model to forecast for the next 5 years. The model predicted that the change in percentage of Dobson units would continue in a sinusoidal downward trend. A possible explanation of this could be the ongoing exponential increase in manufacturing production, greenhouse gases from the increased agriculture sector, more vehicles in transit, or a combination of all of them.