MATH1318 Time Series Analysis - Final Project

Tanish Saajan, s3940991

Introduction

The data represents the daily closing price of the stock for the **Coca Cola company** from **2020 to 2022.**

Required Libraries

```
library(TSA)
```

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

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

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

```
library(fUnitRoots)
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
## Registered S3 methods overwritten by 'forecast':
##   method       from
##   fitted.Arima TSA
##   plot.Arima   TSA
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(fGarch)
```

```
## NOTE: Packages 'fBasics', 'timeDate', and 'timeSeries' are no longer
## attached to the search() path when 'fGarch' is attached.
##
## If needed attach them yourself in your R script by e.g.,
##         require("timeSeries")
```

```
library(rugarch)
```

```
## Loading required package: parallel
```

```
##
## Attaching package: 'rugarch'
```

```
## The following object is masked from 'package:stats':
##
##     sigma
```

```
library(tseries)
```

Required Functions

Creating a **function** that plots the Time Series plot, Histogram, QQ plot,**ACF**, perform the **shapiro test** for the normality and plot the results of **Mc-Leod Test Statistic** for the residuals whenever it is called.

```
sort.score <- function(x, score = c("bic", "aic")){
  if (score == "aic"){
    x[with(x, order(AIC)),]
  } else if (score == "bic") {
    x[with(x, order(BIC)),]
  } else {
    warning('score = "x" only accepts valid arguments ("aic","bic")')
  }
}
residual.analysis <- function(model, std = TRUE,start = 2, class = c("ARIMA","GARC
H","ARMA-GARCH", "fGARCH")[1]){
  library(TSA)
  library(FitAR)
  if (class == "ARIMA"){
    if (std == TRUE){
      res.model = rstandard(model)
    }else{
      res.model = residuals(model)
    }
  }else if (class == "GARCH"){
    res.model = model$residuals[start:model$n.used]
  }else if (class == "ARMA-GARCH"){
    res.model = model@fit$residuals
  }else if (class == "fGARCH"){
    res.model = model@residuals
  }else {
    stop("The argument 'class' must be either 'ARIMA' or 'GARCH' ")
  }

  par(mfrow=c(2,3))
  plot(res.model,type='o',ylab='Standardised residuals', main="Time series plot of
standardised residuals")
  abline(h=0)
  hist(res.model,main="Histogram of standardised residuals")
  qqnorm(res.model,main="QQ plot of standardised residuals")
  qqline(res.model, col = 2)
  acf(res.model,main="ACF of standardised residuals")
  print(shapiro.test(res.model))
  k=0
  LBQPlot(res.model, lag.max = 30, StartLag = k + 1, k = 0, SquaredQ = FALSE)
  par(mfrow=c(1,1))
}
```

Data Import

In this section we read/import the data into R, then saved it as a data frame using the **read.csv()** funtion. The class() function is being used to make sure the class of the dataset is in **Time-Series(TS)**. The Time-Series Plot of Coca Cola Series from 2020 to 2022 is shown as follows:

```
setwd("/Users/tanishsaajan/Documents/Time Series/Final Project")
par(mfrow=c(1,1))
cola<-read.csv("Cola.csv")
class(cola)
```

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

```
c<-tail(cola,600)
c <- na.omit(c)
c<-c$Close
cola<-ts(as.vector(c),frequency=1)
class(cola)
```

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

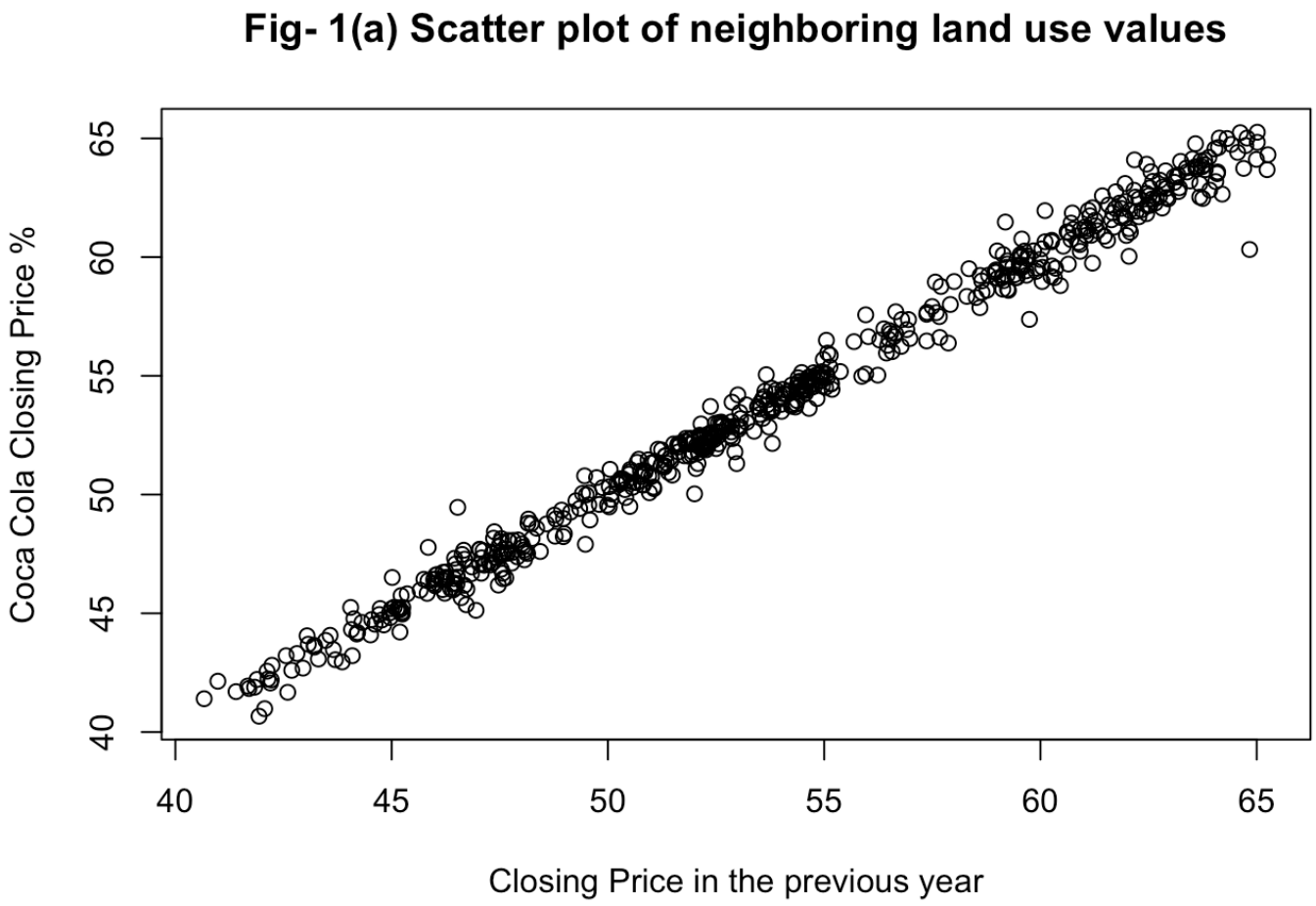
We have set the frequency to **1** because the data is collected on a **daily** basis, and we have verified that the dataset has been successfully transformed to a time series.

Correlation Between the consecutive time points.

```
xlag= cola
xlag1= zlag(xlag)
index = 2:length(xlag1)
cor(xlag[index],xlag1[index])
```

```
## [1] 0.9947777
```

```
plot(y=cola,x= zlag(xlag),ylab='Coca Cola Closing Price %', xlab='Closing Price in
the previous year',
main= "Fig- 1(a) Scatter plot of neighboring land use values")
```



From the above **correlation** and the **Fig- 1(a)** of the **scatter plot** we can see that consecutive time points are highly positively correlated with each other as their value is close to one that is **0.9947777**. , it indicates a **high positive correlation** between consecutive time points in the closing price of the Coca Cola series. This suggests that the current closing value is **highly** dependent on its previous value, indicating a **strong relationship** between adjacent data points.

Here,

$H_0 : B_0$ is **equal** to zero and is statistically **not** significant

$H_1 : B_0$ is **not equal** to zero and is **statistically** significant

$H_0 : B_1$ is **equal** to zero and is **statistically** not significant

$H_1 : B_1$ is **not equal** to zero and is **statistically** significant

```
t<-time(cola)
model123<-lm(cola~t)
summary(model123)
```

```
##
## Call:
## lm(formula = cola ~ t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8742 -1.3574  0.2168  1.5636  5.9058
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.378e+01  1.963e-01  223.02   <2e-16 ***
## t           3.314e-02  5.659e-04   58.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.401 on 598 degrees of freedom
## Multiple R-squared:  0.8515, Adjusted R-squared:  0.8513
## F-statistic:  3430 on 1 and 598 DF,  p-value: < 2.2e-16
```

Here the B_1 indicates that there is **3.314e-02** of increase in the daily closing price of the coca cola series.

We can **reject** null hypothesis for both the coefficients as their p-value is **less** than the 5% significance level and they both are **statistically significant**.

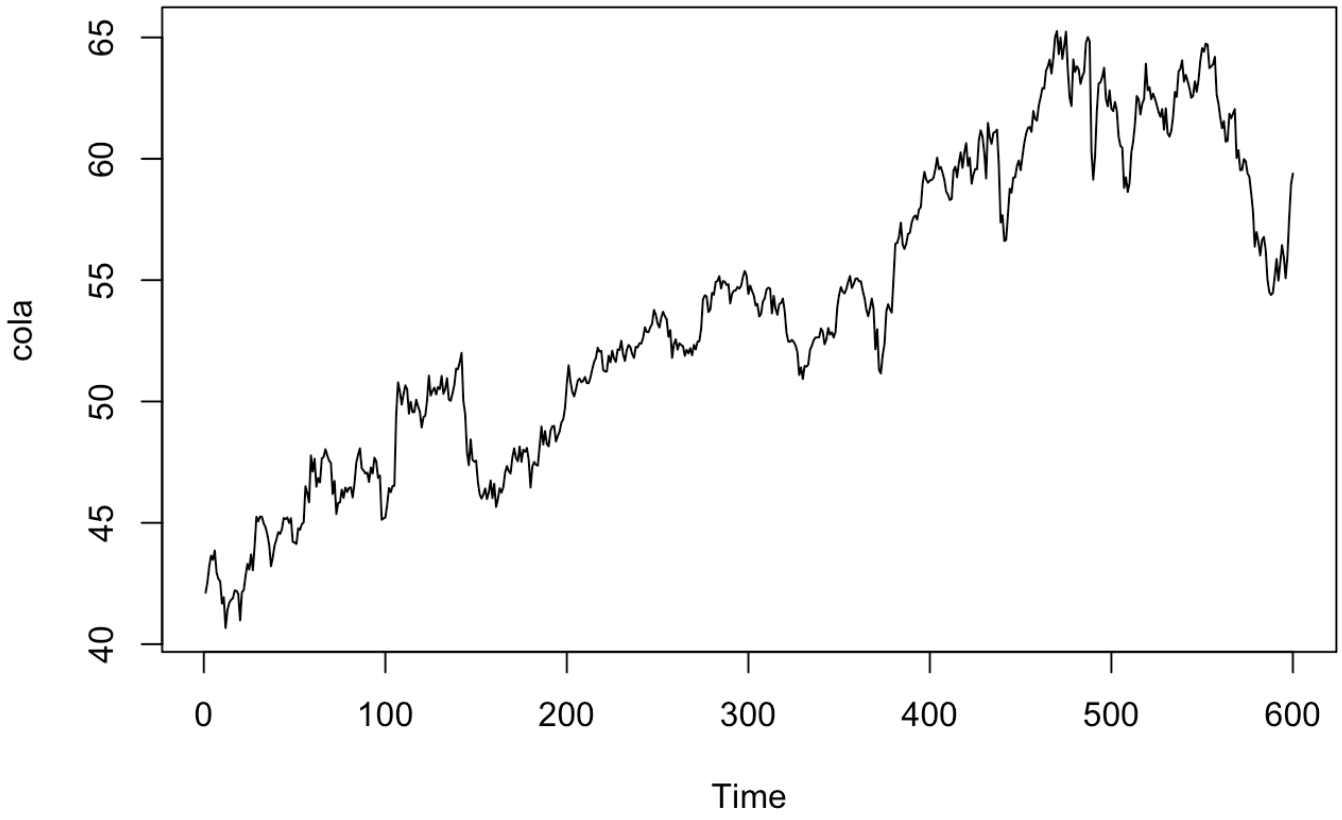
Analyzing the Data:

After converting the data to time series format , we generated a time series plot and will analyse the following

- Trend
- Seasonality
- Variance
- Behavior

```
plot(cola,main="Fig 1.0 - Time series plot of Coca Cola series")
```

Fig 1.0 - Time series plot of Coca Cola series



The initial observations from **the Fig 1.0 are:**

1. **Upward** trend in the series can be observed.
2. There is a **no** obvious **seasonality**, as there are no repeating patterns in the series
3. In the series, there is **2** or more **intervention** or change point.
4. **No** obvious **change** in **variance** can be observed in the series that means the series follow a **deterministic** trend.
5. Furthermore, the succeeding observations imply the presence of **auto-regressive behavior** and fluctuations and bouncing observations around the mean level imply the presence of **moving average behavior**.
6. As it is mostly AR behavior, we may see values of **p>q** in our final model.
7. Furthermore, from fig 1.1 we can see a slowly decaying pattern, which gives the impression that our q may be equal to 0 because a slowly decaying pattern is an indication of an **AR** process from which only the order of p can be determined.

Fitting Regression Models:

As, series **do not** under go **seasonality**, hence we will try and fit **linear** and **quadratic** models.

Here,

$H_0 : B_0$ is **equal** to zero and is statistically **not** significant

$H_1 : B_0$ is **not equal** to zero and is **statistically** significant

$H_0 : B_1$ is **equal** to zero and is **statistically** not significant

$H_1 : B_1$ is **not equal** to zero and is **statistically** significant

1.Linear Model

```
#Linear
tt=time(cola)
model1=lm(cola~tt)
summary(model1)
```

```
##
## Call:
## lm(formula = cola ~ tt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8742 -1.3574  0.2168  1.5636  5.9058
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.378e+01  1.963e-01  223.02   <2e-16 ***
## tt          3.314e-02  5.659e-04   58.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.401 on 598 degrees of freedom
## Multiple R-squared:  0.8515, Adjusted R-squared:  0.8513
## F-statistic:  3430 on 1 and 598 DF,  p-value: < 2.2e-16
```

Labeling the Linear Model as model1.

Summary Statistics of Linear Model

- Intercept,slope of the linear trend model is statistically **significant** at 5% as their values are less than the significance level that is 0.05.That means, we can **reject** null hypothesis and conclude our linear trend coefficient is significant
- p-value is the overall significance and the model has p-value **less** than the significance level **0.05** Hence, the overall **significance** of the model is **normal**.
- According to adjusted R^2, about **85.13**% of the variation in the traders data series is **explained** by the linear time trend.

- The minimum and the maximum errors that our model is showing are **-8.8742** and **5.9058** respectively that implies they have a small range.

Performing **Shapiro Walk Test** to test for **Normally** Distributed **Errors**

H_0: The Errors have a **normal** distribution.

H_1: The Errors does not have a normal distribution.

```
par(mfrow=c(3,2))
plot(cola,ylab='Observation', main = "Fig- 1.1 Fitted linear model to share market
series")
abline(model1)

plot(y=rstudent(model1),x=as.vector(time(c)), xlab='Time',
     ylab='Standardized Residuals', main = "Fig- 1.2 Standardized Residuals from L
inear Model")
hist(rstudent(model1),xlab='Standardized Residuals from'
     , main = "Fig- 1.3 Histogram of standardised residuals
     fitted to to share market series.")
y = rstudent(model1)
qqnorm(y, main = "Fig- 1.4 QQ plot of standardised residuals for
       linear model fitted to share market series.")
qqline(y, col = 2, lwd = 1, lty = 2)

acf(rstudent(model1), main = "Fig- 1.5 ACF of standardized residuals.")
shapiro.test(rstudent(model1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstudent(model1)
## W = 0.95638, p-value = 2.453e-12
```

Fig- 1.1 Fitted linear model to share market series

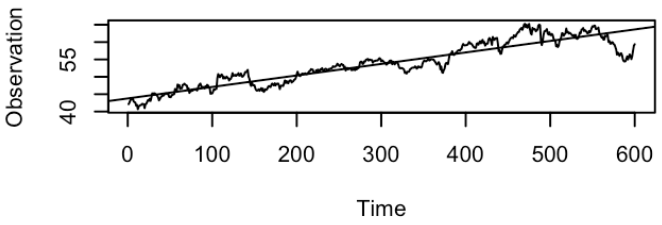


Fig- 1.2 Standardized Residuals from Linear Model

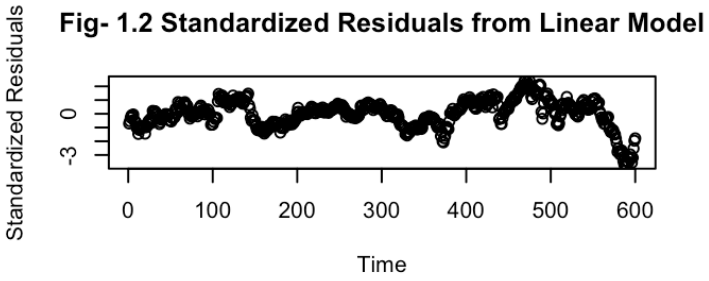


Fig- 1.3 Histogram of standardised residuals fitted to to share market series.

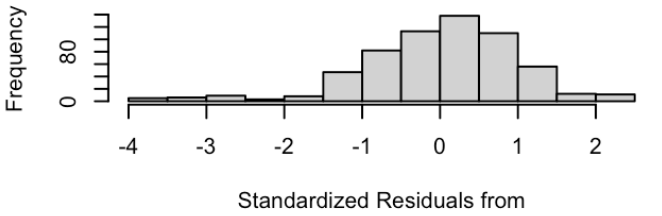


Fig- 1.4 QQ plot of standardised residuals for linear model fitted to share market series.

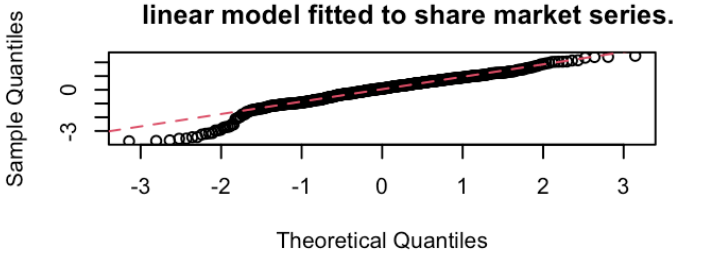
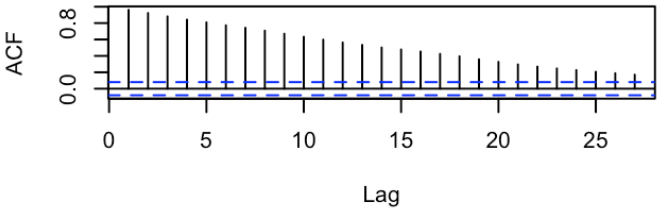


Fig- 1.5 ACF of standardized residuals.



Interpretation of Residual Analysis of Linear Model

- **Fitted Plot:** From the Fig 1.1, we can observe that the Linear model is not able to capture the pattern of the series and hence is capturing noise.
- **Standardized Residuals:** From the Fig 1.2, we can observe that their is significanct amount of change in variance left in the residuals.
- **Histogram** : The Fig- 1.3 concludes the standardized residuals are not fully normally distributed as they are in between -3 and +3.As,our model is not capturing the ending portion of the time frame .Also, there is some outliers that can be captured,
- **QQ plot** : FIg- 1,4 representing QQ- plot give us an overview of the normality.We can see that few of the time series points fall on the reference line, and that there is a slight deviation from the reference line pattern at the front and end of the tail.Also, by looking at the normality test the p-value is ***2.453e-12*** which is less than 0.05.Hence,we can **accept** the null hypothesis that states our data is **not normal**. We can conclude that the QQ plot **does not** fully support a normally distributed stochastic component in the linear trend model.
- **Sample Autocorrelation Function(ACF)** : According to the Fig- 1.5 we can say that trend and seasonality can be observed in the series from the slowly decaying wave downward pattern indicating a non stationary series.Also, we can conclude that there is valuable information left in residuals which our linear model is unable to capture and follow a stochastic trend.

Overall, we see a considerable amount of departure from the reference line, we conclude that the normality assumption **does not** hold for the investment in coca cola series. The Shapiro-Wilk test also confirms this inference with a p-value **less** than **0.05** and we have enough evidence to **reject** the null hypothesis(H_0) that implies that the residuals **do not** follow a **normal** distribution.

2.Quadratic

```
#Quadratic
t2=tt^2
model2 = lm(cola ~ tt+t2)
summary(model2)
```

```
##
## Call:
## lm(formula = cola ~ tt + t2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7059 -1.3468  0.1927  1.6311  5.8776
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.244e+01  2.860e-01 148.366  < 2e-16 ***
## tt           4.648e-02  2.198e-03  21.144  < 2e-16 ***
## t2          -2.219e-05  3.542e-06  -6.265 7.15e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.328 on 597 degrees of freedom
## Multiple R-squared:  0.8607, Adjusted R-squared:  0.8602
## F-statistic:  1844 on 2 and 597 DF,  p-value: < 2.2e-16
```

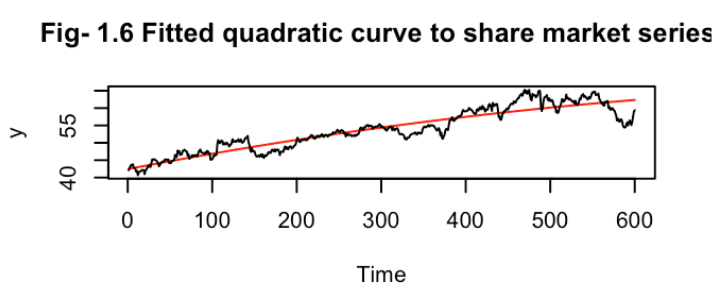
Labeling the Quadratic Model as model2.

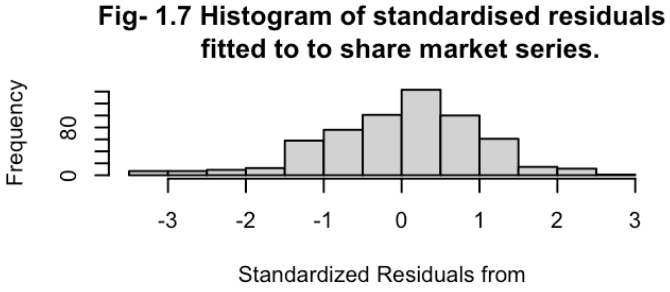
Summary Statistics of Linear Model

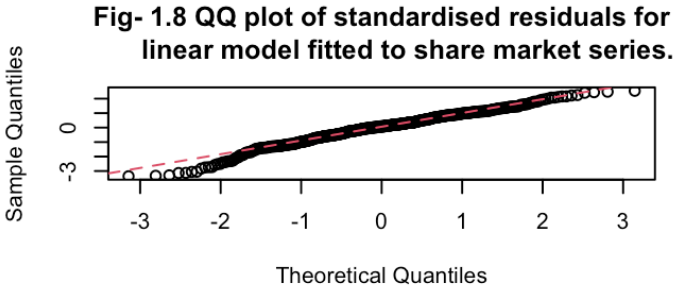
- Intercept,beta 1 and beta2 of the linear trend model is **statistically** significant at 5% as their values are **less** than the significance level that is **0.05**.That means, we can reject null hypothesis Also, our **quadratic** trend coefficient is **significant**
- p-**value** is the overall significance and the model has p-value **less** than the **significance** level **0.05** Hence, the **overall** significance of the model is **normal**.
- According to adjusted R^2, about **86.02%** of the variation in the cola cola series is explained by the **quadratic** time trend.
- The **minimum** and the **maximum** errors that our model is showing are -**7.7059** and **5.8776** respectively that implies they have a small range.

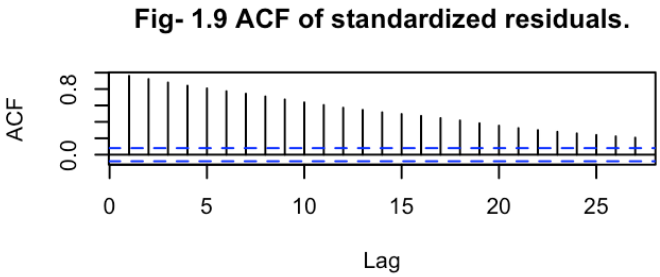
```
par(mfrow=c(3,2))
plot(ts(fitted(model2)), ylim = c(min(c(fitted(model2),
                                        as.vector(cola))), max(c(fitted(model2),as
.vector(cola)))),ylab='y' ,
     main = "Fig- 1.6 Fitted quadratic curve to share market series.",col="red")
lines(as.vector(cola))
hist(rstudent(model2),xlab='Standardized Residuals from'
     , main = "Fig- 1.7 Histogram of standardised residuals
     fitted to to share market series.")
y = rstudent(model2)
qqnorm(y, main = "Fig- 1.8 QQ plot of standardised residuals for
       linear model fitted to share market series.")
qqline(y, col = 2, lwd = 1, lty = 2)

acf(rstudent(model2), main = "Fig- 1.9 ACF of standardized residuals.")
shapiro.test(rstudent(model2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstudent(model2)
## W = 0.98163, p-value = 7.429e-07
```









Interpretation of Residual Analysis of Quadratic Model

- **Residual Plot:** From the Fig 1.6, we can observe that the Quadratic model is **not** able to **fully** capture the pattern of the series, and hence is **capturing noise**.
- **Histogram** : The Fig- 1.7 concludes the standardized residuals are **normally** distributed as they are in between -**3** and +**3**.Also, there is **no** outlier present in the series as the data is distributed between the -3 and +3.
- **QQ plot** : FIg- 1.8 representing QQ- plot give us an overview of the normality.We can see that few of the time series points fall on the reference line, and that there is a slight deviation from the reference line pattern at the front and end of the tail.Also, by looking at the normality test the p-value is ***7.429e-12*** which is less than 0.05.Hence,we can **accept** the null hypothesis that states our data is **not** normal. We can conclude that the QQ plot **does** not fully support a normally distributed stochastic component in the quadratic model.
- **Sample Autocorrelation Function(ACF)** : According to the Fig- 1.9 we can say that trend and seasonality can be observed in the series from the slowly decaying wave downward pattern indicating a non stationary series.Also, we can conclude that there is valuable information left in residuals which our quadratic model is unable to capture and follow a stochastic trend.

Performing **Shapiro Walk Test** to test for **Normally** Distributed **Errors**

H_0 : The Errors have a **normal** distribution.

H_1 : The Errors does not have a normal distribution.

Overall, we see a considerable amount of departure from the reference line, we conclude that the normality assumption does not hold for the investment in coca cola series. The Shapiro-Wilk test also confirms this inference with a p-value less than 0.05 and we have enough evidence to **reject** the null hypothesis(H_0) that implies that the residuals **do not** follow a **normal** distribution.

ACF & PACF Plots of Closing Price of Coca Cola Series:

```
#Series
par(mfrow=c(2,2))
plot(cola,main="Fig 2.0 - Time series plot of Closing Price of Coca Cola series")
acf(cola,lag.max = 100,main = "Fig 2.1 - ACF plot of Coca Cola series.")
#All the auto correlation are significant at seasonal lags,slowing decaying trend
is present,seasonal trend is present signified by D
pacf(cola, main=" Fig 2.2 - PACF plot of Coca Cola series.")
McLeod.Li.test(y=cola, main="Fig 2.3 -McLeod-Li Test Statistics for coca cola")
```

) - Time series plot of Closing Price of Coca C

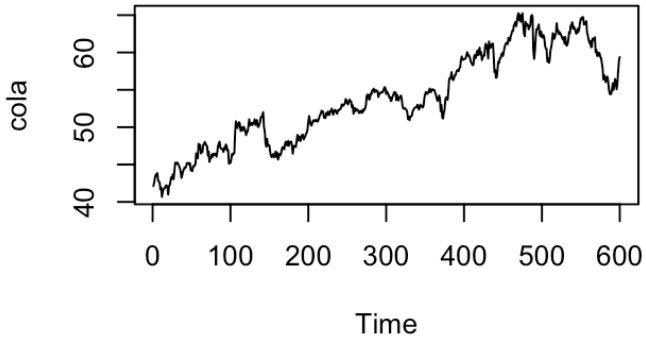


Fig 2.1 - ACF plot of Coca Cola series.

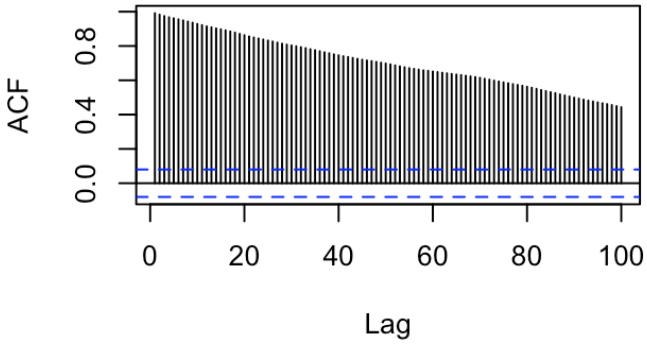


Fig 2.2 - PACF plot of Coca Cola series.

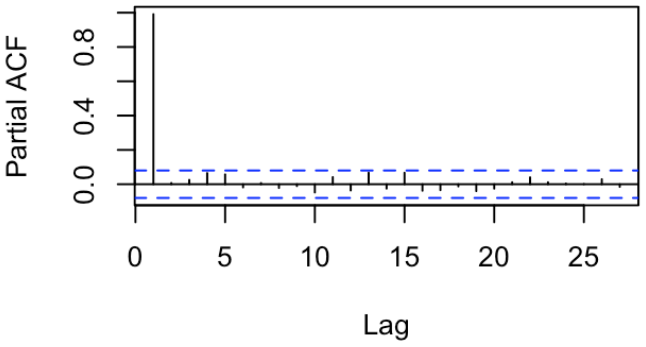
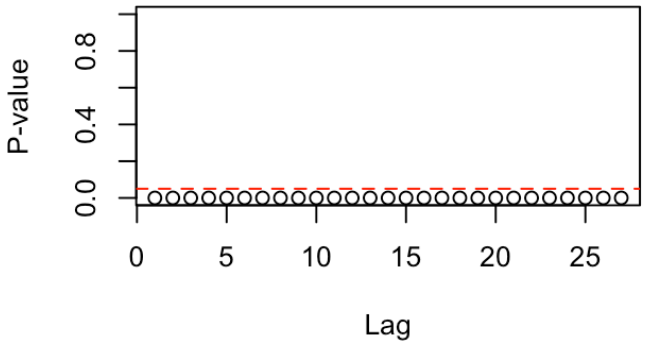


Fig 2.3 -McLeod-Li Test Statistics for coca c



As, there is **no** obvious **seasonality** can be observed from the Fig, 2.0,We have confirmed the presence of seasonality by plotting **ACF** plot.

The presence of a slight wave pattern in **Figure 2.1** indicates the presence of seasonality in the series. This suggests that the series undergo **regular** patterns or fluctuations at specific intervals.

Furthermore, **slowly** decaying pattern in **Fig- 2.1** and there is high first lag value in the partial autocorrelation function (PACF) of Fig- 2.2.Hence, the slowing decaying pattern and high first lag from ACF and PACF, indicates **non stationary** of the series

The highly significant results of the McLeod-Li tests in **Figure 2.3** confirm the presence of high autocorrelation in the series. This indicates that there is a strong relationship between the observations at different lags.

To confirm the existence of non-stationarity in the time series, we will use the Augmented Dickey-Fuller (ADF) unit-root test.

Where,

H0: The time series has a **unit** root (is non-stationary)

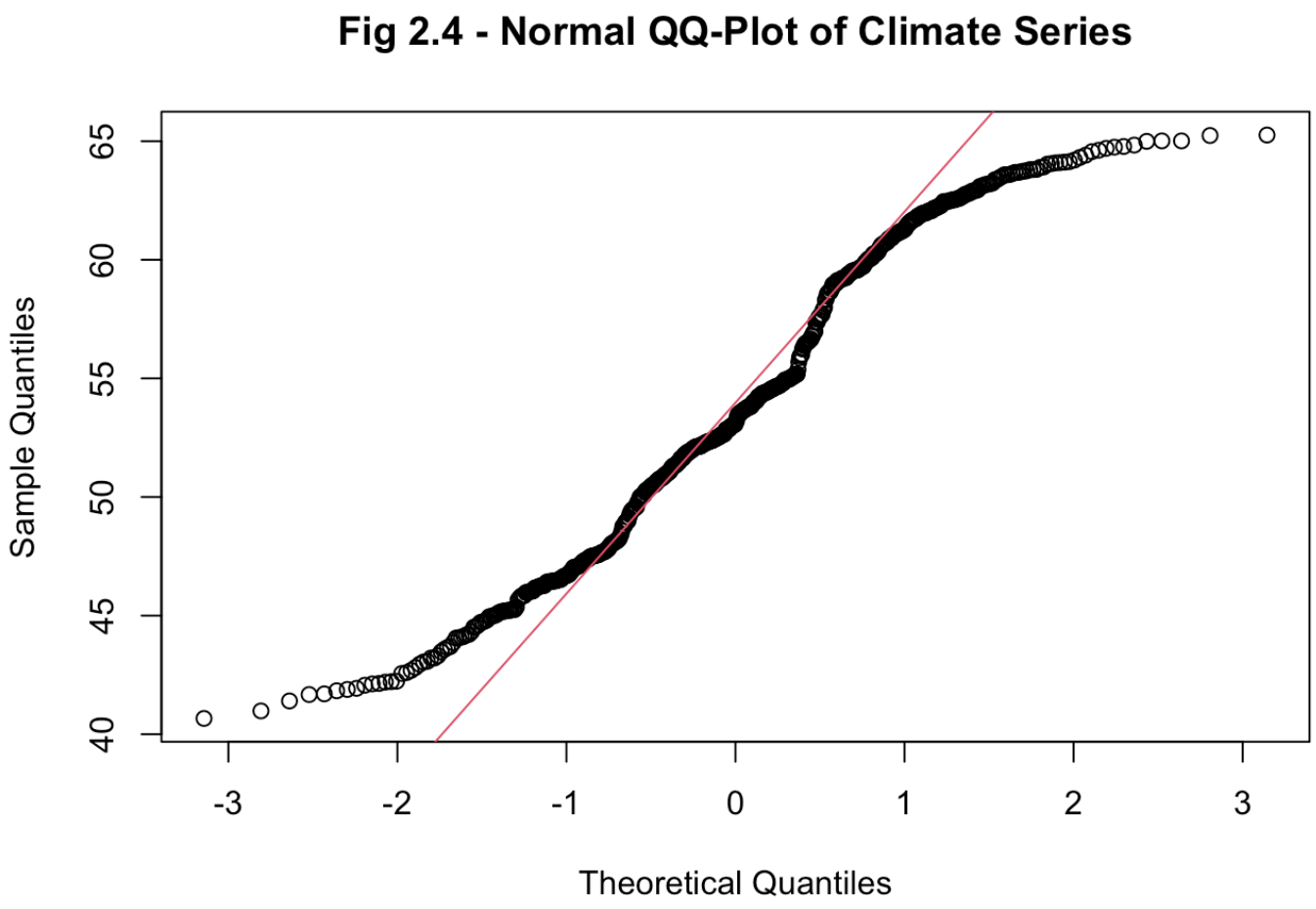
HA: The time series **does not** have a unit root (is stationary)

```
adf.test(cola)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  cola
## Dickey-Fuller = -2.8925, Lag order = 8, p-value = 0.2005
## alternative hypothesis: stationary
```

From the Augmented Dickey-Fuller Test,where the p>0.05 significance level, and we **cannot** reject the null hypothesis and which indicates the series has a **unit** root that is **not stationary** series

```
qqnorm(cola,main="Fig 2.4 - Normal QQ-Plot of Climate Series")
qqline(cola,col=2)
```



By examining the plot shown in **Figure 2.4**, we can see that both ends of the plot are deviated from the normal distribution line, which suggests that the time series is **not normally** distributed.

Hence, we will performing normality test that is **shapiro** test to check if the series follows a **normal** distribution

Where,

H0: The sample **comes** from a normal distribution

HA: The sample **does not** come from a normal distribution

```
shapiro.test(cola)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cola
## W = 0.96685, p-value = 2.141e-10
```

As p-value of the data is **less** than the **significance** level which is **0.05**, we **reject** null hypothesis that states that the data **does not** follow a normal distribution or **does not** exhibits **normality**.

Transformation

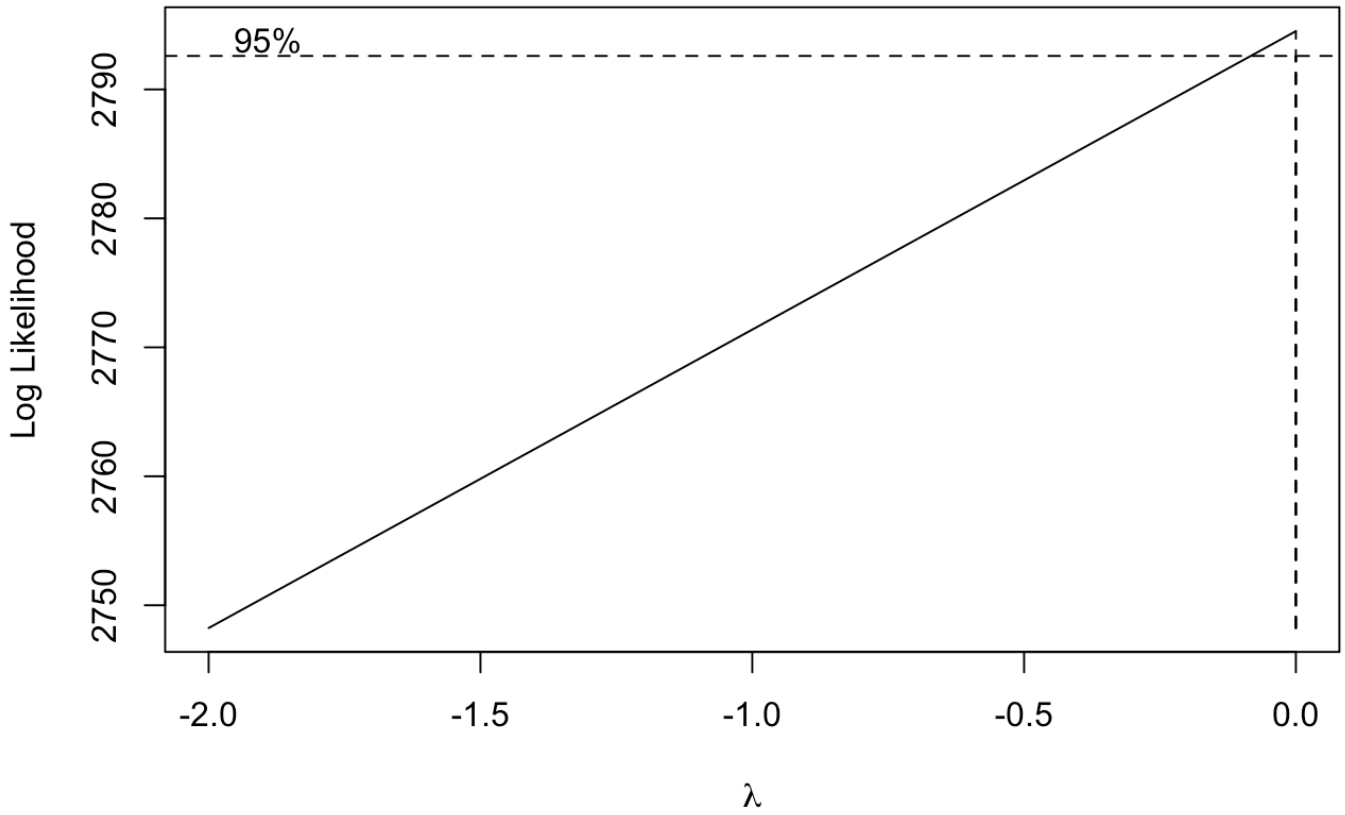
Data transformation are important tool for proper statistical analysis as it helps to stabilize the **variance** and the **normal distribution** of the series.

Because our series is not normally distributed in this case, we will use **box-cox** as the transformation family that depends on lambda, where lambda gives us different transformations such as **log** transformation when it approaches zero and power transformation when it is away.

As, the values in data are all positive we **might** see the value of lambda **greater** than 1.

We will also compare the transformation of the **Box-Cox** with the **log transformation** if the ideal value of lambda is close to zero.

```
BC <- BoxCox.ar(cola,lambda = seq(-2, 0.5, 2))
```




```
#,lambda = seq(-1, 0.5, 0.01) If you get an error.
lambda <- BC$lambda[which(max(BC$loglike) == BC$loglike)]
lambda
```

```
## [1] 0
```

Based on this interval and the Box-Cox transformation formula, we can conclude that the ideal value of lambda is **zero**, which suggests that we need to apply a **log** transformation to the time series data.

As, the series undergo seasonality we will use Seasonal Auto regressive Integrated Moving Average (**SARIMA**) models.

Where, SARIMA models have two sets of orders **(p,d,q)** for the **regular** part of the model and **(P,D,Q)** for the seasonal part of the model and a **frequency**.

To start, since there is slowly decaying pattern just at the periods, we will fit **SARIMA(0,0,0)x(0,1,0)1** model and display time series, ACF and PACF plots of the residuals/standardised residuals.

```
par(mfrow=c(2,2))
m1.temp = Arima(cola,order=c(0,0,0), seasonal=list(order=c(0,1,0), period=1))
res.m1 = rstandard(m1.temp)
plot(res.m1,xlab='Time',ylab='Residuals',
                                  main="Fig-2.5 Time series plot of the residuals f
or Cola.")
acf(res.m1, main="Fig-2.6 ACF after first seasonal dif")
pacf(res.m1,main="Fig-2.7 PACF after first seasonal dif")
#McLeod.Li.test(y=res.m1, main="Fig-2.8 McLeod-Li Test Statistics for coca cola")
```

Fig-2.5 Time series plot of the residuals for C

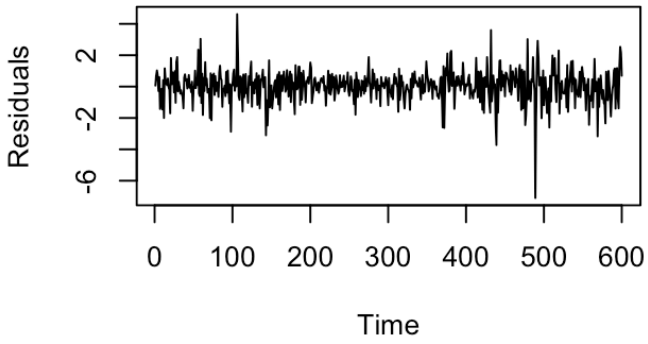


Fig-2.6 ACF after first seasonal dif

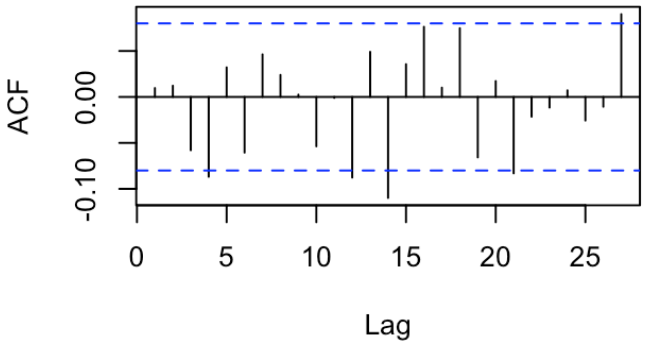
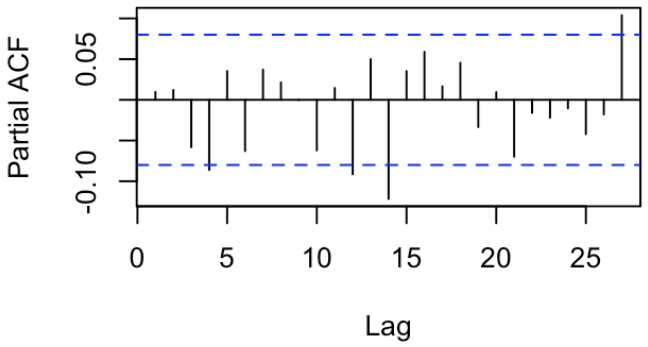


Fig-2.7 PACF after first seasonal dif



From the output above we can conclude that,from **Fig- 2.5** the seasonal trend has been **removed** that was present in the original series.

Furthermore, their is **no** trend in the **Fig 2.6** and **Fig 2.7** of the ACF and the PACF of the series as their is **no slowing** pattern in the ACF and **no first high value** in the PACF.

Hence, their is **no** need to do further **differencing** as the first differencing has **worked** for us.

Since there are significant autocorrelation at the first seasonal lag in both ACF and PACF the order of P from the PACF is **3** and order of Q from ACF is 3.

Fitting **SARIMA(0,0,0)x(3,1,3)1 1** model; and displaying time series, ACF and PACF plots of the residuals/standardized residuals.

```
m2.cola = Arima(cola,order=c(0,0,0), seasonal=list(order=c(3,1,3), period=1))
res.m2 = residuals(m2.cola);

par(mfrow=c(2,2))
plot(res.m2,xlab='Time',ylab='Residuals',main="Fig- 2.8 Time series plot of the re
siduals")
acf(res.m2, lag.max = 48, main = "Fig- 2.9 ACF of the residuals for 2nd diff")
pacf(res.m2, lag.max = 48, main = "Fig- 3.0 PACF of the residuals")
#McLeod.Li.test(y=res.m2, main="Fig- 3.1 McLeod-Li Test Statistics for cola")
```

Fig- 2.8 Time series plot of the residuals

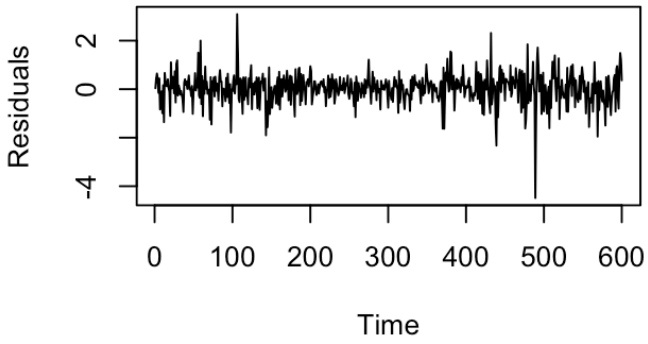


Fig- 2.9 ACF of the residuals for 2nd diff

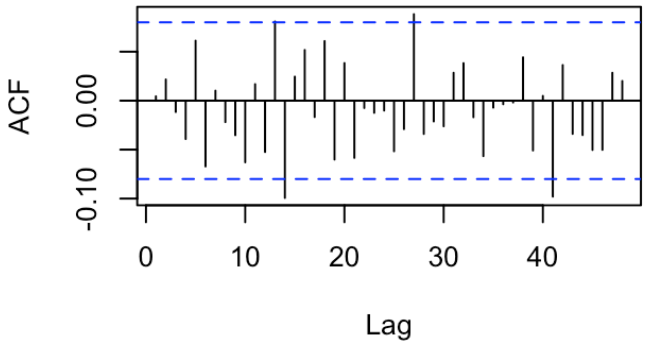
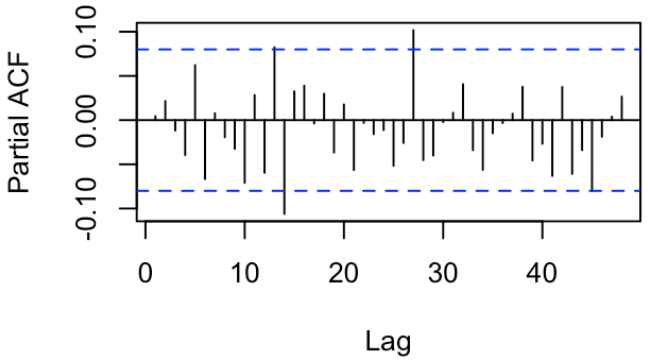


Fig- 3.0 PACF of the residuals



From the output above we can conclude that,from **Fig- 2.9** their is still **some** change in **variance** present in the model that is **due** to the **change point** present.

Furthermore,**Fig 3.0** and **Fig 3.1** of the ACF and the PACF of the series is **better** than the previous model that is **model1** as their is **less** significant **autocorrelation** at the first seasonal lag in both ACF and PACF and the second seasonal lag in PACF,

Since there are significant autocorrelation at the first seasonal lag in both ACF and PACF and the second seasonal lag in PACF, the **P = 2** from the PACF and **Q = 2** from the **ACF** plot.

As, the lambda, **suggest** for the **log** transformation, we will try and fit the log transformation series into the model, in order to **remove** the **significant** auto correlation which are present in the model

```
#It is coming change point,due to which series go different, variation and seasona
lity
#p=2,q=2

m3.landing = Arima(log(cola),order=c(0,0,0),seasonal=list(order=c(3,1,3), period=1
))
res.m3 = residuals(m3.landing);
par(mfrow=c(2,2))
plot(res.m3,xlab='Time',ylab='Residuals',main="Fig- 3.1 Time series plot of the re
siduals")
#Stablised Variance
acf(res.m3, lag.max = 36, main = " Fig- 3.2 ACF of the log residuals for model3")
pacf(res.m3,main = "Fig- 3.3 PACF of the log residuals for model3")
#McLeod.Li.test(y=res.m3, main="Fig- 3.6 McLeod-Li Test Statistics for cola")
```

Fig- 3.1 Time series plot of the residuals

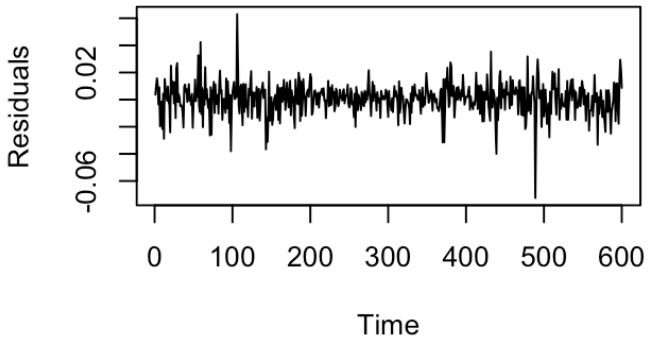


Fig- 3.2 ACF of the log residuals for mode

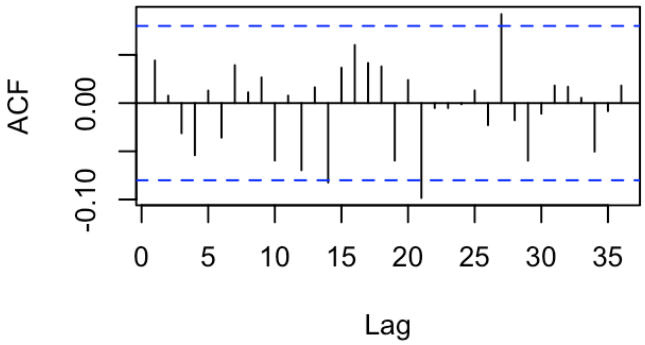
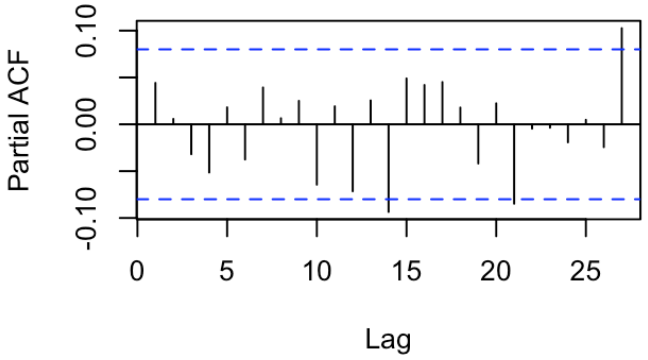


Fig- 3.3 PACF of the log residuals for mode

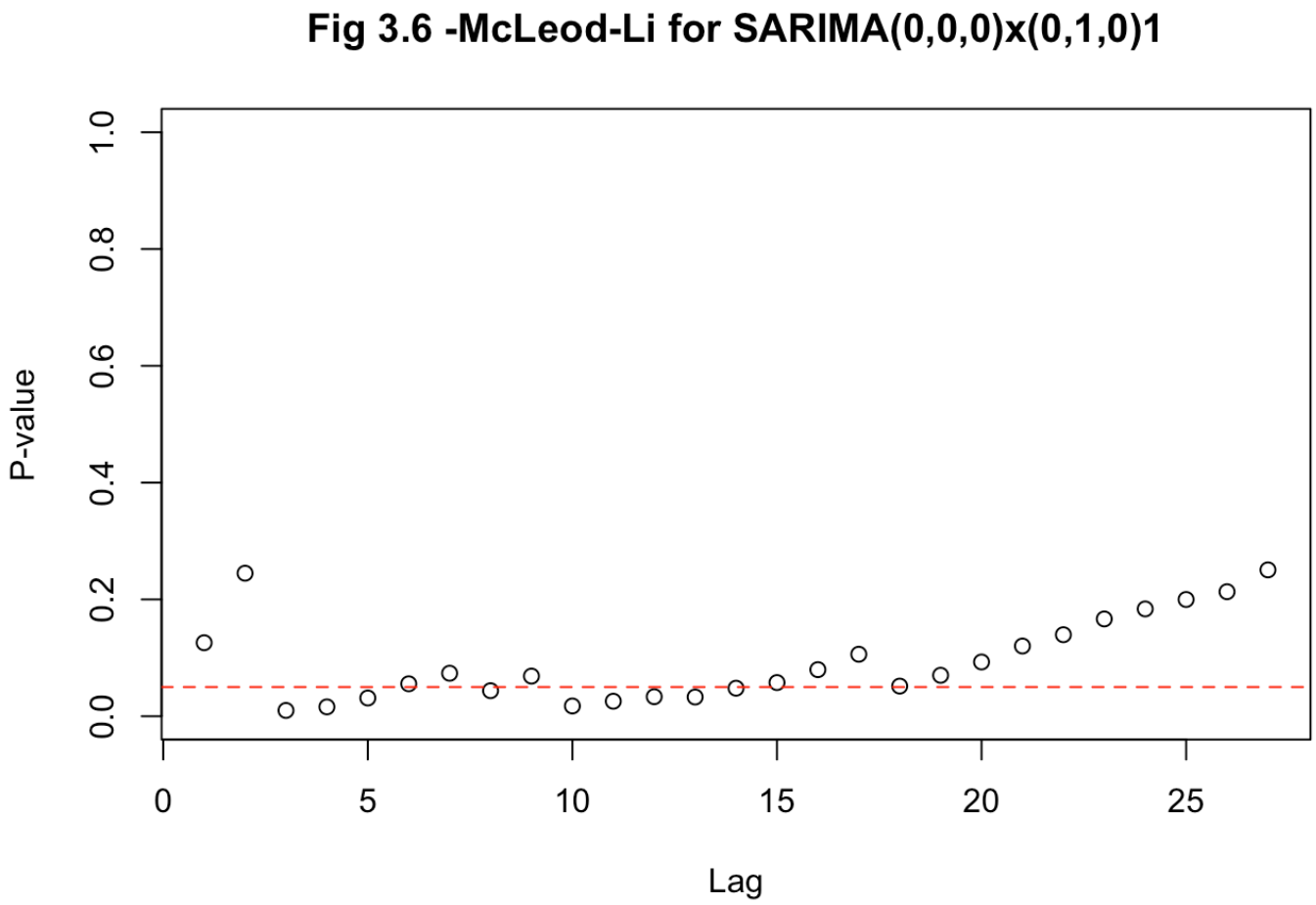


From the output above we can conclude that,from **Fig- 3.3** their is still **some** change in **variance** present in the model that is due to the change point present.But is **better** than the **model1** and **model2** as the limits for the **yaxis** has been significantly **decreased**

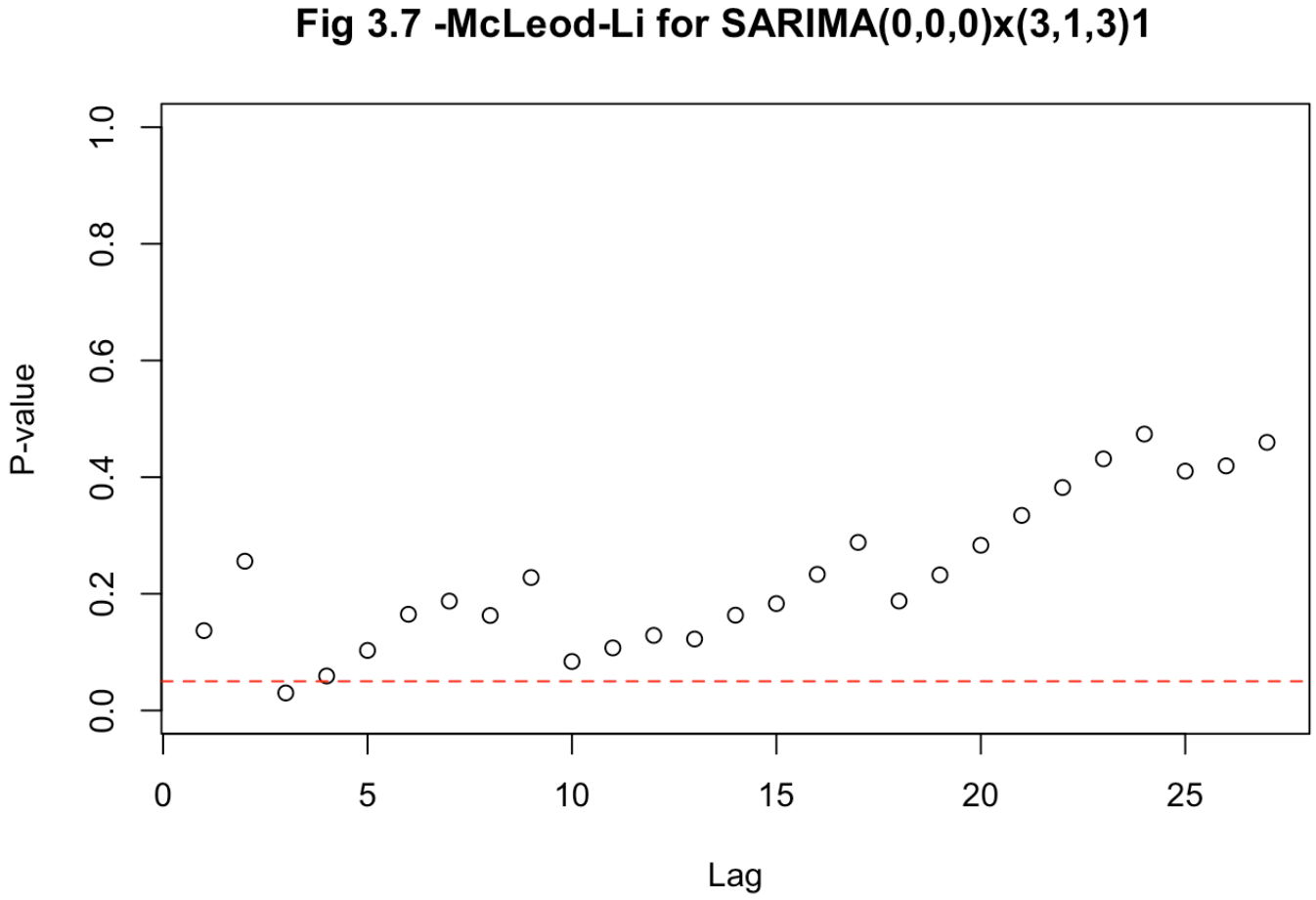
Furthermore**,Fig 3.4** and **Fig 3.5** of the ACF and the PACF of the series is **better** than the previous models as their only **one** significant **autocorrelation** at the first seasonal lag in both ACF and PACF and the second seasonal lag in PACF,

Comparing the **McLeod**-Li test among the **3** models

```
McLeod.Li.test(y=res.m1, main="Fig 3.6 -McLeod-Li for SARIMA(0,0,0)x(0,1,0)1")
```

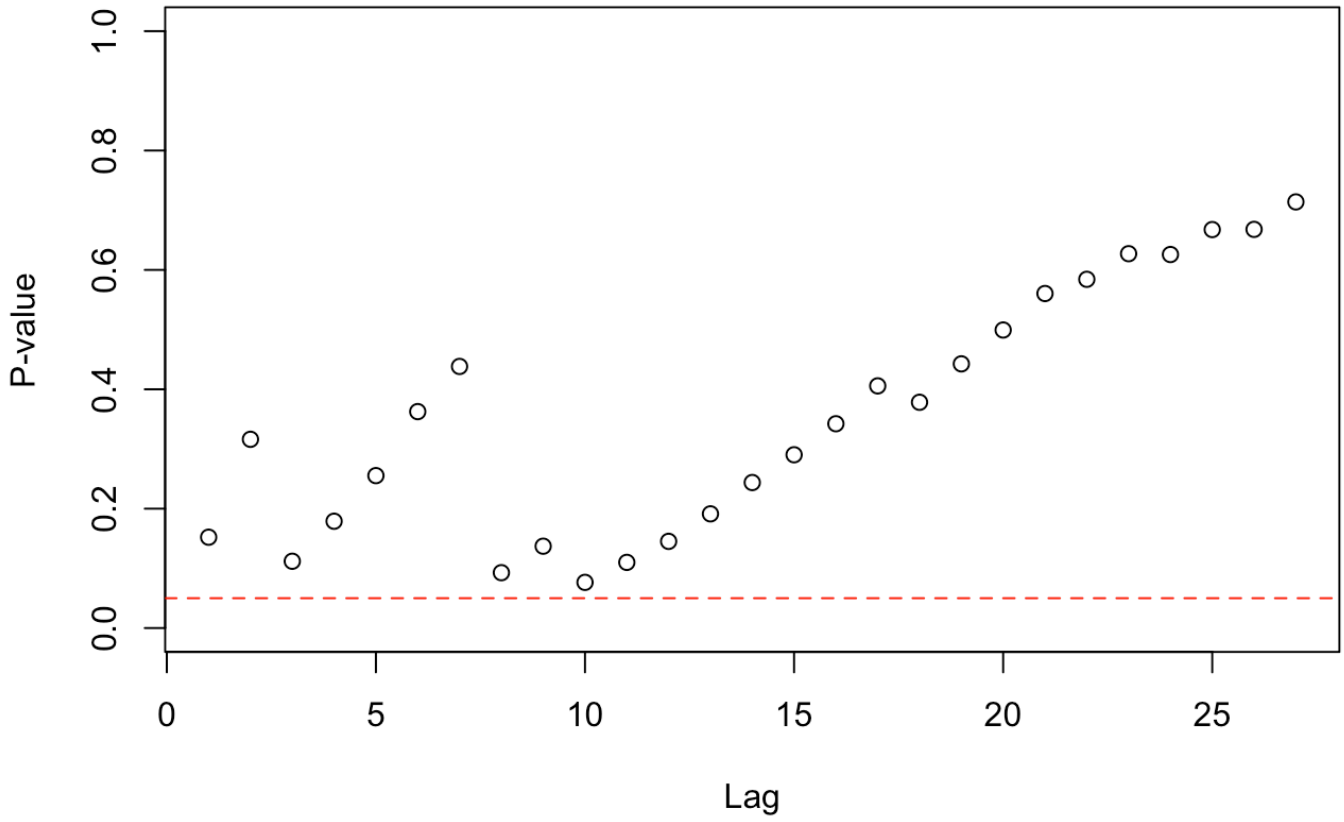



```
McLeod.Li.test(y=res.m2, main="Fig 3.7 -McLeod-Li for SARIMA(0,0,0)x(3,1,3)1  ")
```




```
McLeod.Li.test(y=res.m3, main="Fig 3.8 -McLeod-Li for Log series SARIMA(0,0,0)x(3,
1,3)1")
```

Fig 3.8 -McLeod-Li for Log series SARIMA(0,0,0)x(3,1,3)1



The above plots compares the amount of auto correlation left in the residuals of the model

From **Fig- 3.6** we can conclude that many points are under the reference line indicating still their is **valuable** information **left** in the residual of the model

Furthermore, from **Fig- 3.7** we can conclude that still there is some points are **under** the **reference** line indicating still their is valuable information **left** in the residual of the model.But is **better** than the previous model (**model1**)

Additionally, from **Fig- 3.8** we can conclude that **no** points are under the reference line indicating still their is **no** valuable information left in the residual of the model.Hence, it suggest it is the **best** model.

Hence, the significant auto correlation suggested by the ACF and the PACF in the **Fig 3.4** and **Fig 3.5**, are not important which has been **confirmed** by the McLeod Li Test Statistics.

Hence, the above comparison confirm **model3** is the **best** model. But,instead of the last model that captures all autocorrelation, we use one **previous** model to specify the set of **possible** models.

We will apply **shaprio**-wilk test to check the **normality** of the series

Here

H0: The series has a **normal** distribution.

HA: The series **does not** have a normal distribution.

We will apply the **ADF unit-root test** to test the existence of non-stationarity with this series.

Where,

H0: The time series has a **unit** root (is non-stationary)

HA: The time series **does not** have a unit root (is stationary)

For **pp test:**

H0: The time series is **non-stationary**

HA: The time series is **stationary**

For **kpss test:**

H0: The time series is **stationary**

HA: The time series is **non-stationary**

```
shapiro.test(res.m2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.m2
## W = 0.94654, p-value = 6.755e-14
```

```
adf.test(res.m2)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  res.m2
## Dickey-Fuller = -8.4864, Lag order = 8, p-value = 0.01
## alternative hypothesis: stationary
```

```
pp.test(res.m2)
```

```
##
##  Phillips-Perron Unit Root Test
##
## data:  res.m2
## Dickey-Fuller Z(alpha) = -595.42, Truncation lag parameter = 6, p-value
## = 0.01
## alternative hypothesis: stationary
```

```
kpss.test(res.m2)
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  res.m2
## KPSS Level = 0.078206, Truncation lag parameter = 6, p-value = 0.1
```

The p-value of the Shapiro test is **less** than 0.05, indicating that we have enough evidence to reject the normality hypothesis, which states that the series does not exhibit normal distribution.

The p-value obtained from an Augmented Dickey-Fuller (ADF) test is **less** than 0.05, it indicates that there is sufficient evidence to **reject** the null hypothesis that states the time series data **does not** have a unit root and is a **stationary** series.

The p-value obtained from a PP test is **less** than 0.05, it indicates that there is sufficient evidence to reject the null hypothesis that states the data is **stationary**.

The p-value obtained from a KPSS (Kwiatkowski-Phillips-Schmidt-Shin) test is **greater** than the significance level, it indicates that there is insufficient evidence to reject the null hypothesis of **stationarity**.

Model Specification

```
knitr::include_graphics("0Diff.png")
```

Model	Auto-Correlation in Residual	ML Signi	CSS Signi
m005ml	No	No	No
m200ml	Yes	No	No
m202ml	No	No	No
m504ml	No	No	No
m400ml	Yes	No	No
m000ml	Yes	No	No
m001ml	No	No	No

Initally, the order of **d=0**,however when we fit the models with d=0, their is **no CSS** significant model that can be used to forecast the Closing price of the Coca Cola Series.

Hence, we have **set d=1** and fitted the models

```
par(mfrow=c(1,2))
acf(res.m2, lag.max = 48, main = "Fig- 3.9 ACF of the residuals for model2")
pacf(res.m2, lag.max = 48, main = "Fig- 4.0 PACF of the residuals for model2")
```

Fig- 3.9 ACF of the residuals for mod

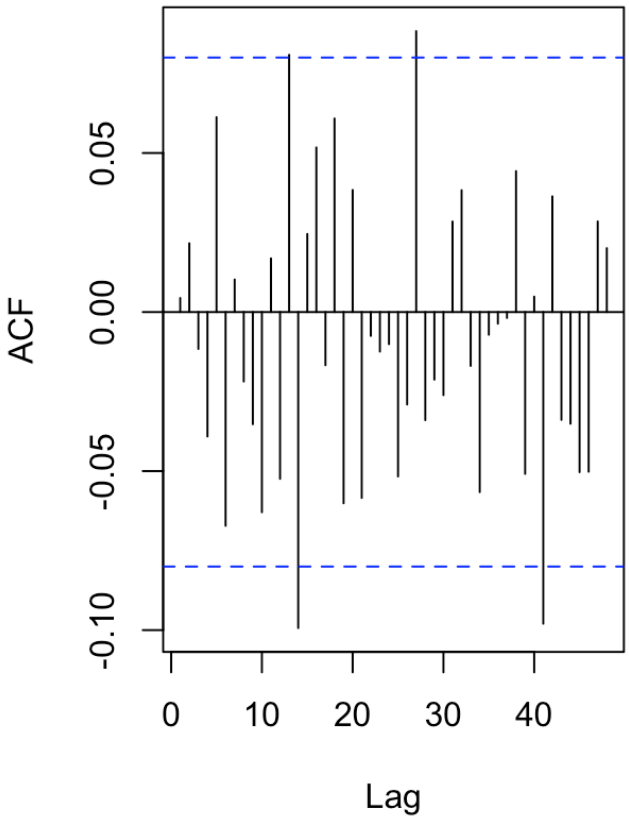
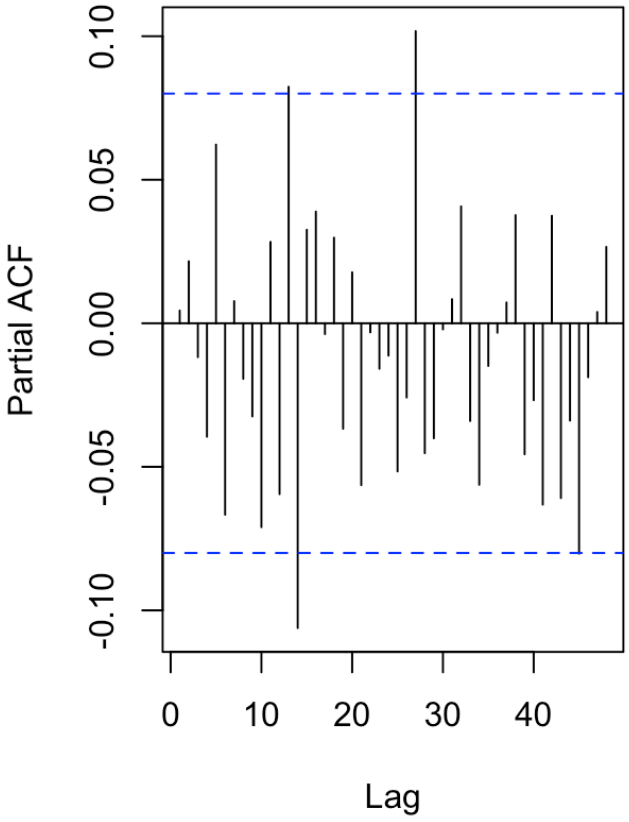


Fig- 4.0 PACF of the residuals for mo



So, possible models from here are:

SARIMA(2,1,2)x(3,1,3)1

```
eacf(res.m2)
```

```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 o o o o o o o o o o o  o  o  x
## 1 x o o o o o o o o o o  o  o  o
## 2 x x o o o o o o o o o  o  o  o
## 3 x x o o o o o o o o o  o  o  o
## 4 x x x x o o o o o o o  o  o  o
## 5 x x x x o o o o o o o  o  o  o
## 6 x x x x o x o o o o o  o  o  x
## 7 x x o x x x o o o o o  o  o  o
```

Our top left vertex is at **0,0** with neighbors **0,1** and **1,1**

The top left '**o'** symbol in EACF is located at the intersection of **AR = 0** and **MA = 0.** Then following the vertex downward, AR would be 1 and 1 as well. The set of possible models becomes **SARIMA(0,1,0)x(3,1,3)1 SARIMA(0,1,1)x(3,1,3)1 SARIMA(1,1,1)x(3,1,3)1**

Our updated **total models** are

SARIMA(2,1,2)x(3,1,3)1

SARIMA(0,1,0)x(3,1,3)1 , SARIMA(0,1,1)x(3,1,3)1

```
plot(armasubsets(y=res.m2, nar=5 , nma=5, y.name='p', ar.method='ols'))
mtext("Fig 4.1- BIC table/Armasubsets", side = 1, line = 1, cex = 1.5)
```

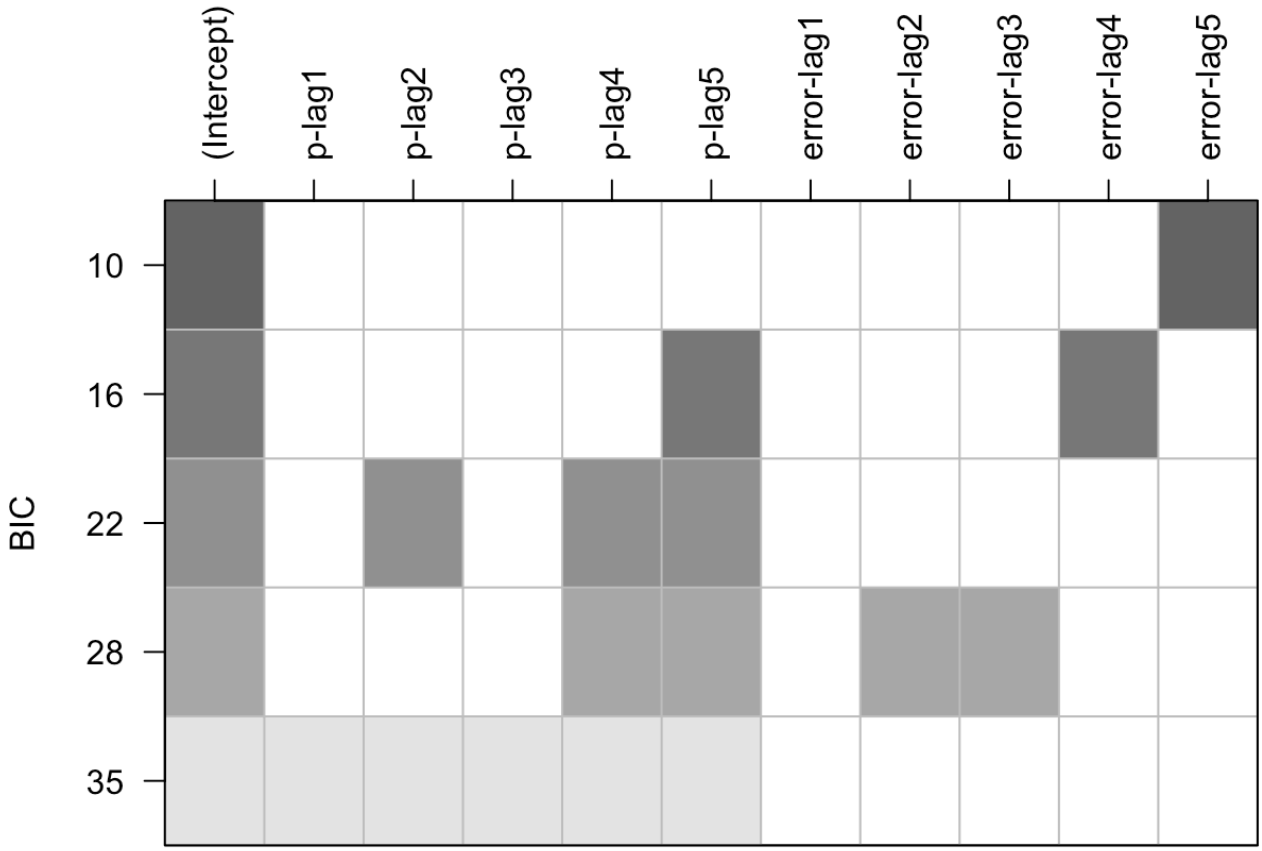


Fig 4.1- BIC table/Armasubsets

From the Fig-4.1, we read the models **SARIMA(2,1,0)x(3,1,3)1 SARIMA(5,1,4)x(3,1,3)1 1 SARIMA(0,1,5)x(3,1,3)1 SARIMA(4,1,0)x(3,1,3)1**

Our updated total models are

SARIMA(2,1,2)x(3,1,3)1 1

SARIMA(0,1,0)x(3,1,3)1 , SARIMA(0,1,1)x(3,1,3)1

SARIMA(2,1,0)x(3,1,3)1 , SARIMA(5,1,4)x(3,1,3)1 , SARIMA(0,1,5)x(3,1,3)1 , SARIMA(4,1,0)x(3,1,3)1

Now, we will fit these **7** models and find their parameter estimates and related significance tests.

Parameter Estimation

Here, we are using the **Conditional Sum of Squares (CSS) and Maximum Likelihood (ML)** method to test for the **significance** of the model.

```
#SARIMA(0,1,0)x(3,1,3)1
m010ml = Arima(cola,order=c(0,1,0), seasonal=list(order=c(3,1,3),period=1),method
= "ML")
coeftest(m010ml)
```

```
##
## z test of coefficients:
##
##        Estimate Std. Error z value  Pr(>|z|)
## sar1 -0.3289119  0.0408486 -8.0520 8.147e-16 ***
## sar2 -0.9662326        NaN     NaN       NaN
## sar3 -0.0012572  0.0398993 -0.0315    0.9749
## sma1 -0.6617814        NaN     NaN       NaN
## sma2  0.6570743        NaN     NaN       NaN
## sma3 -0.9950736        NaN     NaN       NaN
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m010css = Arima(cola,order=c(0,1,0), seasonal=list(order=c(3,1,3),period=1),method
= "CSS")
coeftest(m010css)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## sar1 -0.336127   0.024348 -13.8051 < 2.2e-16 ***
## sar2 -0.107608   0.012515  -8.5982 < 2.2e-16 ***
## sar3 -0.022729   0.027588  -0.8239 0.4099986
## sma1 -0.650409        NaN      NaN       NaN
## sma2 -0.215338   0.061923  -3.4775 0.0005061 ***
## sma3 -0.151527   0.041608  -3.6417 0.0002708 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
residual.analysis(model =m010css,010)
```

```
## Loading required package: lattice
```

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

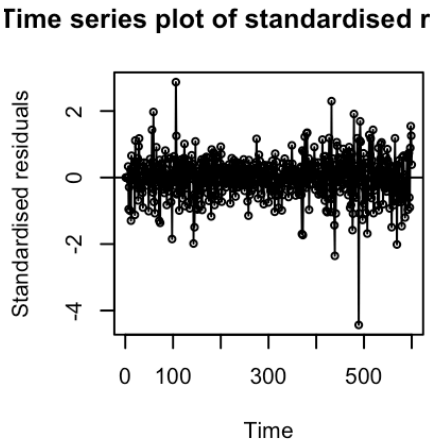
```
## Loading required package: ltsa
```

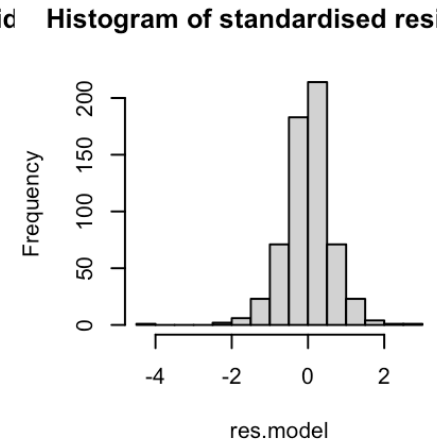
```
## Loading required package: bestglm
```

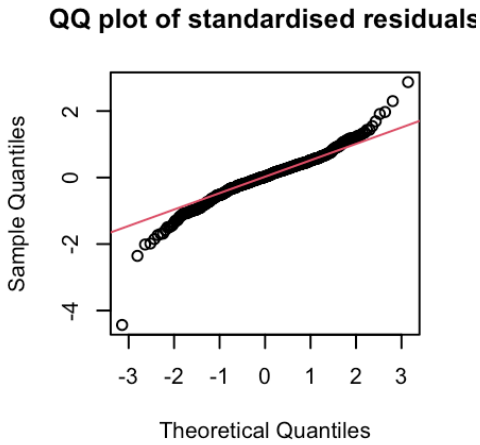
```
##
## Attaching package: 'FitAR'
```

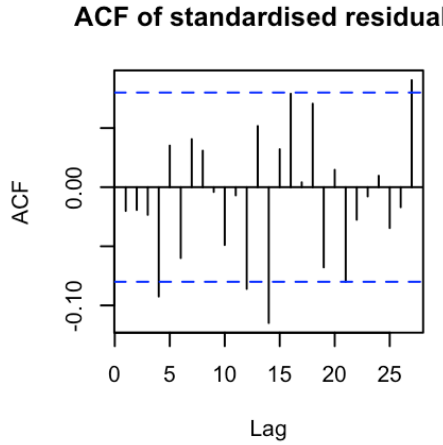
```
## The following object is masked from 'package:forecast':
##
##     BoxCox
```

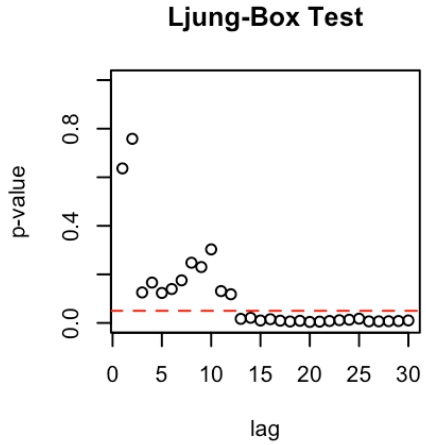
```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.95095, p-value = 3.178e-13
```











From the above results, we can see that for both ML and CSS model, not all coefficients have p-values less than 0.05, indicating that we can reject the null hypothesis and conclude that they not are statistically significant at the 5% level. The negative signs on sar1, sar2, sar3,sma2,sma2,sma3 indicate a negative impact on the dependent variable

From the histogram, qqplot and the shapiro test for the model SARIMA(0,1,0)x(3,1,3)1 , from the histogram we can determine that the distribution of the data is normal and their are no outliers as the value lies between -3 and 3,from the qqplot we can determine that from start and the end bits the points significantly deviated from the reference line indicating the series might not be normal

Additionally ,the shaipro-walk test **confirms** that the residuals are not **normally** distributed as the test has a p-value less than the significance of 5 percent, Hence, we cannot **reject** the null hypothesis and can conclude that the residuals are not **normally** distributed.

We can conclude from the ACF plot that there are significant lags .The Ljung-Box Statistic will confirm whether or not the significant values are important or not.

Hence, we can conclude from the Ljung-Box Statistic that there are many points that are significant at lag 5 as it lies below the confidence interval and the other p-values are greater than the 5% interval at multiple lags and can be concluded that there is large amount of auto correlation left in the residual.

Overall, we can say that the lags which was significant in the ACF plot is important as first lag has a p-value less than the 5% significance level.Also, the output is not very good because there is high significant auto correlation left in the residuals for the model of order SARIMA(0,1,0)x(3,1,3)1

```
#SARIMA(2,1,2)x(3,1,3)1  1
m212ml = Arima(cola,order=c(2,1,2), seasonal=list(order=c(3,1,3),period=1),method
= "ML")
coeftest(m212ml)
```

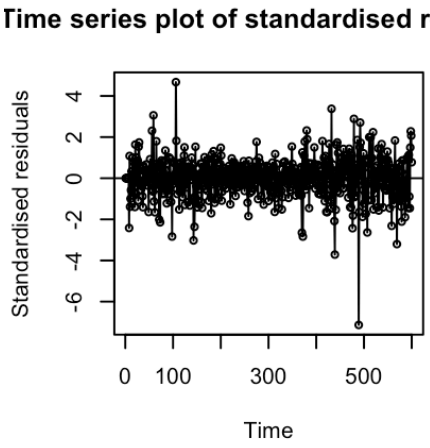
```
##
## z test of coefficients:
##
##        Estimate Std. Error  z value Pr(>|z|)
## ar1   0.1060965        NaN      NaN      NaN
## ar2  -0.1872213  0.1421056  -1.3175  0.18768
## ma1  -0.1268699  0.0728491  -1.7415  0.08159 .
## ma2  -0.8706349  0.0702435 -12.3945  < 2e-16 ***
## sar1 -0.2816374        NaN      NaN      NaN
## sar2  0.8279482  0.0051669 160.2397  < 2e-16 ***
## sar3  0.3818408        NaN      NaN      NaN
## sma1 -0.6874482        NaN      NaN      NaN
## sma2 -0.0747536  0.3213862  -0.2326  0.81607
## sma3 -0.2374848        NaN      NaN      NaN
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

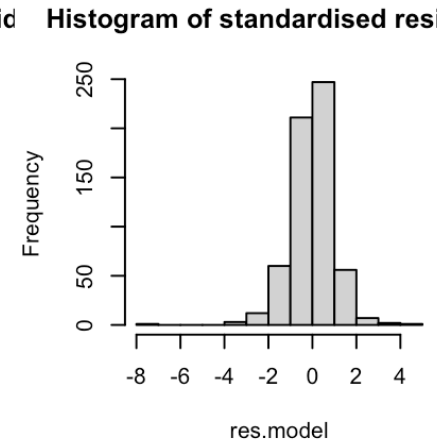
```
m212css = Arima(cola,order=c(2,1,2), seasonal=list(order=c(3,1,3),period=1),method
= "CSS")
coeftest(m212css)
```

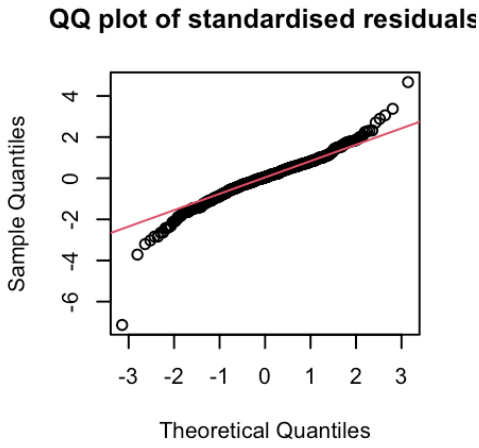
```
##
## z test of coefficients:
##
##        Estimate Std. Error  z value  Pr(>|z|)
## ar1  -0.7500007  0.0495054 -15.1499 < 2.2e-16 ***
## ar2  -0.4482406        NaN      NaN       NaN
## ma1   1.0204817  0.0862539  11.8311 < 2.2e-16 ***
## ma2   0.5063100  0.1085563   4.6640 3.101e-06 ***
## sar1 -1.2664519  0.0449732 -28.1601 < 2.2e-16 ***
## sar2 -0.7612060        NaN      NaN       NaN
## sar3 -0.0198624  0.0250873  -0.7917   0.42852
## sma1  0.0061718        NaN      NaN       NaN
## sma2 -0.3664136  0.1136189  -3.2249   0.00126 **
## sma3 -0.6711431  0.0659895 -10.1705 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

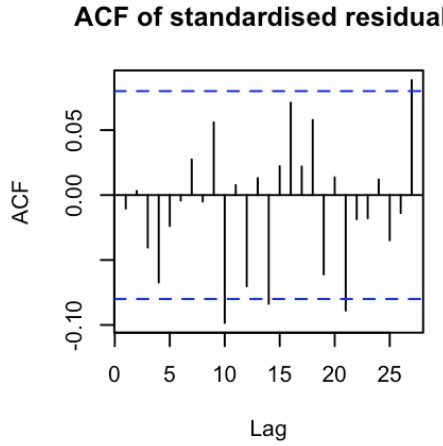
```
residual.analysis(model =m212css)
```

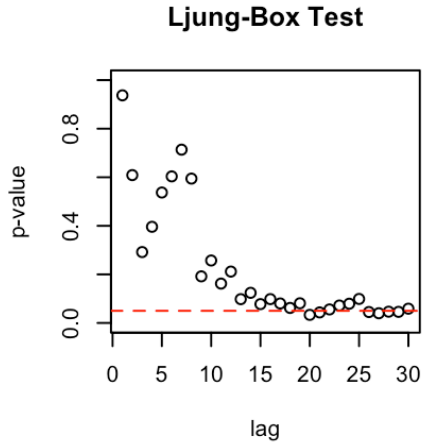
```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.95145, p-value = 3.806e-13
```











From the above results, we can see that for both ML and CSS model,we can see that not all coefficients have p-values less than 0.05, indicating that we can reject the null hypothesis and conclude that they not are statistically significant at the 5% level.

From the histogram, qqplot and the shapiro test for the model SARIMA(0,1,0)x(3,1,3)1 ,from the histogram we can determine that the distribution of the data is normal and their are no outliers as the value lies between -3 and 3,from the qqplot we can determine that from start and the end bits the points significantly deviated from the reference line indicating the series might not be normal Additionally ,the shaipro-walk test **confirms** that the residuals are not **normally** distributed as the test has a p-value less than the significance of 5 percent, Hence, we cannot **reject** the null hypothesis and can conclude that the residuals are not **normally** distributed.

We can conclude from the ACF plot that there are significant lags .The Ljung-Box Statistic will confirm whether or not the significant values are important or not.

Hence, we can conclude from the Ljung-Box Statistic that there are many points that are significant at lag 5 as it lies below the confidence interval and the other p-values are greater than the 5% interval at multiple lags and can be concluded that there is significant amount of auto correlation left in the residual.

Overall, we can say that the significant lags present in the ACF plot are important, and that there is still **significant** autocorrelation **left** in the residuals for the model of order SARIMA(2,1,2)x(3,1,3)1 1 1 because there are p-values are **partially on/below** the Confidence Interval.

```
#SARIMA(0,1,1)x(3,1,3)1
m011ml = Arima(cola,order=c(0,1,1), seasonal=list(order=c(3,1,3),period=1),method
= "ML")
coeftest(m011ml)
```

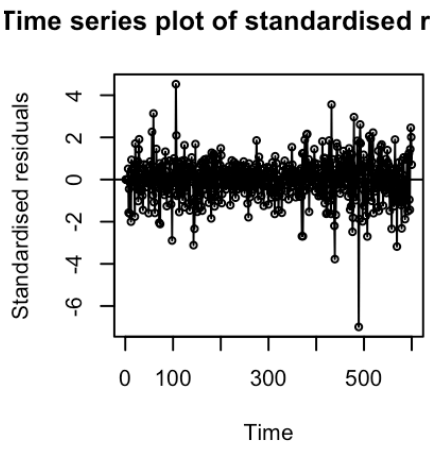
```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## ma1  -0.999964   0.029825 -33.5276 < 2.2e-16 ***
## sar1  0.690400   0.129904   5.3147 1.068e-07 ***
## sar2 -0.451429        NaN      NaN       NaN
## sar3  0.703416        NaN      NaN       NaN
## sma1 -0.681660   0.119160  -5.7205 1.062e-08 ***
## sma2  0.456564        NaN      NaN       NaN
## sma3 -0.774053        NaN      NaN       NaN
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

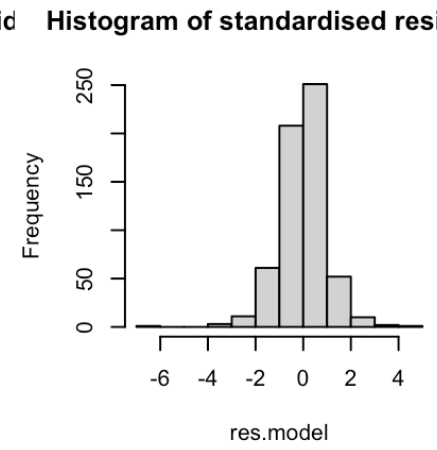
```
m011css = Arima(cola,order=c(0,1,1), seasonal=list(order=c(3,1,3),period=1),method
= "CSS")
coeftest(m011css)
```

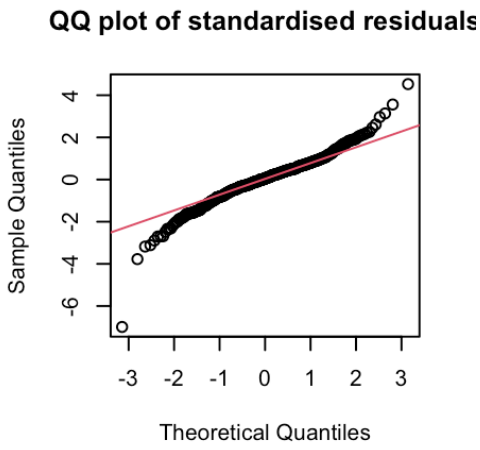
```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## ma1  -0.282021   0.040764  -6.9184 4.568e-12 ***
## sar1 -0.290690   0.020589 -14.1185 < 2.2e-16 ***
## sar2 -0.195763   0.024121  -8.1159 4.823e-16 ***
## sar3  0.067759   0.019888   3.4071 0.0006566 ***
## sma1 -0.440612        NaN      NaN       NaN
## sma2 -0.177859   0.049282  -3.6090 0.0003074 ***
## sma3 -0.403128   0.025272 -15.9514 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

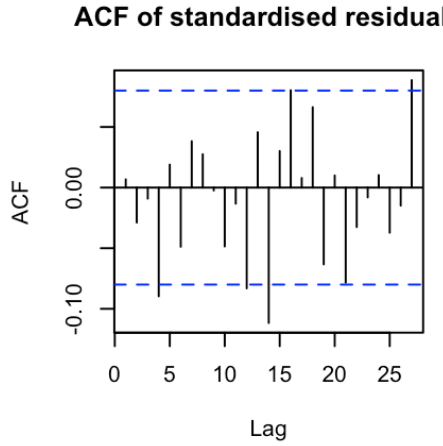
```
residual.analysis(model =m011css)
```

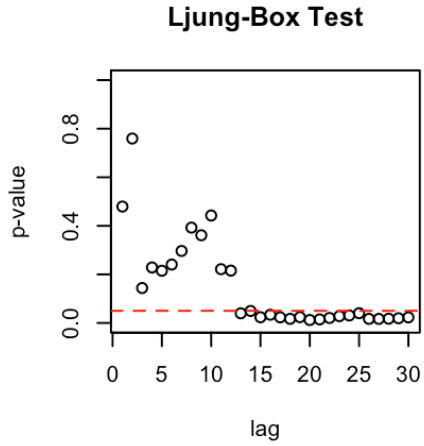
```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.95134, p-value = 3.669e-13
```











From the above results, we can see that for both ML and CSS model,we can see that almost all coefficients have p-values less than 0.05, indicating that we cannot reject the null hypothesis and conclude that they are statistically significant at the 5% level.

From the histogram, qqplot and the shapiro test for the model SARIMA(0,1,0)x(3,1,3)1 ,from the histogram we can determine that the distribution of the data is normal and their are no outliers as the value lies between -3 and 3,from the qqplot we can determine that from start and the end bits the points significantly deviated from the reference line indicating the series might not be normal

Additionally ,the shaipro-walk test **confirms** that the residuals are not **normally** distributed as the test has a p-value less than the significance of 5 percent, Hence, we cannot **reject** the null hypothesis and can conclude that the residuals are not **normally** distributed.

We can conclude from the ACF plot that there are significant lags .The Ljung-Box Statistic will confirm whether or not the significant values are important or not.

Hence, we can conclude from the Ljung-Box Statistic that there are many points that are significant at lag 5 as it lies below the confidence interval and the other p-values are greater than the 5% interval at multiple lags and can be concluded that there is significant amount of auto correlation left in the residual.

Overall, we can say that the lags which was significant in the ACF plot is important as first lag has a p-value less than the 5% significance level.Also, the output is not very good because there is high significant auto correlation left in the residuals for the model of order SARIMA(2,1,2)x(3,1,3)1 .

```
#SARIMA(2,1,0)x(3,1,3)1
m210ml = Arima(cola,order=c(2,1,0), seasonal=list(order=c(3,1,3),period=1),method
= "ML")
coeftest(m210ml)
```

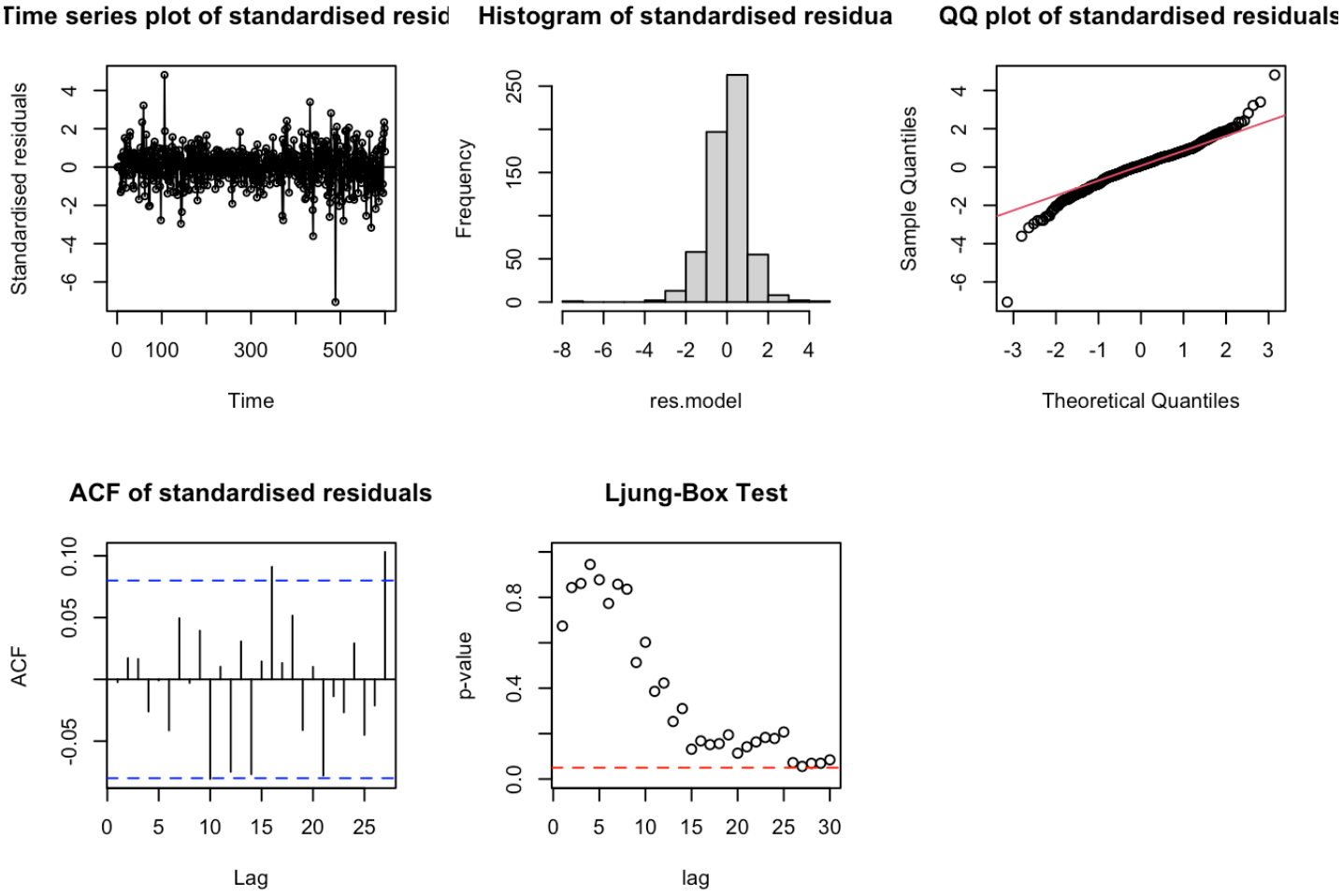
```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## ar1  -1.362174   0.078561 -17.3390 < 2.2e-16 ***
## ar2  -0.478465   0.071215  -6.7186 1.835e-11 ***
## sar1  1.164934        NaN      NaN       NaN
## sar2 -0.536304        NaN      NaN       NaN
## sar3  0.110893        NaN      NaN       NaN
## sma1 -0.783725        NaN      NaN       NaN
## sma2 -0.781690        NaN      NaN       NaN
## sma3  0.565417        NaN      NaN       NaN
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m210css = Arima(cola,order=c(2,1,0), seasonal=list(order=c(3,1,3),period=1),method
= "CSS")
coeftest(m210css)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## ar1   0.507902   0.083679   6.0696 1.282e-09 ***
## ar2  -0.237891   0.066845  -3.5588 0.0003725 ***
## sar1 -1.985109   0.089882 -22.0858 < 2.2e-16 ***
## sar2 -1.638825   0.130946 -12.5152 < 2.2e-16 ***
## sar3 -0.434713   0.072945  -5.9595 2.530e-09 ***
## sma1  0.506236   0.080478   6.2904 3.167e-10 ***
## sma2 -0.593705   0.025081 -23.6713 < 2.2e-16 ***
## sma3 -0.890354   0.077289 -11.5198 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
residual.analysis(model =m210css)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.95096, p-value = 3.185e-13
```



From the above results, we can see that for both ML and CSS model,we can see that almost all coefficients have p-values less than 0.05, indicating that we cannot reject the null hypothesis and conclude that they are statistically significant at the 5% level.

From the histogram, qqplot and the shapiro test for the model SARIMA(0,1,0)x(3,1,3)1 ,from the histogram we can determine that the distribution of the data is normal and their are no outliers as the value lies between -3 and 3,from the qqplot we can determine that from start and the end bits the points significantly deviated from the reference line indicating the series might not be normal

Additionally ,the shaipro-walk test **confirms** that the residuals are not **normally** distributed as the test has a p-value less than the significance of 5 percent, Hence, we cannot **reject** the null hypothesis and can conclude that the residuals are not **normally** distributed.

We can conclude from the ACF plot that there are significant lags .The Ljung-Box Statistic will confirm whether or not the significant values are important or not.

Hence, we can conclude from the Ljung-Box Statistic that there are few points that are significant at lag 5 as it lies below the confidence interval and the other p-values are greater than the 5% interval at multiple lags and can be concluded that there is few significant amount of auto correlation left in the residual.

Overall, we can say that the significant lags present in the ACF plot are important, and that there is still **significant** autocorrelation **left** in the residuals for the model of order SARIMA(2,1,2)x(3,1,3)1 because there are p-values are **partially on/below** the Confidence Interval.

```
#SARIMA(5,1,4)x(3,1,3)1
m514ml = Arima(cola,order=c(5,1,4), seasonal=list(order=c(3,1,3),period=1),method
= "ML")
coeftest(m514ml)
```

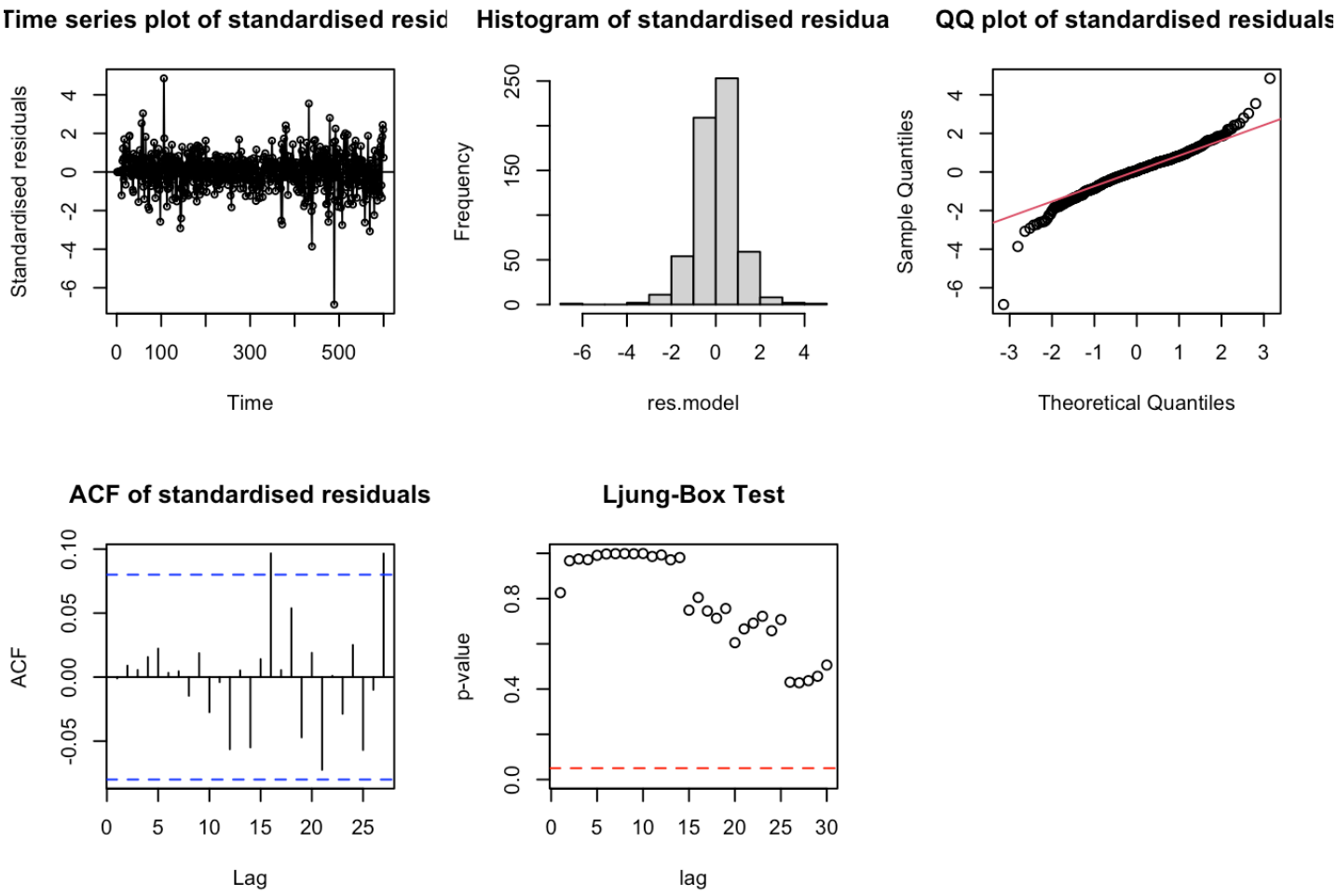
```
##
## z test of coefficients:
##
##        Estimate Std. Error  z value  Pr(>|z|)
## ar1   0.1801638        NaN      NaN       NaN
## ar2   0.0099335        NaN      NaN       NaN
## ar3  -0.5023854        NaN      NaN       NaN
## ar4   0.4571080        NaN      NaN       NaN
## ar5   0.3149180        NaN      NaN       NaN
## ma1  -0.7073306  0.0234885 -30.1139 < 2.2e-16 ***
## ma2   0.1600581  0.0317267   5.0449 4.538e-07 ***
## ma3   0.4447692  0.0339607  13.0966 < 2.2e-16 ***
## ma4  -0.8962072  0.0301146 -29.7599 < 2.2e-16 ***
## sar1  0.1847620  0.3555975   0.5196   0.60335
## sar2 -0.4611732        NaN      NaN       NaN
## sar3  0.1149896        NaN      NaN       NaN
## sma1 -0.6470405  0.3126883  -2.0693   0.03852 *
## sma2  0.2428994        NaN      NaN       NaN
## sma3 -0.4186843        NaN      NaN       NaN
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m514css = Arima(cola,order=c(5,1,4), seasonal=list(order=c(3,1,3),period=1),method
= "CSS")
coeftest(m514css)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
## ar1  -1.566680   0.235842 -6.6429 3.075e-11 ***
## ar2  -0.872343   0.409363 -2.1310  0.033091 *
## ar3  -0.140239   0.312663 -0.4485  0.653771
## ar4  -0.303640   0.124247 -2.4438  0.014532 *
## ar5  -0.252730   0.070675 -3.5759  0.000349 ***
## ma1   0.542842   0.190436  2.8505  0.004365 **
## ma2  -0.582546   0.117669 -4.9507 7.394e-07 ***
## ma3  -0.878582   0.123738 -7.1003 1.244e-12 ***
## ma4  -0.018685   0.183443 -0.1019  0.918868
## sar1 -0.353615   0.151945 -2.3273  0.019951 *
## sar2 -0.453250   0.163751 -2.7679  0.005641 **
## sar3 -0.540762   0.057247 -9.4461 < 2.2e-16 ***
## sma1  0.436128   0.094895  4.5959 4.309e-06 ***
## sma2  0.369911   0.062249  5.9425 2.808e-09 ***
## sma3  0.722084        NaN     NaN       NaN
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
residual.analysis(model =m514css)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.95254, p-value = 5.686e-13
```



From the above results,we can see that for both ML and CSS model, we can see that almost all coefficients have p-values less than 0.05, indicating that we can reject the null hypothesis and conclude that they are statistically significant at the 5% level.

From the histogram, qqplot and the shapiro test for the model SARIMA(0,1,0)x(3,1,3)1 ,from the histogram we can determine that the distribution of the data is normal and their are no outliers as the value lies between -3 and 3,from the qqplot we can determine that from start and the end bits the points significantly deviated from the reference line indicating the series might not be normal

Additionally ,the shaipro-walk test **confirms** that the residuals are not **normally** distributed as the test has a p-value less than the significance of 5 percent, Hence, we cannot **reject** the null hypothesis and can conclude that the residuals are not **normally** distributed.

We can conclude from the ACF plot that there are significant lags .The Ljung-Box Statistic will confirm whether or not the significant values are important or not.

Hence, we can conclude from the Ljung-Box Statistic that all p-values are **insignificant** because all of the p-values are greater than the 5% interval at multiple lags.

Overall, we can conclude that the lags, which were significant in the ACF plot, is not important in the residuals for the model SARIMA(2,1,2)x(3,1,3)1 because there are **no** p-values within the Confidence Interval.

```
#SARIMA(0,1,5)x(3,1,3)1
m015ml = Arima(cola,order=c(0,1,5), seasonal=list(order=c(3,1,3),period=1),method
= "ML")
coeftest(m015ml)
```

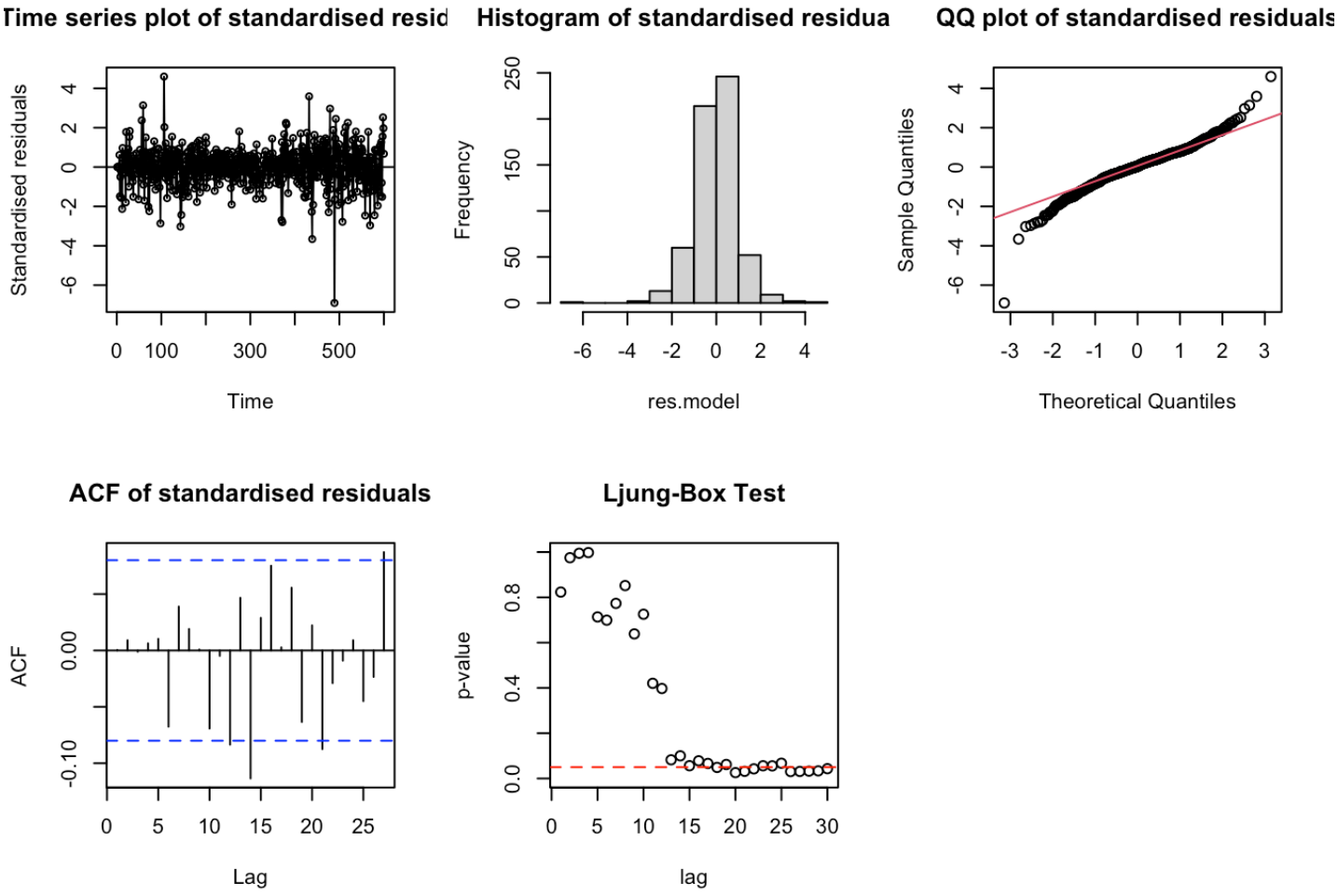
```
##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
## ma1  -0.028171   0.156750 -0.1797 0.8573741
## ma2   0.226863        NaN     NaN       NaN
## ma3   0.667654   0.082357  8.1069 5.194e-16 ***
## ma4  -0.385781   0.111863 -3.4487 0.0005633 ***
## ma5  -0.108612   0.105670 -1.0278 0.3040261
## sar1 -0.284750        NaN     NaN       NaN
## sar2 -0.498296        NaN     NaN       NaN
## sar3 -0.758625   0.090637 -8.3700 < 2.2e-16 ***
## sma1 -0.673756   0.194111 -3.4710 0.0005186 ***
## sma2 -0.022031   0.191969 -0.1148 0.9086318
## sma3 -0.304109   0.068628 -4.4313 9.368e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m015css = Arima(cola,order=c(0,1,5), seasonal=list(order=c(3,1,3),period=1),method
= "CSS")
coeftest(m015css)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value Pr(>|z|)
## ma1  -0.436196   0.174589 -2.4984  0.01248 *
## ma2  -0.282792   0.262427 -1.0776  0.28121
## ma3  -0.138819        NaN     NaN      NaN
## ma4  -0.195808        NaN     NaN      NaN
## ma5   0.031205   0.038480  0.8109  0.41740
## sar1 -0.264799        NaN     NaN      NaN
## sar2 -0.117347   0.013558 -8.6552  < 2e-16 ***
## sar3 -0.081022        NaN     NaN      NaN
## sma1 -0.304674   0.182238 -1.6719  0.09455 .
## sma2 -0.012119   0.200365 -0.0605  0.95177
## sma3 -0.047482   0.059594 -0.7968  0.42559
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
residual.analysis(model =m015css)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.95201, p-value = 4.683e-13
```



From the above results,we can see that for both ML and CSS model, we can see that almost all coefficients have p-values greater than 0.05, indicating that we cannot reject the null hypothesis and conclude that they are not statistically significant at the 5% level.

From the histogram, qqplot and the shapiro test for the model SARIMA(0,1,5)x(3,1,3)1 ,from the histogram we can determine that the distribution of the data is normal and their are no outliers as the value lies between -3 and 3,from the qqplot we can determine that from start and the end bits the points significantly deviated from the reference line indicating the series might not be normal

Additionally ,the shaipro-walk test **confirms** that the residuals are not **normally** distributed as the test has a p-value less than the significance of 5 percent, Hence, we cannot **reject** the null hypothesis and can conclude that the residuals are not **normally** distributed.

We can conclude from the ACF plot that there are significant lags .The Ljung-Box Statistic will confirm whether or not the significant values are important or not.

Hence, we can conclude from the Ljung-Box Statistic that there are few points that are significant at lag 5 as it lies below the confidence interval and the other p-values are greater than the 5% interval at multiple lags and can be concluded that there is few significant amount of auto correlation left in the residual.

Overall, we can say that the lags which was significant in the ACF plot is important as first lag has a p-value less than the 5% significance level.Also, the output is not very good because there is high significant auto correlation left in the residuals for the model of order ARIMA(2,1,2)x(3,1,3)1 .

```
#SARIMA(4,1,0)x(3,1,3)1

m410ml = Arima(cola,order=c(4,1,0), seasonal=list(order=c(3,1,3),period=1),method
= "ML")
coeftest(m410ml)
```

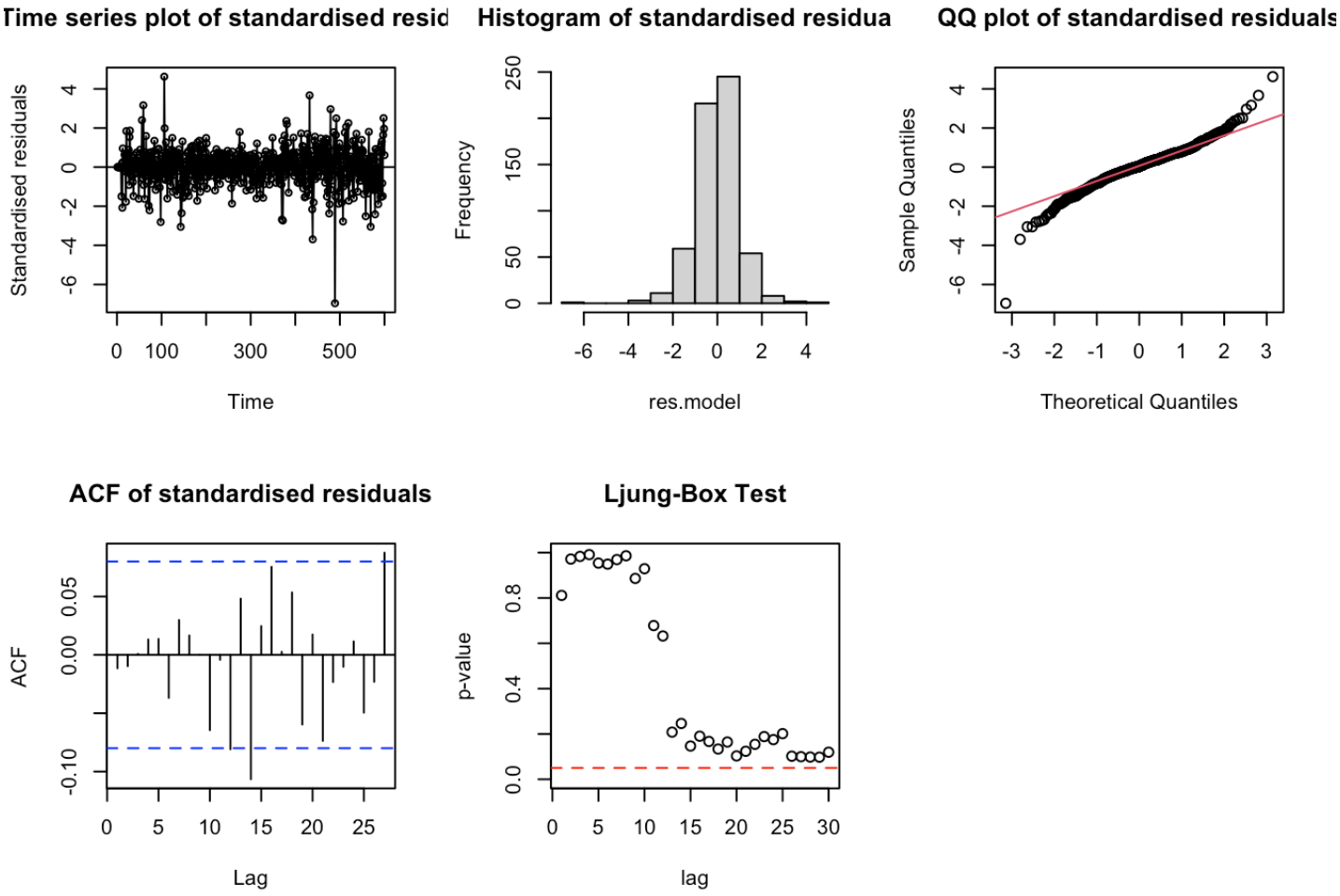
```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## ar1  -0.235980   0.469204  -0.5029    0.6150
## ar2   0.859283   0.136504   6.2949 3.076e-10 ***
## ar3   0.319793   0.385378   0.8298    0.4066
## ar4  -0.048475   0.155548  -0.3116    0.7553
## sar1  0.359164   0.492961   0.7286    0.4663
## sar2 -0.095324   0.284073  -0.3356    0.7372
## sar3 -0.054032   0.148050  -0.3650    0.7151
## sma1 -1.104995   0.078368 -14.1001 < 2.2e-16 ***
## sma2 -0.740401   0.152037  -4.8699 1.117e-06 ***
## sma3  0.845560   0.075879  11.1435 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m410css = Arima(cola,order=c(4,1,0), seasonal=list(order=c(3,1,3),period=1),method
= "CSS")
coeftest(m410css)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value Pr(>|z|)
## ar1  -0.177623        NaN     NaN      NaN
## ar2   0.282828        NaN     NaN      NaN
## ar3  -0.017213   0.071093 -0.2421   0.8087
## ar4  -0.120587        NaN     NaN      NaN
## sar1 -0.078106        NaN     NaN      NaN
## sar2 -0.047001        NaN     NaN      NaN
## sar3 -0.063476   0.068503 -0.9266   0.3541
## sma1 -0.731202        NaN     NaN      NaN
## sma2 -0.482082        NaN     NaN      NaN
## sma3  0.205817        NaN     NaN      NaN
```

```
residual.analysis(model =m410css)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.9511, p-value = 3.361e-13
```



From the above results, we can see that for both ML and CSS model,we can see that almost all coefficients have p-values greater than 0.05, indicating that we cannot reject the null hypothesis and conclude that they are not statistically significant at the 5% level.

From the histogram, qqplot and the shapiro test for the model SARIMA(0,1,5)x(3,1,3)1 ,from the histogram we can determine that the distribution of the data is normal and their are no outliers as the value lies between -3 and 3,from the qqplot we can determine that from start and the end bits the points significantly deviated from the reference line indicating the series might not be normal

Additionally ,the shaipro-walk test **confirms** that the residuals are not **normally** distributed as the test has a p-value less than the significance of 5 percent, Hence, we cannot **reject** the null hypothesis and can conclude that the residuals are not **normally** distributed.

We can conclude from the ACF plot that there are significant lags .The Ljung-Box Statistic will confirm whether or not the significant values are important or not.

Hence, we can conclude from the Ljung-Box Statistic that all p-values are **insignificant** because all of the p-values are greater than the 5% interval at multiple lags.

Overall, we can conclude that the lags, which were significant in the ACF plot, is not important in the residuals for the model SARIMA(4,1,0)x(3,1,3)1 because there are **no** p-values within the Confidence Interval.

We are using **AIC and BIC** as they are **selection** tools for the **goodness** of fit.

```
sc.AIC = AIC(m010ml,m011ml,m210ml,m514ml,m015ml,m410ml,m212ml)

sc.BIC = BIC(m010ml,m011ml,m210ml,m514ml,m015ml,m410ml,m212ml)

sort.score(sc.AIC, score = "aic")
```

```
##         df      AIC
## m015ml 12 1163.223
## m210ml  9 1166.872
## m514ml 16 1169.100
## m410ml 11 1169.504
## m212ml 11 1169.505
## m010ml  7 1169.585
## m011ml  8 1170.825
```

```
sort.score(sc.BIC, score = "bic")
```

```
##         df      BIC
## m010ml  7 1200.341
## m011ml  8 1205.973
## m210ml  9 1206.414
## m015ml 12 1215.946
## m410ml 11 1217.833
## m212ml 11 1217.835
## m514ml 16 1239.397
```

From both the AIC and BIC results it is clear that the model that has the **lowest** AIC is **SARIMA(0,1,5)x(3,1,3)1** , whereas model with **minimum** BIC is **SARIMA(0,1,0)x(3,1,3)1** .

Hence, we will check the error measures and the residual assumption for this models.

```
Sm010ml <- accuracy(m010ml)[1:7]
Sm011ml <- accuracy(m011ml)[1:7]
#Sm111ml <- accuracy(m111ml)[1:7]
Sm210ml <- accuracy(m210ml)[1:7]
Sm514ml <- accuracy(m514ml)[1:7]
Sm015ml <- accuracy(m015ml)[1:7]
Sm410ml <- accuracy(m410ml)[1:7]
Sm212ml <- accuracy(m212ml)[1:7]
#Sm112ml <- accuracy(m112ml)[1:7]
df.Smodels <- data.frame(
  rbind(Sm010ml, Sm011ml, Sm210ml,
        Sm514ml, Sm015ml, Sm410ml,Sm212ml)
)
colnames(df.Smodels) <- c("ME", "RMSE", "MAE", "MPE", "MAPE",
                          "MASE", "ACF1")

rownames(df.Smodels) <- c("SARIMA(0,1,0)x(3,1,3)1  ", "SARIMA(0,1,1)x(3,1,3)1  ","
SARIMA(2,1,0)x(3,1,3)1  ", "SARIMA(5,1,4)x(3,1,3)1  ", "SARIMA(0,1,5)x(3,1,3)1  ",
"SARIMA(4,1,0)x(3,1,3)1  ","SARIMA(2,1,2)x(3,1,3)1  1  ")
round(df.Smodels,  digits = 3)
```

```
##                                ME  RMSE   MAE    MPE  MAPE  MASE   ACF1
## SARIMA(0,1,0)x(3,1,3)1     -0.016 0.631 0.455 -0.036 0.851 0.992  0.002
## SARIMA(0,1,1)x(3,1,3)1     -0.026 0.628 0.454 -0.055 0.847 0.988 -0.014
## SARIMA(2,1,0)x(3,1,3)1     -0.018 0.628 0.455 -0.039 0.849 0.991 -0.002
## SARIMA(5,1,4)x(3,1,3)1     -0.022 0.620 0.450 -0.047 0.840 0.979 -0.002
## SARIMA(0,1,5)x(3,1,3)1     -0.018 0.622 0.452 -0.039 0.844 0.984 -0.001
## SARIMA(4,1,0)x(3,1,3)1     -0.024 0.627 0.455 -0.051 0.849 0.990 -0.003
## SARIMA(2,1,2)x(3,1,3)1  1  -0.025 0.625 0.454 -0.055 0.846 0.988 -0.003
```

```
knitr::include_graphics("Models1.png")
```

Model	AIC	BIC	Auto-Correlation in Residual	ML Signi	CSS Signi
m015ml	1163.223	1215.946	No	No	No
m210ml	1166.872	1206.414	Yes	No	Yes
m212ml	1169.505	1217.835	Yes	No	No
m514ml	1169.1	1239.397	No	No	Yes
m410ml	1169.504	1217.833	No	No	No
m010ml	1169.585	1200.341	Yes	No	No
m011ml	1170.825	1205.973	Yes	No	Yes

In terms of error measures, the results show that the models with the lowest RMSE and MAE is **SARIMA(5,1,4)x(3,1,3)1 1** as 0.620 and 0.450 respectively.

Model with the **lowest** ME are **SARIMA(0,1,0)x(3,1,3)1** , **SARIMA(2,1,0)x(3,1,3)1** and **SARIMA(0,1,5)x(3,1,3)1** but the model **SARIMA(0,1,0)x(3,1,3)1** is an **insignificant** CSS model with significant amount of auto correlation left in the residual, whereas model **SARIMA(0,1,5)x(3,1,3)1** is an **insignificant** CSS model.And, model **SARIMA(2,1,0)x(3,1,3)1** has a significant CSS model, but significant amount of auto correlation left in the residual.

The next best model with the lowest ME is are **SARIMA(5,1,4)x(3,1,3)1** , where **SARIMA(5,1,4)x(3,1,3)1 1** is an **significant** CSS model as coefficients have p-values less than the significance level.

As a result model **SARIMA(5,1,4)x(3,1,3)1 1** is the best model with **low** error measures and its residuals do not show any **no autocorrelation.**In addition to this, both AIC and BIC, as well as error measurements, indicate that **SARIMA(5,1,4)x(3,1,3)1 1** is the **optimal** model.

Overfitting Models

We will check if the overfitting models of **SARIMA(5,1,4)x(3,1,3)1** are significant or not.

We want the overfitting models to be insignificant as that would suggest that our best model with the lowest AIC and BIC captures the pattern or trend rather than noise.For this model, the overfitting models are **SARIMA(5,1,5)x(3,1,3)1** and **SARIMA(6,1,4)x(3,1,3)1 .**

We will fit these models to see if we get **significant** results for overfitting parameters

```
#Overfitting
m515ml = Arima(cola,order=c(5,1,5),seasonal=list(order=c(3,1,3), period=1),
               lambda = 0, method = "ML")
coeftest(m515ml)
```

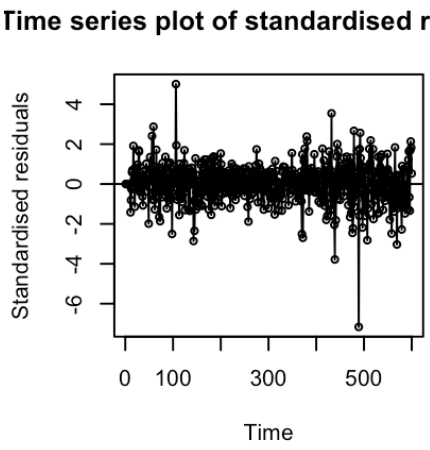
```
##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
## ar1   0.357219        NaN     NaN       NaN
## ar2  -0.124913   0.089713 -1.3924 0.1638117
## ar3   0.191852   0.069798  2.7487 0.0059834 **
## ar4   0.798163   0.106025  7.5281 5.150e-14 ***
## ar5  -0.268825        NaN     NaN       NaN
## ma1  -0.690610   0.083155 -8.3050 < 2.2e-16 ***
## ma2   0.060563   0.070631  0.8575 0.3911891
## ma3  -0.157203   0.042356 -3.7115 0.0002061 ***
## ma4  -0.802481   0.083585 -9.6008 < 2.2e-16 ***
## ma5   0.595642   0.084171  7.0766 1.478e-12 ***
## sar1 -0.709459   0.248486 -2.8551 0.0043020 **
## sar2 -0.068262   0.228002 -0.2994 0.7646414
## sar3  0.338676        NaN     NaN       NaN
## sma1  0.048578   0.287915  0.1687 0.8660138
## sma2 -0.420156   0.113045 -3.7167 0.0002018 ***
## sma3 -0.621588   0.302272 -2.0564 0.0397452 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

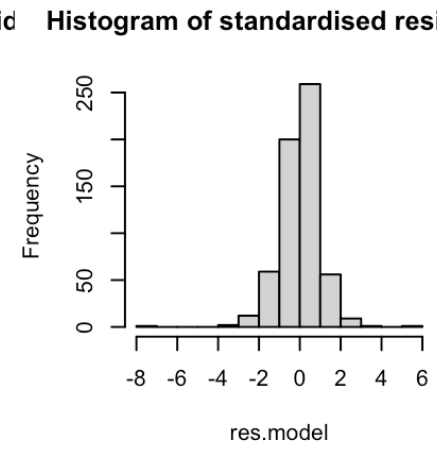
```
m515css = Arima(c,order=c(5,1,5),seasonal=list(order=c(3,1,3), period=1),
                method = "CSS")

coeftest(m515css)
```

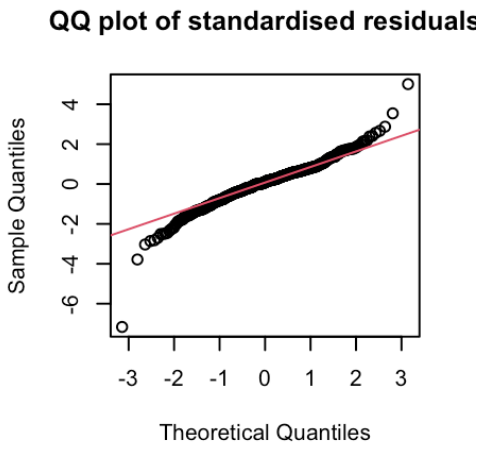
```
##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
## ar1  -0.675712   0.308014 -2.1938 0.0282517 *
## ar2   0.374434   0.021933 17.0714 < 2.2e-16 ***
## ar3   0.154325   0.126369  1.2212 0.2219991
## ar4  -0.938521        NaN     NaN       NaN
## ar5  -0.579846   0.194454 -2.9819 0.0028645 **
## ma1  -0.999483   0.134320 -7.4411 9.988e-14 ***
## ma2  -0.422369   0.164049 -2.5746 0.0100342 *
## ma3   0.544993   0.114343  4.7663 1.876e-06 ***
## ma4   0.714986   0.174979  4.0861 4.386e-05 ***
## ma5  -0.829249        NaN     NaN       NaN
## sar1 -0.803257   0.497261 -1.6154 0.1062318
## sar2 -0.115480   0.411245 -0.2808 0.7788587
## sar3 -0.075054   0.054909 -1.3669 0.1716649
## sma1  1.518733   0.437246  3.4734 0.0005139 ***
## sma2  0.766944   0.739084  1.0377 0.2994119
## sma3  0.193999   0.321786  0.6029 0.5465871
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

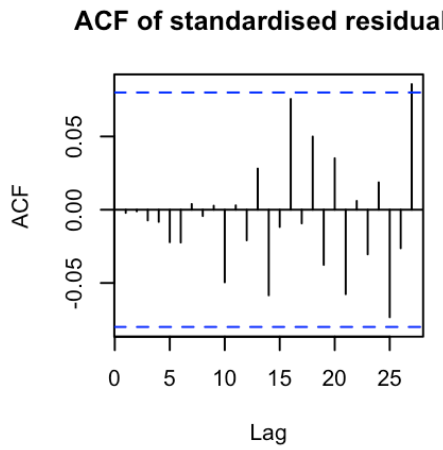
```
residual.analysis(model =m515css)
```

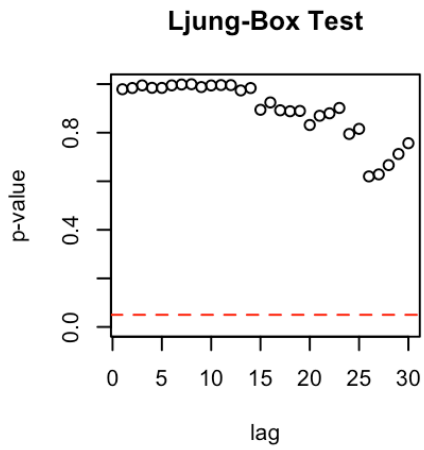
```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.94713, p-value = 8.25e-14
```












```
m614ml = Arima(cola,order=c(6,1,4),seasonal=list(order=c(3,1,3), period=1),
               lambda = 0, method = "ML")
coeftest(m614ml)
```

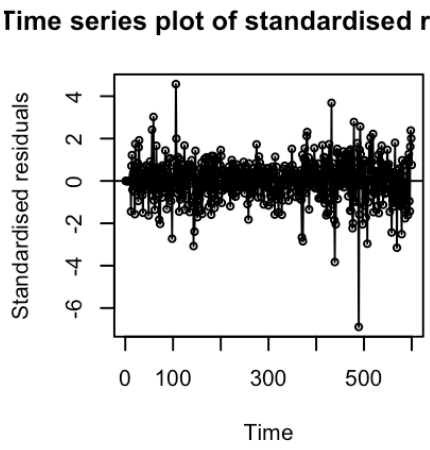
```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## ar1  -0.611193   0.065947  -9.2679 < 2.2e-16 ***
## ar2  -0.096234   0.083703  -1.1497  0.250263
## ar3   0.122680        NaN      NaN       NaN
## ar4   0.190628   0.072605   2.6256  0.008651 **
## ar5   0.611374        NaN      NaN       NaN
## ar6   0.626559        NaN      NaN       NaN
## ma1  -0.628295   0.033807 -18.5848 < 2.2e-16 ***
## ma2   0.055134        NaN      NaN       NaN
## ma3   0.429524   0.022331  19.2347 < 2.2e-16 ***
## ma4  -0.847031        NaN      NaN       NaN
## sar1  0.015240   0.075329   0.2023  0.839677
## sar2 -0.036274   0.077552  -0.4677  0.639972
## sar3 -0.091262   0.055665  -1.6395  0.101111
## sma1  0.228366        NaN      NaN       NaN
## sma2 -0.352776   0.042069  -8.3856 < 2.2e-16 ***
## sma3 -0.873180   0.035985 -24.2650 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

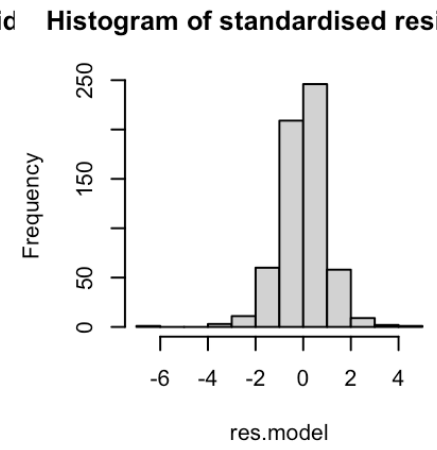
```
m614css = Arima(c,order=c(6,1,4),seasonal=list(order=c(3,1,3), period=1),
                method = "CSS")

coeftest(m614css)
```

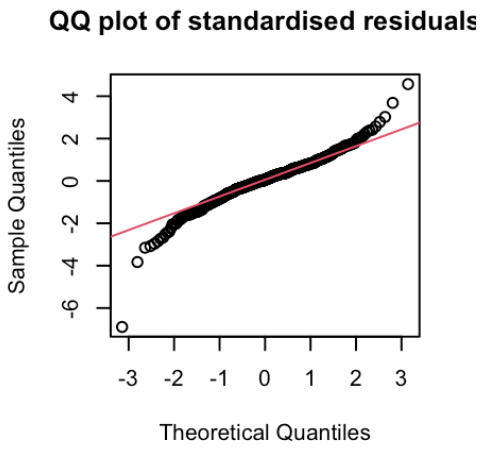
```
##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
## ar1  -0.411145   0.080712 -5.0940 3.506e-07 ***
## ar2  -0.078156   0.172925 -0.4520   0.65130
## ar3  -0.035675        NaN     NaN       NaN
## ar4   0.452594   0.174608  2.5921   0.00954 **
## ar5   0.080960   0.359454  0.2252   0.82180
## ar6  -0.079522   0.203687 -0.3904   0.69623
## ma1   0.058721   0.445622  0.1318   0.89516
## ma2  -0.190050        NaN     NaN       NaN
## ma3  -0.214330   0.422950 -0.5067   0.61233
## ma4  -0.682177        NaN     NaN       NaN
## sar1 -0.143913   0.098746 -1.4574   0.14501
## sar2  0.121256   0.275847  0.4396   0.66024
## sar3 -0.030793        NaN     NaN       NaN
## sma1 -0.485600   0.418952 -1.1591   0.24642
## sma2 -0.345752   0.041273 -8.3771 < 2.2e-16 ***
## sma3  0.078468   0.658456  0.1192   0.90514
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

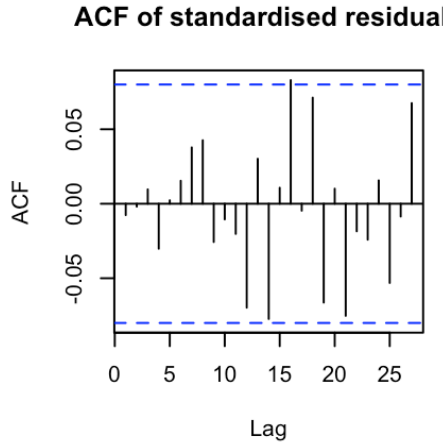
```
residual.analysis(model =m614css)
```

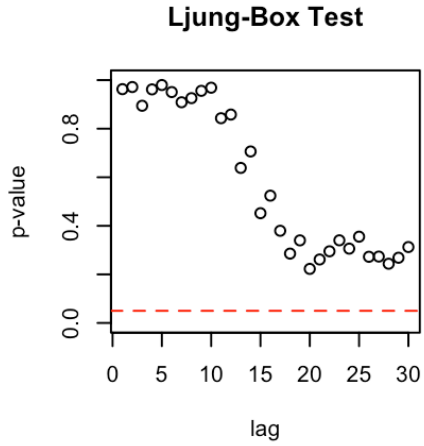
```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.95144, p-value = 3.805e-13
```











We can conclude that even after **increasing** and **decreasing** the value of the order by the 1 at a time,the overfitting model become **SARIMA(5,1,5)x(3,1,3)1** and **SARIMA(6,1,4)x(3,1,3)1 .**Additional AR() and MA() models both give **insignificant** coefficients

Therefore, these models are overfitting models and imply suitability of our original **SARIMA(5,1,4)x(3,1,3)1** model.

We are using **AIC and BIC** as they are **selection** tools for the **goodness** of fit.

```
sc.AIC = AIC(m010ml,m011ml,m210ml,m514ml,m015ml,m410ml,m212ml,m515ml,m614ml)

sc.BIC = BIC(m010ml,m011ml,m210ml,m514ml,m015ml,m410ml,m212ml,m515ml,m614ml)

knitr::include_graphics("Models.png")
```

Model	AIC	BIC	Auto-Correlation in Residual	ML Signi	CSS Signi
m015ml	1163.223	1215.946	No	No	No
m210ml	1166.872	1206.414	Yes	No	Yes
m212ml	1169.505	1217.835	Yes	No	No
m514ml	1169.1	1239.397	No	No	Yes
m410ml	1169.504	1217.833	No	No	No
m010ml	1169.585	1200.341	Yes	No	No
m011ml	1170.825	1205.973	Yes	No	Yes
m515ml(Overfitted)	-3595.39	-3520.699	No	No	No
m614ml(Overfitted)	-3598.776	-3524.085	No	No	No

The overfitting models have better AIC and BIC than the other models,though their is **no** significant auto-correlation left in the residuals, However, but these overfitting has **insignificant** CSS models as coefficients have **p**-values **more** than the significance level.

Hence, The **optimal** model is **SARIMA(5,1,4)x(3,1,3)1** .

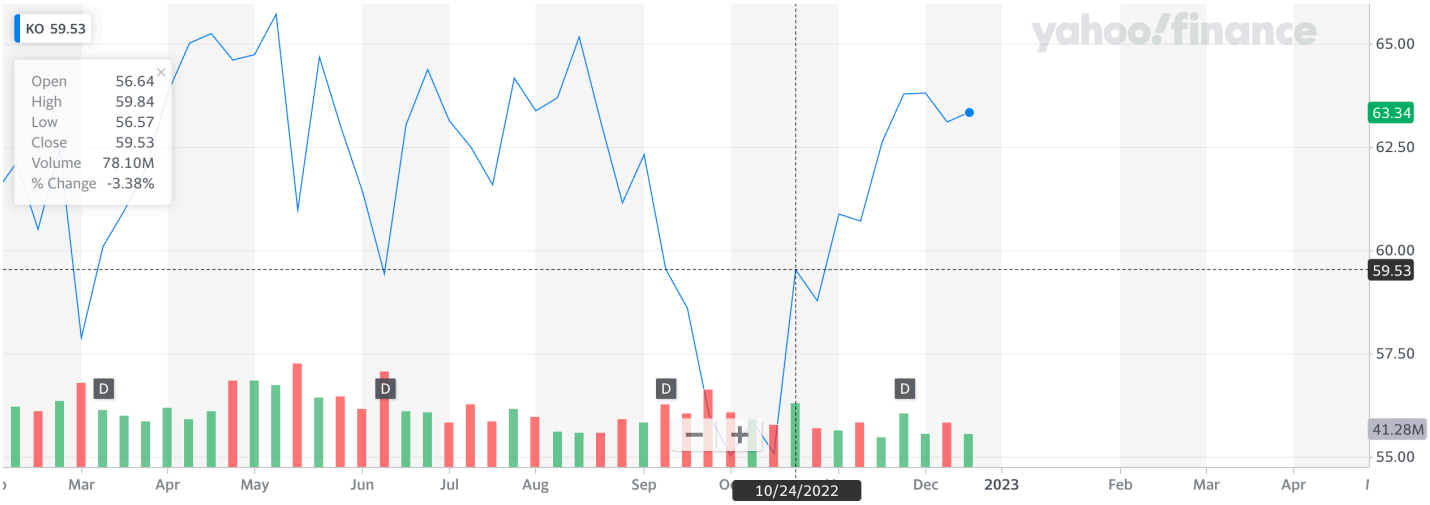
Prediction for the next 50 units that is 50 days.

We are using our CSS model for the prediction for the **next 50 days** as presenting a forecast for 10 days will not be largely represented in the graph.

Furthermore, we will compare the actual trend of series with the trend predicted by our model.

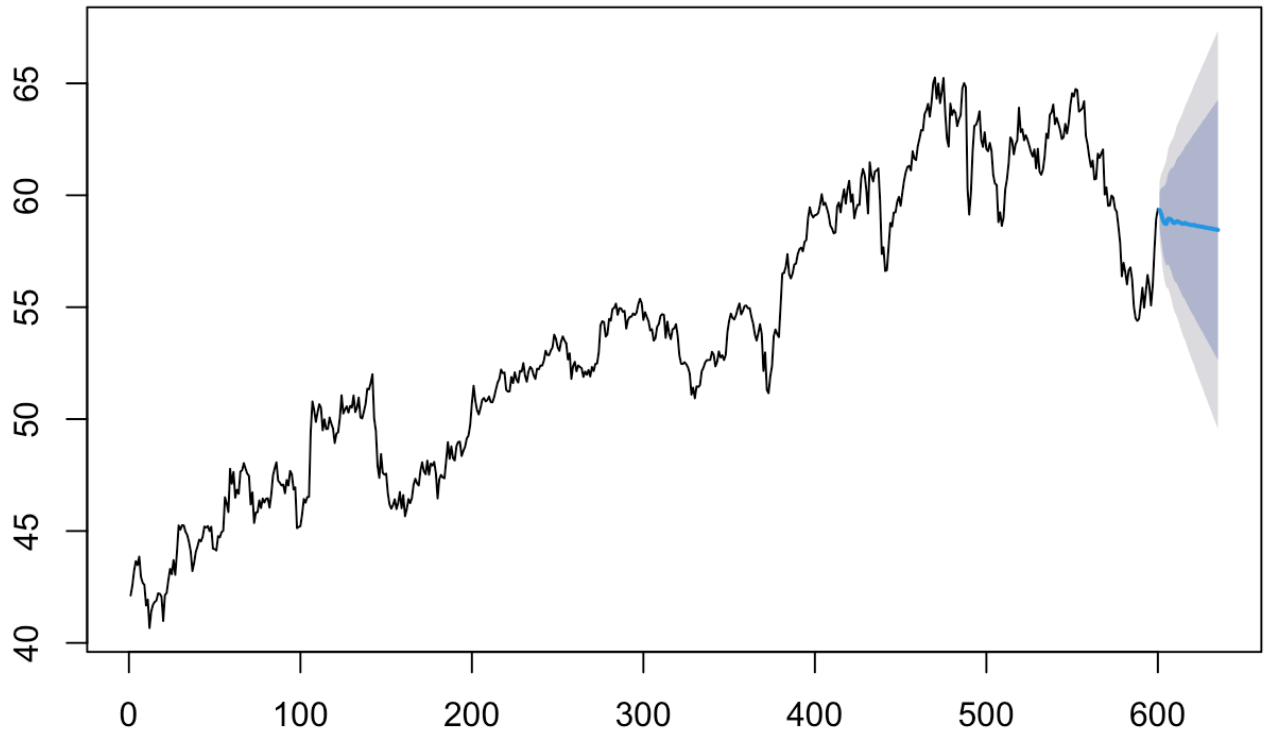
```
## [1] "Fig 4.2 Orginal Series of Coca Cola Data"
```

```
knitr::include_graphics("Original.png")
```




```
par(mfrow=c(1,1))
pred3=forecast(m514css, h = 35)
plot(pred3,main="Fig- 4.3 Forecasts from SARIMA model (5,1,4)x(3,1,3)1  ")
```

Fig- 4.3 Forecasts from SARIMA model (5,1,4)x(3,1,3)1



From the **Fig: 4.2** and **Fig: 4.3**, we can determine that the **SARIMA(5,1,4)x(3,1,3)1** has followed the trend of the original series **very** closely.

Conclusion

We analysed the Coca Cola series by using various analysis. As the data was collected **daily**, we first **transformed** the raw data into a time series with a frequency of **1**.In the next step we have checked the correlation between the consecutive time points which indicated that they are highly correlated to reach other. And, in-order to check the trend, seasonality and the change point in the series we have plotted the time series plot of the coca cola series, and presence of **seasonality** was **confirmed** from the plot. we have analysed **ACF ,PACF, McLeod test** of the series to check for the presence of **trend** ,**seasonality** and the auto correlation left in the residuals that may lead to **non-stationarity** series.We confirmed the existence of non-stationarity by performing an **ADF** test, and Shapiro wilk test which indicated the series **does not** exhibits normal distribution.

Then we have fitted a **first differenced** model with the original series to **capture** the order of the **seasonal** part of the model and a frequency.After this, we have added the order of the seasonal part and again fitted the model, as the their were still **some** significant **auto correlation** left in the residuals, we have used the transformed series using **log** transformation as the optimum value of **lambda** was **zero** to make the series

stationary and the result were **stable** as their was **no** significant **auto correlation** present in the model.hence,we have used the **previous** model instead of the last model that **captures** all **autocorrelation** to specify the set of possible models.

We used various methods such as **ACF, PACF, EACF, Armasubsets , AIC and BIC** to identify **7** possible models. Then, we performed parameter estimation and diagnostic checking for each model. Also, we have taken the first **differenced** series and have set order of **d=1** as their was **no** significant **CSS model** which can be used to forecast when it was **set** to **zero**.Also, we have only selected the models which were having significant coefficients in their CSS model as the data were **not normally** distributed after the **transformation** and the **differencing**.Then, we selected the model with **minimum AIC and BIC** value .We also checked the significance of the **overfitting** model to avoid model selection bias. We eliminated models with **auto-correlation** in their residuals to avoid **inaccurate** predictions.

We **compared** error measures among all possible models and disregarded models with those with remaining autocorrelation in their residuals. Hence, the best model from both the error measures and the model with lowest AIC, BIC was **SARIMA(5,1,4)x(3,1,3)1** which has almost all coefficents significant with p-values less than the 5% significance level and having **insignificant** coefficients in it's over-fitting model indicates that our model is **not** capturing **noise**.Also, our final model confirms the assumption that was made from **Figure 1.0** that p is greater than q, as **p=5** and **p=4** for our final model which is **SARIMA(5,1,4)x(3,1,3)1** .

After comparing the actual trend of the series with the forecasted trend by our model, we can conclude that our model were **able** to predict the actual trend of the series **very closely** and **accurately** as both grpah shows a decrease trend.

Reference

- Yahoo Finance. (n.d.). Coca-Cola Company (The) (KO) Historical Data. Retrieved from **https://finance.yahoo.com/quote/KO/history?p=KO** (https://finance.yahoo.com/quote/KO/history?p=KO)