An Assignment on Time Series Analysis and Forecasting

Course Name: Time Series Analysis and Forecasting

Course Code: PM-ASDS10

Submitted to:

Dr. Rumana Rois

Professor

Department of Statistics, JU

Submitted by:

Mohammad Saiduzzaman Sayed

ID: 20215063

Batch: 5th

Sec: A

Professional Masters in

Applied Statistics and Data Science (PM-ASDS)

JAHANGIRNAGAR UNIVERSITY

Assignment 1

Selected the R Dataset:

```
63%%6

## [1] 3
```

My ID is 20215063 and last two digit is 63. So, 63 %% 6 = 3. Hence, our given dataset is UKgas.

Dataset Details:

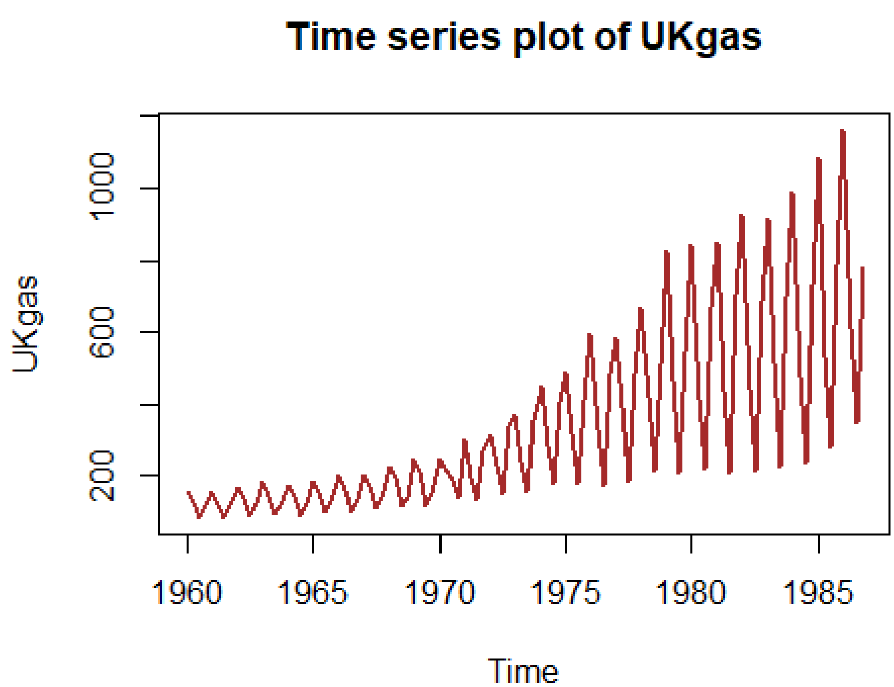
UK Quarterly Gas Consumption

Description: Quarterly UK gas consumption from 1960Q1 to 1986Q4, in millions of therms.
Usage: UKgae
Format: A quarterly time series of length 108.

1. Plots

Time-Plot:

```
plot(UKgas, col='brown', lwd=2, main='Time series plot of UKgas')
```



Comment on Time-plot:

The above time-plot describes some main features of data which is given below:

Trends: The above time series plot shows a clear upward trend.

Seasonal pattern: These data show a multiplicative seasonal pattern because the pattern repeats every 4 quarters. The frequency is 4, and therefore period is quarterly, so a seasonal component is present.

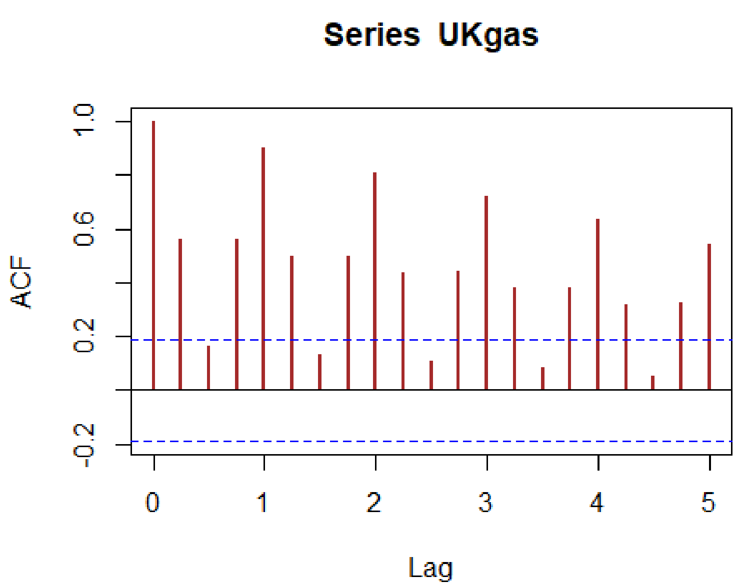
Sharp Changes: The above time series plot shows that there is no sharp change in our data.

Outliers: Following the above time series plot, we can assume that there are some outliers in the data. We can check outliers with the help of forecast package and tsoutliers function in R.

```
tsoutliers(UKgas)

## $index
##  [1]  44  57  65  69  73  77  79  81  83  84  85  87  88  89  91  93  95
97  98
## [20]  99 101 103 105 107
##
## $replacements
##  [1] 301.1369 358.9973 404.5556 475.4278 492.0752 531.2751 394.3690 524.91
69
##  [9] 383.0432 565.7420 547.0516 402.9917 585.6832 559.4504 453.9700 643.24
21
## [17] 495.6343 670.9894 559.0268 534.9950 676.2419 551.6219 744.3583 588.41
11
```

ACF(AutoCorrelation Function) Plot:

```
acf(UKgas, col='brown', lwd=2)
```

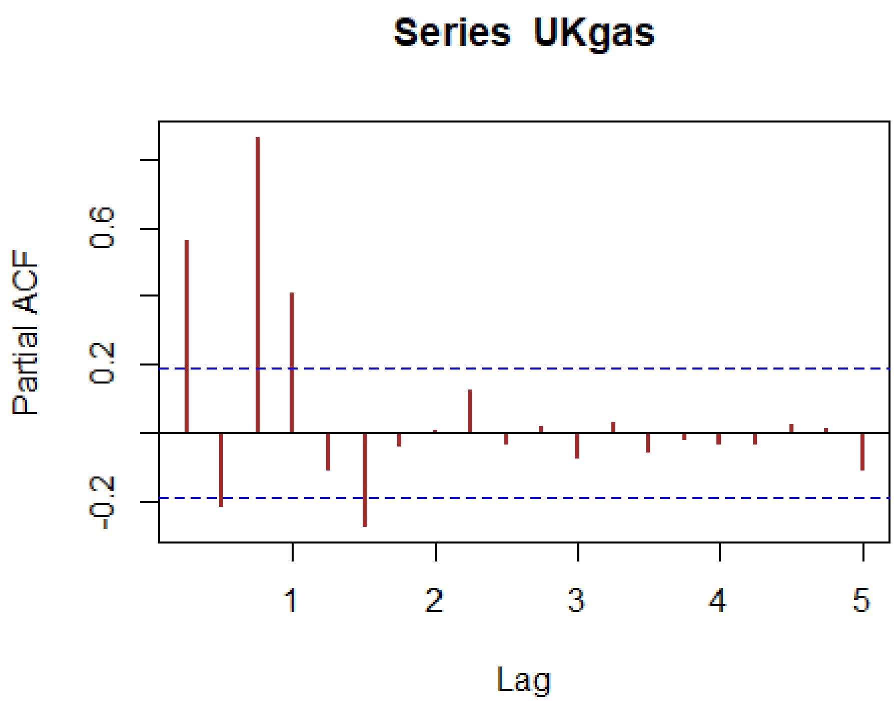


Comment on ACF plot:

ACF plot shows that first two significant spike and the coefficient is high at lag 4,8,12. So, here we can say the MA(q)=2.

PACF(Partial AutoCorrelation Function) Plot:

```
pacf(UKgas, col='brown', lwd=2)
```

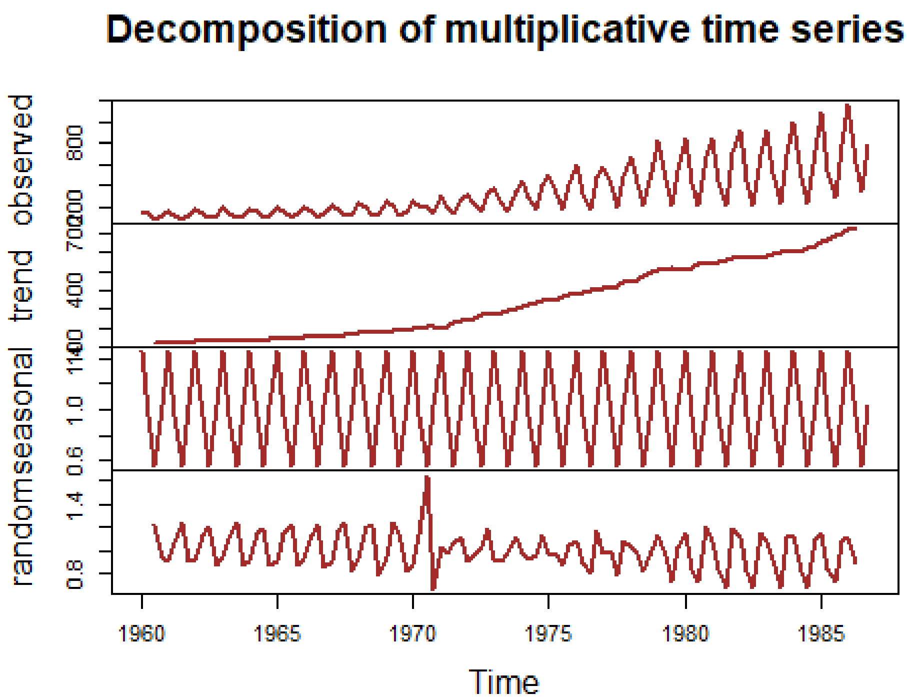


Comment on PACF plot:

PACF plot shows that first three significant spike and the coefficient is high at lag 4,8,12. So, here we can say the AR(p)=3.

Decomposition Plot:

```
plot(decompose(UKgas, type = c("multiplicative")), col='brown', lwd=2)
```

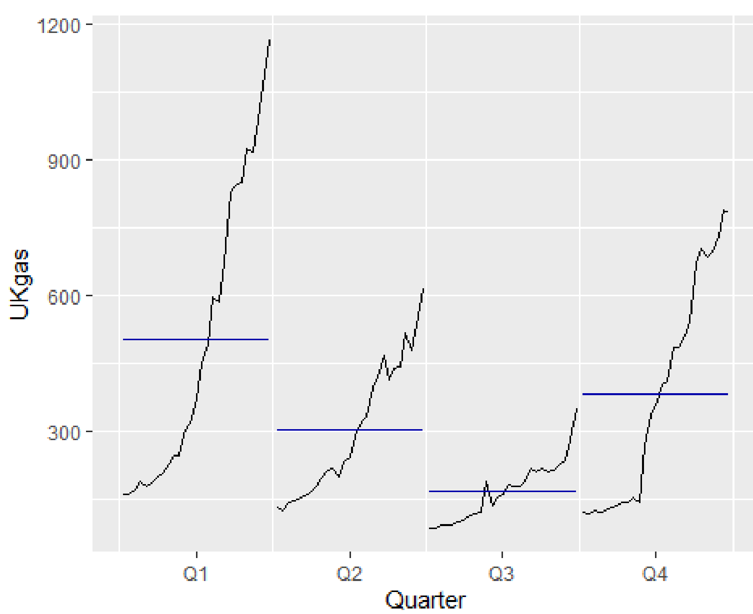


Comment on Decomposition plot:

Decomposition plot shows that clear upward trend, multiplicative seasonality and random fluctuation.

Seasonal-Subseries Plot:

```
ggsubseriesplot(UKgas)
```



Comment on seasonal subseries plot:

A seasonal subseries plot is used to determine if there is significant seasonality in a time series. For our quarterly data, all the Q1 values are plotted, then all the Q2 values, and so on. So, we can clearly see that from the above seasonal subseries data, there is seasonality present in UKgas data.

Estimated Outliers:

```
new_UKgas<-tsclean(UKgas)
```

Here, we estimated outliers with tsclean function in forecast package.

Splitting the data into training (70%) and test (30%) data sets:

Training Data:

```
traindata<-ts(new_UKgas[0:75], frequency = 4, start=c(1960,1))
```

Test Data:

```
testdata<-ts(new_UKgas[76:length(UKgas)], frequency = 4, start=c(1978,4))
```

2. Stationary test

Checking Stationary:

Augmented Dickey-Fuller (ADF) test and The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test can be used to test for stationary. These tests are available in tseries package with the function adf.test() for the ADF test and kpss.test() for the KPSS test.

Augmented Dickey-Fuller Test:

```
adf.test(traindata)

##
##  Augmented Dickey-Fuller Test
##
## data:  traindata
## Dickey-Fuller = -2.1996, Lag order = 4, p-value = 0.4941
## alternative hypothesis: stationary
```

In the ADF test,
H_0 : the series is not stationary
H_a : the series is stationary
Hence, a small p-value (i.e., less than Alpha=0.05) suggests that the series is stationary. Here, the p-value(0.5043) is greater than Alpha(0.05). So, we can not reject H0. That means the ADF test shows that the UKgas dataset is not stationary.

Kwiatkowski-Phillips-Schmidt-Shin Test:

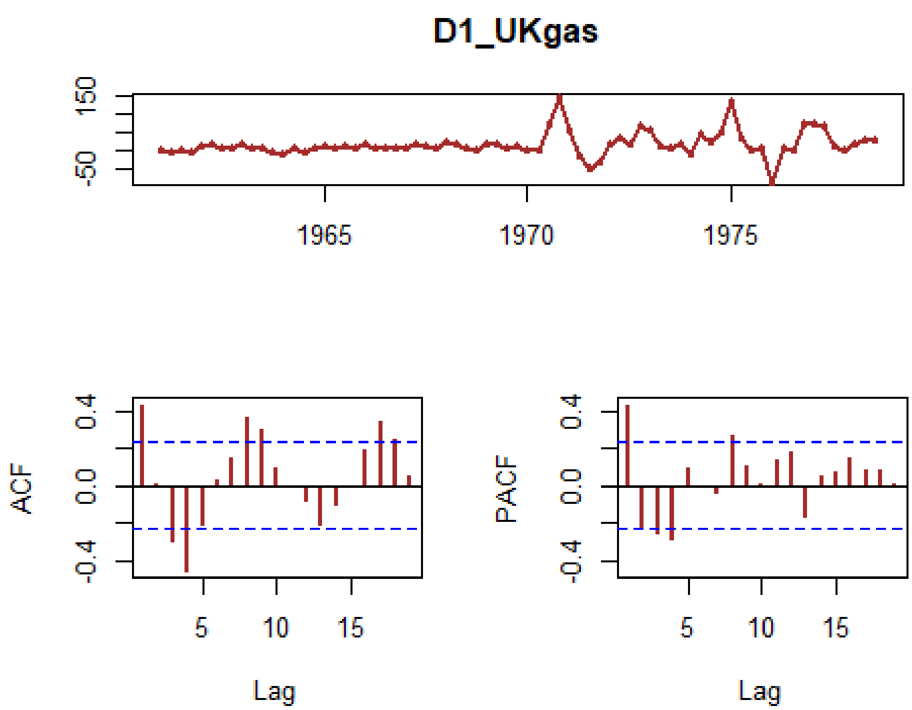
```
kpss.test(traindata)
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  traindata
## KPSS Level = 1.7857, Truncation lag parameter = 3, p-value = 0.01
```

In the KPSS test,
H_0 : the series is stationary
H_a : the series is not stationary
Hence, a small p-value suggests that the series is not stationary and a differencing is required. Here, the p-value(0.01) is less than Alpha(0.05). So, we can reject H0. That means the KPSS test shows that the UKgas dataset is not stationary.

So, here the seasonal differences are required. Now, we can check stationary after taking 1st differences.

```
D1_UKgas<-diff(traindata, lag=4)
tsdisplay(D1_UKgas, col='brown', lwd=2)
```




```
adf.test(D1_UKgas)

##
##  Augmented Dickey-Fuller Test
##
## data:  D1_UKgas
## Dickey-Fuller = -5.7039, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

```
kpss.test(D1_UKgas)

##
##  KPSS Test for Level Stationarity
##
## data:  D1_UKgas
## KPSS Level = 0.36104, Truncation lag parameter = 3, p-value = 0.09395
```

After taking 1st difference the ADF test shows that the p-value is less than Alpha(0.05) means the data is stationary and the kpss test shows that the p-value is greater than Alpha(0.05) means the data is stationary. So, we can say that one seasonal and 1 nonseasonal difference is enough for the M1 model UKgas dataset to get stationary series.

Here, we can see ACF and PACF plot our seasonal spike is AR=1, MA=1. And non seasonal spike is AR=3, MA=1. So the nonseasonal order is (p=3,d=1,q=1) and seasonal order is (P=1,D=1,Q=1).

Fit the M1 model:

```
M1<-Arima(traindata, order = c(3, 1, 1), seasonal = c(1, 1, 1), method="ML")
summary(M1)

## Series: traindata
## ARIMA(3,1,1)(1,1,1)[4]
##
## Coefficients:
##           ar1      ar2      ar3     ma1     sar1     sma1
##       -0.8951  -0.8121  -0.8816  0.3557  -0.5754  -0.6343
## s.e.   0.0726   0.0920   0.0698  0.1419   0.1029   0.1185
##
## sigma^2 = 694.8:  log likelihood = -327.29
## AIC=668.58   AICc=670.39   BIC=684.32
##
## Training set error measures:
##                    ME     RMSE      MAE       MPE    MAPE     MASE
ACF1
## Training set 3.581664 24.34885 14.11413 0.7893229 5.71129 0.631229 -0.0428
8031
```

Mathematical Expression of M1 model:

ARIMA (3,1,1)(1,1,1)[4]

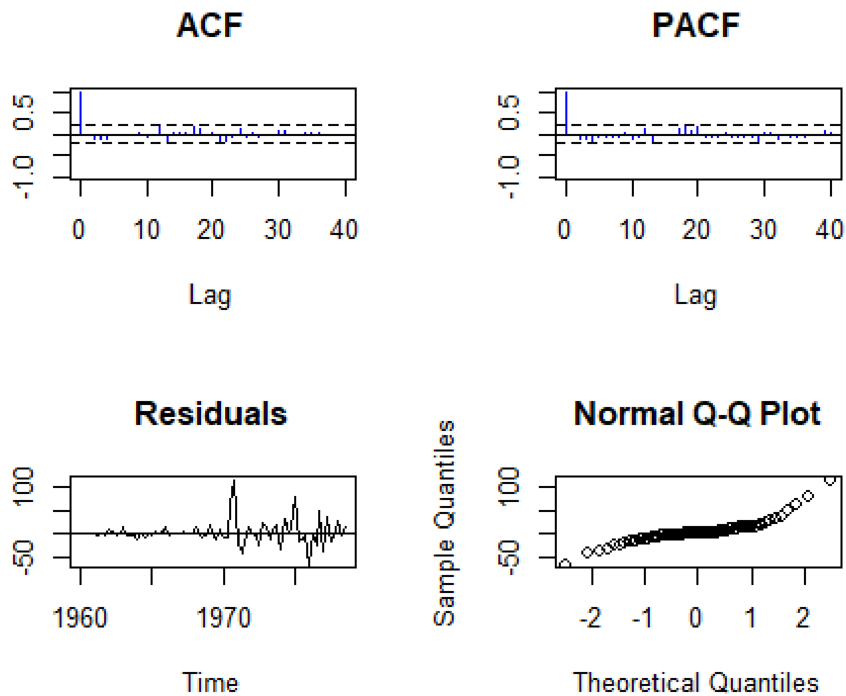
Estimated Model:

$$(1 + 0.8951B + 0.8121B^2 + 0.8816B^3)(1 + 0.5754B^4)(1 - B)(1 - B^4)y_t = (1 + 0.3557B)(1 - 0.6343B^4)Z_t, \qquad where\ \{Z_t\} \sim WN(0, \hat{\sigma}^2 = 694.8)$$

Residuals Diagnostic Checking:

```
test(M1$residuals)
```

```
## Null hypothesis: Residuals are iid noise.
## Test                        Distribution Statistic   p-value
## Ljung-Box Q                Q ~ chisq(20)     13.19    0.8689
## McLeod-Li Q                Q ~ chisq(20)      13.4    0.8598
## Turning points T    (T-48.7)/3.6 ~ N(0,1)        49    0.9264
## Diff signs S          (S-37)/2.5 ~ N(0,1)        37         1
## Rank P          (P-1387.5)/109.3 ~ N(0,1)      1442    0.6181
```



Assumptions:

i. {et} uncorrelated. Here, we can see the ACF and PACF plot of residuals diagnostics, there is no significant spike. So, we can say that the {et} is uncorrelated.

ii. {et} have mean zero. From our Residuals plot, we can say that mean zero but variance in not much constant.

iii. {et} is normally distributed. Here, the Normal Q-Q plot shows that the residuals are approximately normally distributed because the data is near 45 degrees.

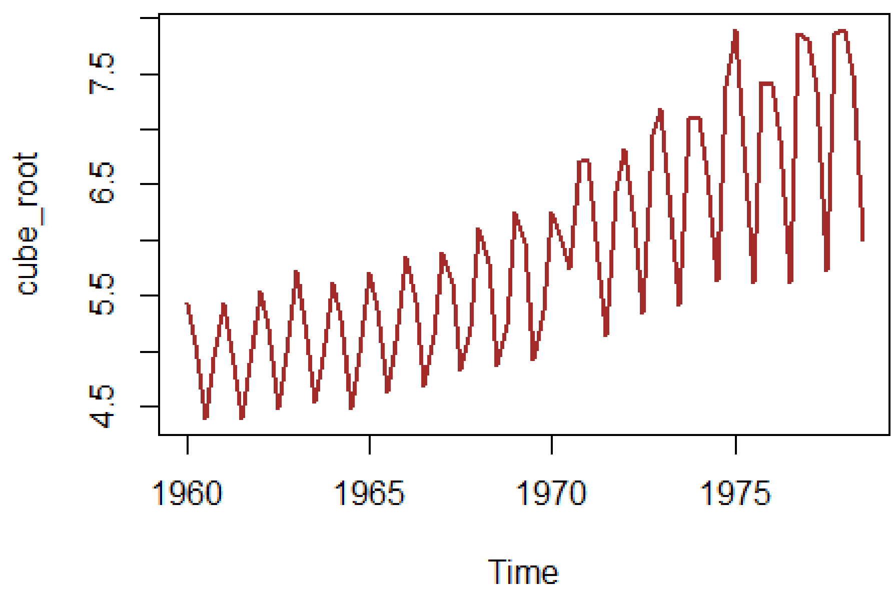
Here, H_0: Residuals are iid noise.
We can see the summary of the residuals test here all p-value is greater than Alpha (0.05).
So, here we can not reject H_0. So, the M1 model satisfied all assumptions of residuals.

3. Transformation

As it has increasing upward trend and multiplicative seasonality, cube root transformation is perfect for this data set.

Cube Root Transformation:

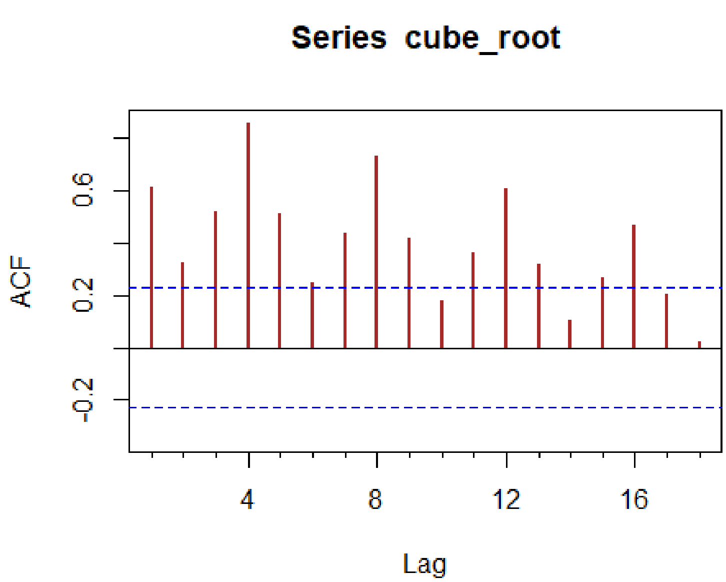
```
cube_root<-traindata**(1/3)
plot(cube_root, col='brown', lwd=2)
```



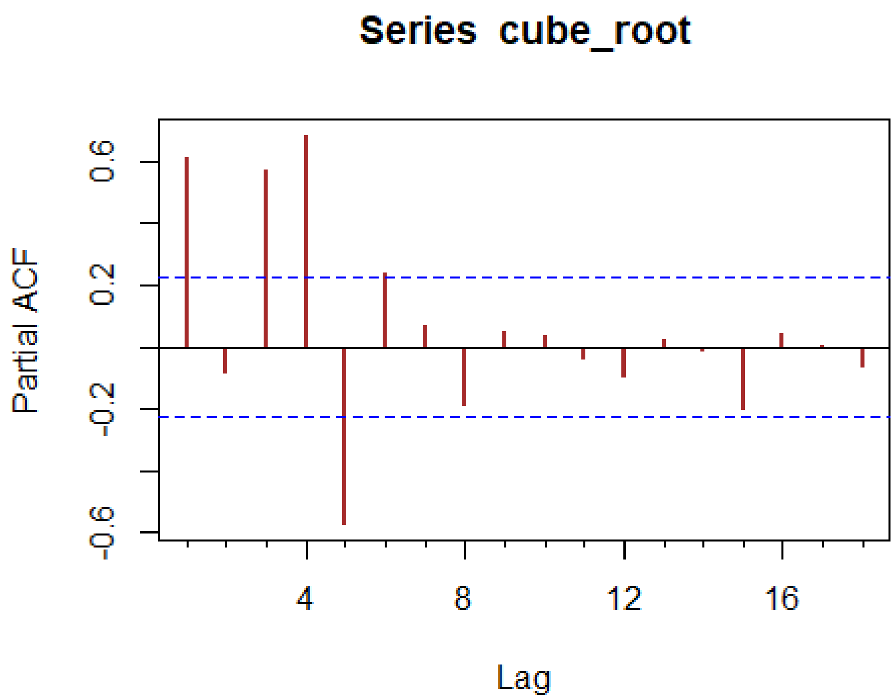
After cube root transformation, the time plot shows that a upward dumped trend and additive seasonality.

ACF and PACF plot:

```
Acf(cube_root, col='brown', lwd=2)
```



```
Pacf(cube_root, col='brown', lwd=2)
```

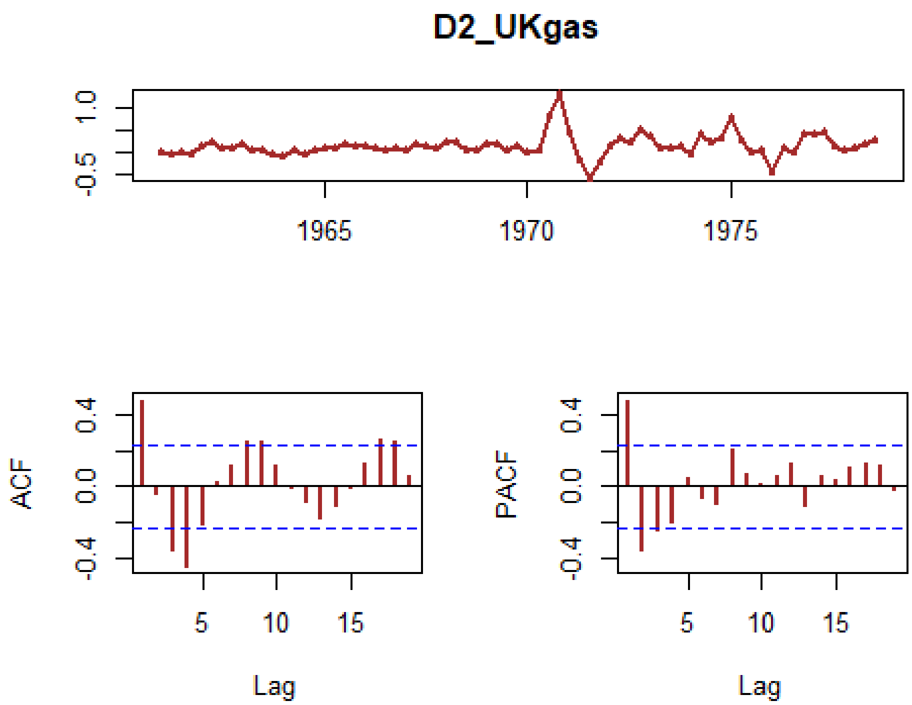


ADF and KPSS Test:

```
adf.test(cube_root)

##
##  Augmented Dickey-Fuller Test
##
## data:  cube_root
## Dickey-Fuller = -3.044, Lag order = 4, p-value = 0.1496
## alternative hypothesis: stationary

kpss.test(cube_root)

##
##  KPSS Test for Level Stationarity
##
## data:  cube_root
## KPSS Level = 1.8497, Truncation lag parameter = 3, p-value = 0.01
```

Here, the ADF test shows that the p-value is greater than Alpha(0.05) means the data is not stationary and the kpss test shows that the p-value is less than Alpha(0.05) means the data is not stationary. So, here 1 seasonal differences is required.

```
D2_UKgas<-diff(cube_root, lag=4)
tsdisplay(D2_UKgas, col='brown', lwd=2)
```




```
adf.test(D2_UKgas)

##
##  Augmented Dickey-Fuller Test
##
## data:  D2_UKgas
## Dickey-Fuller = -5.0527, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary

kpss.test(D2_UKgas)

##
##  KPSS Test for Level Stationarity
##
## data:  D2_UKgas
## KPSS Level = 0.22712, Truncation lag parameter = 3, p-value = 0.1
```

After taking 1st difference the ADF test shows that the p-value is less than Alpha(0.05) means the data is stationary and the kpss test shows that the p-value is greater than Alpha(0.05) means the data is stationary. So, we can say that one seasonal and 0 nonseasonal difference is enough for the M2 model UKgas dataset to get stationary series.

Here, we can see ACF and PACF plot our seasonal spike is AR=1, MA=1. And non seasonal spike is AR=3, MA=1. So the nonseasonal order is (p=3,d=0,q=1) and seasonal order is (P=1,D=1,Q=0).

Fit the M2 model:

```
M2<-Arima(cube_root, order = c(3, 0, 1), seasonal = c(1, 1, 0))
summary(M2)

## Series: cube_root
## ARIMA(3,0,1)(1,1,0)[4]
##
## Coefficients:
##           ar1     ar2      ar3     ma1      sar1
##       -0.1912  0.5485  -0.2200  0.9363  -0.2429
## s.e.   0.1263  0.1093   0.1209  0.0627   0.1272
##
## sigma^2 = 0.05167:  log likelihood = 6.47
## AIC=-0.93   AICc=0.38   BIC=12.64
##
## Training set error measures:
##                      ME     RMSE      MAE      MPE     MAPE      MASE
## Training set 0.07121495 0.213234 0.139853 1.078315 2.232152 0.7428787
##                     ACF1
## Training set -0.07558433
```

Mathematical Expression of M2 model:

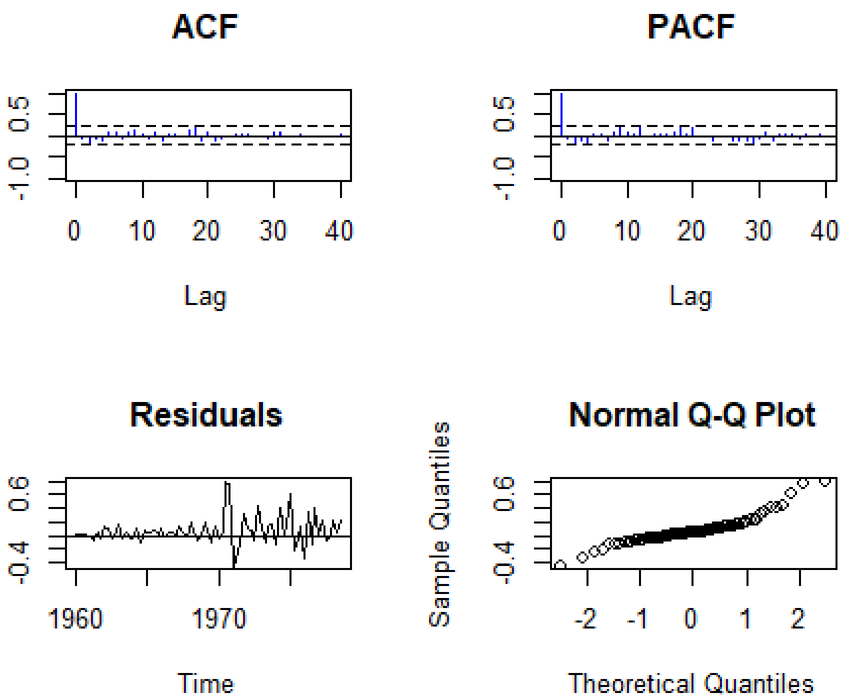
ARIMA (3,0,1)(1,1,0)[4]

Estimated Model:

$$(1 + 0.1912B - 0.5485B^2 + 0.2200B^3)(1 + 0.2429B^4)(1 - B^4)y_t = (1 + 0.9363B)Z_t,$$
$$where\ \{Z_t\} \sim WN(0, \hat{\sigma}^2 = 0.05167)$$

Residuals Diagnostic Checking:

```
test(M2$residuals)

## Null hypothesis: Residuals are iid noise.
## Test                         Distribution Statistic   p-value
## Ljung-Box Q                 Q ~ chisq(20)     16.82    0.6647
## McLeod-Li Q                 Q ~ chisq(20)     25.92    0.1686
## Turning points T   (T-48.7)/3.6 ~ N(0,1)         51    0.5177
## Diff signs S         (S-37)/2.5 ~ N(0,1)         37         1
## Rank P          (P-1387.5)/109.3 ~ N(0,1)      1563    0.1084
```



Assumptions:

i. {et} uncorrelated. Here, we can see the ACF and PACF plot of residuals diagnostics, there is no significant spike. So, we can say that the {et} is uncorrelated.

ii. {et} have mean zero. From our Residuals plot, we can say that mean zero but variance in not much constant.

iii. {et} is normally distributed. Here, the Normal Q-Q plot shows that the residuals are approximately normally distributed because the data is near 45 degrees.

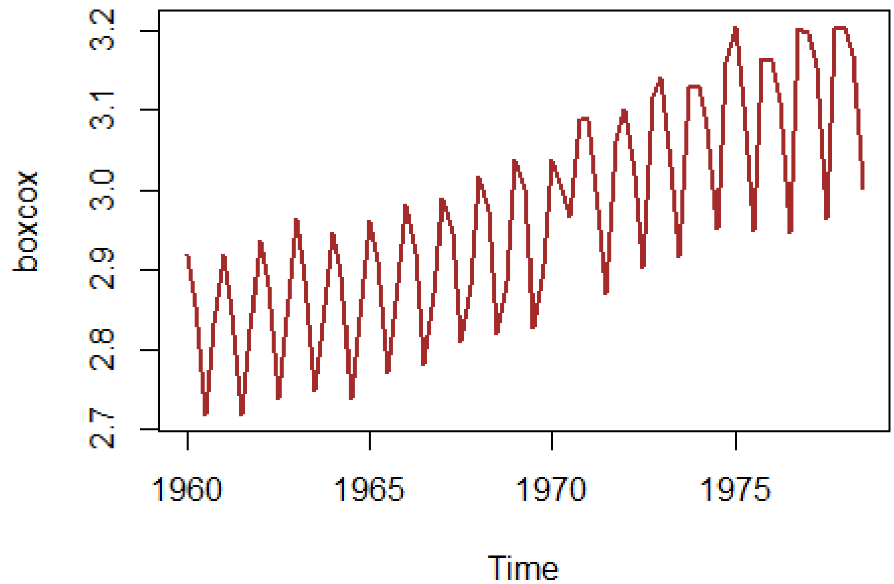
Here, H_0: Residuals are iid noise.

We can see the summary of the residuals test here without one other p-value is greater than Alpha (0.05). So, here we can not reject H_0. So, the M2 model satisfied all assumptions of residuals.

4. Box-cox Transformation:

Optimum lambda value:

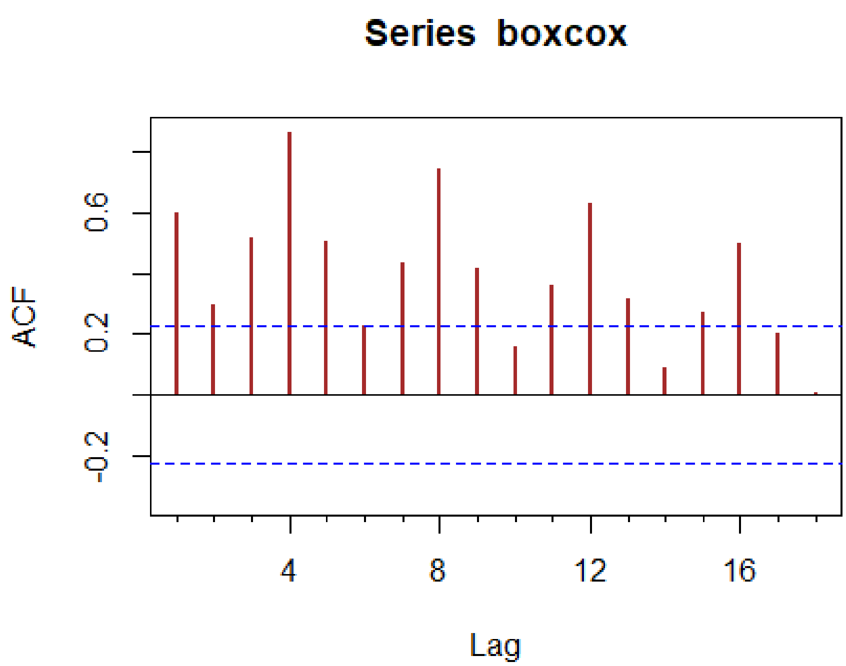
```
lambda<-BoxCox.lambda(traindata)
lambda

## [1] -0.2426249

boxcox<-BoxCox(traindata, lambda = lambda)
plot(boxcox, col='brown', lwd=2)
```



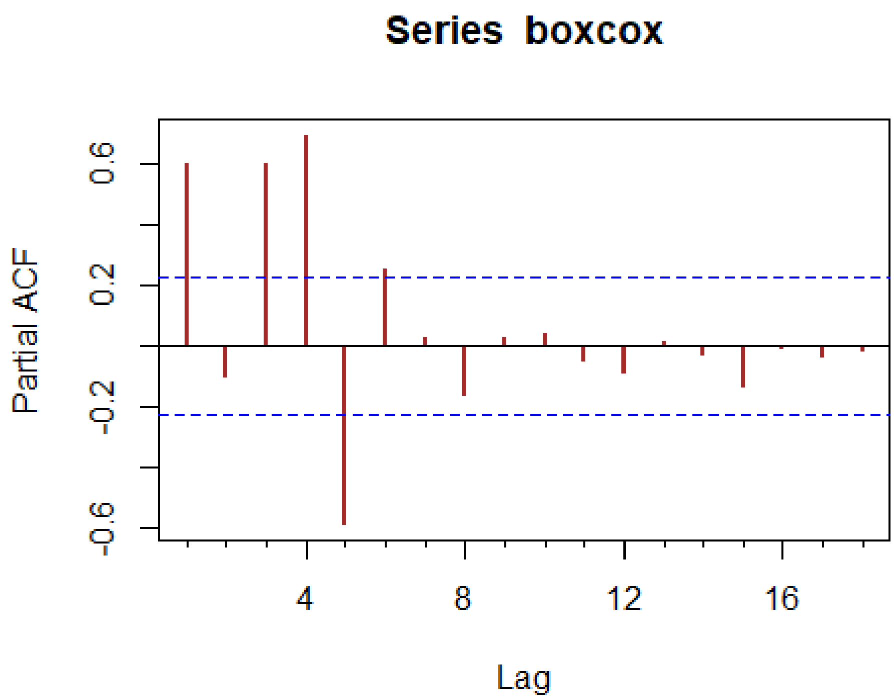
After box-cox transformation the time plot shows that upward dumped trend.

ACF and PACF plot:

```
Acf(boxcox, col='brown', lwd=2)
```



```
Pacf(boxcox, col='brown', lwd=2)
```

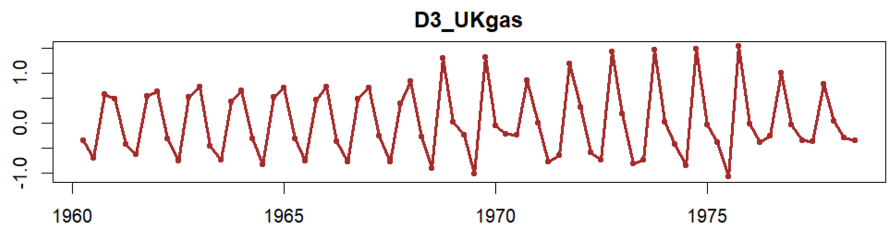


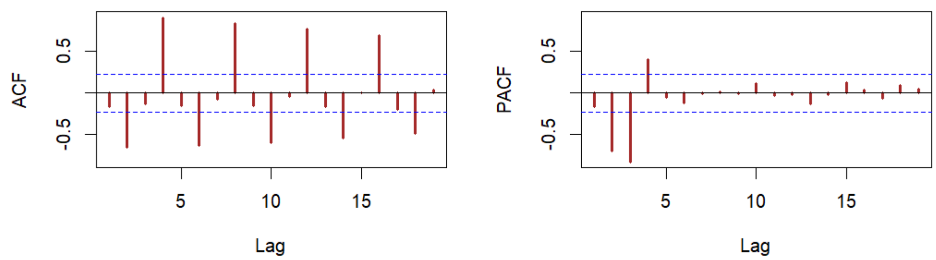
ADF and KPSS Test:

```
adf.test(boxcox)

##
##  Augmented Dickey-Fuller Test
##
## data:  boxcox
## Dickey-Fuller = -3.8995, Lag order = 4, p-value = 0.01897
## alternative hypothesis: stationary

kpss.test(boxcox)

##
##  KPSS Test for Level Stationarity
##
## data:  boxcox
## KPSS Level = 1.8815, Truncation lag parameter = 3, p-value = 0.01
```

Here, the ADF test shows that the p-value is less than Alpha(0.05) means the data is stationary and the kpss test shows that the p-value is less than Alpha(0.05) means the data is not stationary. So, here nonseasonal differences is required.

```
D3_UKgas<-diff(boxcox, differences=1)
tsdisplay(D3_UKgas, col='brown', lwd=2)
```






```
adf.test(D3_UKgas)

##
##  Augmented Dickey-Fuller Test
##
## data:  D3_UKgas
## Dickey-Fuller = -4.9066, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary

kpss.test(D3_UKgas)

##
##  KPSS Test for Level Stationarity
##
## data:  D3_UKgas
## KPSS Level = 0.27197, Truncation lag parameter = 3, p-value = 0.1
```

After taking 1st difference the ADF test shows that the p-value is less than Alpha(0.05) means the data is stationary and the kpss test shows that the p-value is greater than Alpha(0.05) means the data is stationary. So, we can say that one nonseasonal difference is enough for the M3 model UKgas dataset to get stationary series.

Here, we can see ACF and PACF plot our seasonal spike is AR=1, MA=0. And seasonal order is (P=1,D=1,Q=0). And nonseasonal order is (p=1,d=0,q=0)

Fit the M3 model:

```
M3<-Arima(boxcox, order = c(1, 0, 0), seasonal = c(1, 1, 0), include.drift =
TRUE)
summary(M3)

## Series: boxcox
## ARIMA(1,0,0)(1,1,0)[4] with drift
##
```

```
## Coefficients:
##          ar1      sar1   drift
##       0.4441   -0.3453  0.0045
## s.e.  0.1083    0.1111  0.0011
##
## sigma^2 = 0.0007967:  log likelihood = 153.75
## AIC=-299.51   AICc=-298.9   BIC=-290.46
##
## Training set error measures:
##                         ME        RMSE         MAE          MPE      MAPE
## Training set 4.318478e-05 0.02687573 0.01663033 -0.008349374 0.5558844
##                   MASE      ACF1
## Training set 0.6596294 0.1756305
```

Mathematical Expression of M3 model:

ARIMA (1,0,0)(1,1,0)[4]

Estimated Model:

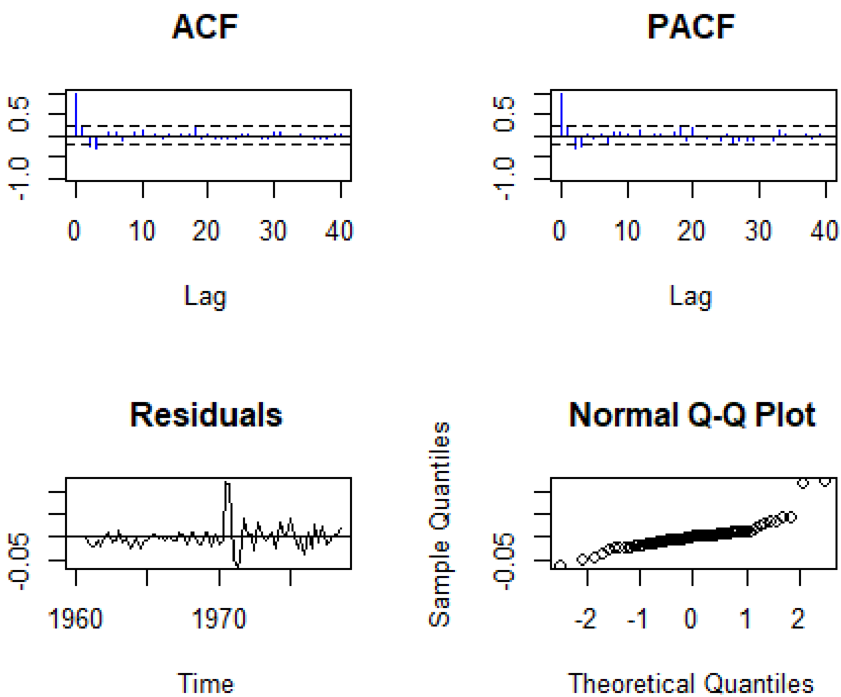
$$(1-0.4441B)(1+0.3453B^4)(1-B^4)y_t = Z_t, \qquad where\ \{Z_t\} \sim WN(0, \hat{\sigma}^2 = 0.0007967)$$

Mean Form:

$$(1-0.4441B)(1+0.3453B^4)(y_t' - 0.0045) = Z_t, where\ \{Z_t\} \sim WN(0, \hat{\sigma}^2 = 0.0007967)$$

Residuals Diagnostic Checking:

```
test(M3$residuals)

## Null hypothesis: Residuals are iid noise.
## Test                        Distribution Statistic   p-value
## Ljung-Box Q                Q ~ chisq(20)     23.96    0.2443
## McLeod-Li Q                Q ~ chisq(20)     29.78    0.0736
## Turning points T   (T-48.7)/3.6 ~ N(0,1)        51    0.5177
## Diff signs S         (S-37)/2.5 ~ N(0,1)        37         1
## Rank P          (P-1387.5)/109.3 ~ N(0,1)      1508    0.2703
```



Assumptions:

i. {et} uncorrelated. Here, we can see the ACF and PACF plot of residuals diagnostics, there is no significant spike. So, we can say that the {et} is uncorrelated.

ii. {et} have mean zero. From our Residuals plot, we can say that mean zero but variance in constant.

iii. {et} is normally distributed. Here, the Normal Q-Q plot shows that the residuals are approximately normally distributed because the data is near 45 degrees.

Here, H_0: Residuals are iid noise.
We can see the summary of the residuals test here all p-value is greater than Alpha (0.05). So, here we can not reject H_0. So, the M3 model satisfied all assumptions of residuals.

5. Model Selection Criteria:

Akaike Information Criterion:

$$AIC(\beta) = -2 \ln L_X(\beta, S_X(\beta)/n) + 2(p + q + 1)$$

Akaike Information Criterion corrected:

$$AICc(\beta) = -2 \ln L_X(\beta, S_X(\beta)/n) + \frac{2(p + q + 1)n}{(n - p - q - 2)}$$

Bayesian Information Criterion:

$$BIC = (n-p-q)\ln\left[\frac{n\hat{\sigma}^2}{(n-p-q)}\right] + n\left(1 + ln\sqrt{2\pi}\right) + (p+q)ln\left[\frac{\sum_{t=1}^{n} X_t^2 - n\hat{\sigma}^2}{p+q}\right]$$

Model	AIC	AICc	BIC
M1	668.58	670.39	684.32
M2	-0.93	0.38	12.64
M3	-299.51	-298.9	-290.46

For the same family as ARIMA whose AIC and AICc value among other models is the lowest is the best model. Here, M3 is the best fitted model because its AIC and AICc value is lowest. And, M3 model fulfilled more underlying assumptions of residuals than others.

6.Forecasting accuracy measures based on the test data set:

$$ME = \frac{1}{n}\sum_{t=1}^{n} e_t$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t|$$

$$MSE = \frac{1}{n}\sum_{t=1}^{n} e_t^2$$

$$MPE = \frac{1}{n}\sum_{t=1}^{n} PE_t$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} |PE_t|$$

$$PE_t = \left(\frac{Y_t - F_t}{Y_t}\right) * 100$$

For M1 model,

```
forecasting_mean1<-forecast::forecast(M1,h=length(testdata))$mean
error1<-testdata-forecasting_mean1
MSE1<-sum(error1**2)/length(testdata)
MSE1
```

```
## [1] 9871.195

RMSE1<-sqrt(MSE1)
RMSE1

## [1] 99.35389

MAE1<-sum(abs(error1))/length(testdata)
MAE1

## [1] 76.52547

PE1<-(100*error1)/testdata
MPE1<-sum(PE1)/length(testdata)
MPE1

## [1] 5.465914

MAPE1<-sum(abs(PE1))/length(testdata)
MAPE1

## [1] 15.15148
```

For M2 model,

```
forecasting_mean2<-forecast::forecast(M2,h=length(testdata))$mean
error2<-testdata-forecasting_mean2
MSE2<-sum(error2**2)/length(testdata)
MSE2

## [1] 320412.3

RMSE2<-sqrt(MSE2)
RMSE2

## [1] 566.0498

MAE2<-sum(abs(error2))/length(testdata)
MAE2

## [1] 555.3293

PE2<-(100*error2)/testdata
MPE2<-sum(PE2)/length(testdata)
MPE2

## [1] 98.65587

MAPE2<-sum(abs(PE2))/length(testdata)
MAPE2

## [1] 98.65587
```

For M3 model,

```
forecasting_mean<-forecast::forecast(M3,h=length(testdata))$mean
newf<-(lambda*forecasting_mean+1)**(1/lambda)
error<-testdata-forecasting_mean

MSE<-sum(error**2)/length(testdata)
MSE

## [1] 325123.1

RMSE<-sqrt(MSE)
RMSE

## [1] 570.1957

MAE<-sum(abs(error))/length(testdata)
MAE

## [1] 559.4744

PE<-(100*error)/testdata
MPE<-sum(PE)/length(testdata)
MPE

## [1] 99.40564

MAPE<-sum(abs(PE))/length(testdata)
MAPE

## [1] 99.40564
```

Back Transformation of predicted value of M2 model:

```
M2_back<-(M2$fitted)**3
```

Back Transformation of predicted value of M3 model:

```
M3_back<-InvBoxCox(M3$fitted, lambda = lambda)
```

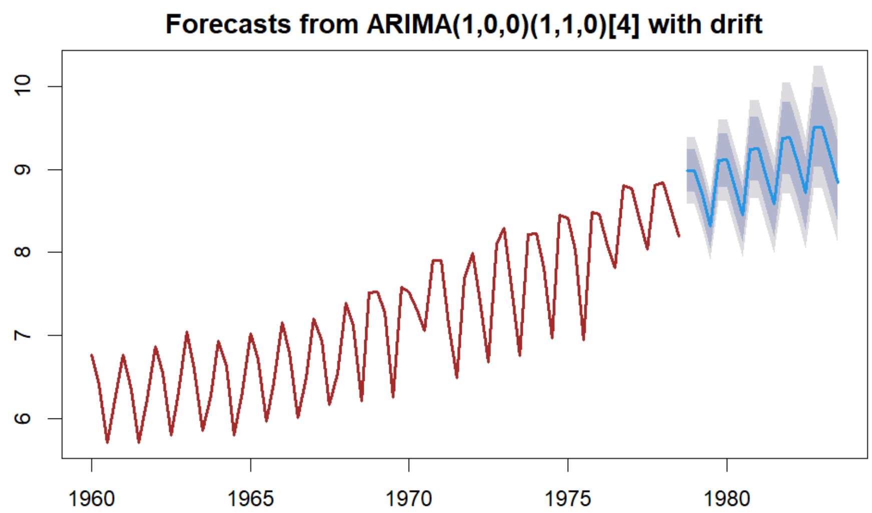
Model	**MSE**	**RMSE**	**MAE**	**MPE**	**MAPE**
M1	9871.195	99.35389	76.52547	5.465914	15.15148
M2	320412.3	566.0498	555.3293	98.65587	98.65587
M3	325123.1	570.1957	559.4744	99.40564	99.40564

After checking forecasting accuracy measures for the test data set, here the M1 model forecasting accuracy measures values are lower than other models. So, based on that M1 model is best fitted model for forecasting.

7. Forecasting

Using the best model M3, forecasting 10 points ahead:

```
forecasting<-forecast::forecast(M3,h=20)
plot(forecasting)
```



THE END