Time Series Modeling in R

Introduction

Dataset

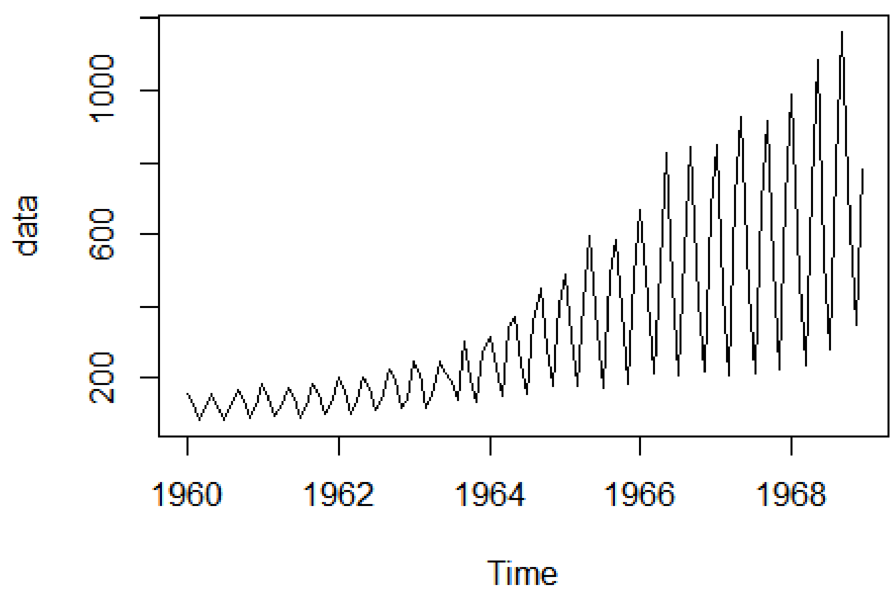
The dataset chosen for this study is the UK gas dataset from the R "datasets" package. The data discusses the quarterly UK gas consumption from 1960 Quarter 1 to 1986 Quarter 4, in millions of terms. The data is a quarterly time series data of length 108

Procedure

Import the required dataset

```
View(uspop)
UKgas

##         Qtr1   Qtr2   Qtr3   Qtr4
## 1960  160.1  129.7   84.8  120.1
## 1961  160.1  124.9   84.8  116.9
## 1962  169.7  140.9   89.7  123.3
## 1963  187.3  144.1   92.9  120.1
## 1964  176.1  147.3   89.7  123.3
## 1965  185.7  155.3   99.3  131.3
## 1966  200.1  161.7  102.5  136.1
## 1967  204.9  176.1  112.1  140.9
## 1968  227.3  195.3  115.3  142.5
## 1969  244.9  214.5  118.5  153.7
## 1970  244.9  216.1  188.9  142.5
## 1971  301.0  196.9  136.1  267.3
## 1972  317.0  230.5  152.1  336.2
## 1973  371.4  240.1  158.5  355.4
## 1974  449.9  286.6  179.3  403.4
## 1975  491.5  321.8  177.7  409.8
## 1976  593.9  329.8  176.1  483.5
## 1977  584.3  395.4  187.3  485.1
## 1978  669.2  421.0  216.1  509.1
## 1979  827.7  467.5  209.7  542.7
## 1980  840.5  414.6  217.7  670.8
## 1981  848.5  437.0  209.7  701.2
## 1982  925.3  443.4  214.5  683.6
## 1983  917.3  515.5  224.1  694.8
## 1984  989.4  477.1  233.7  730.0
## 1985 1087.0  534.7  281.8  787.6
## 1986 1163.9  613.1  347.4  782.8
```

```
data=UKgas
data=ts(data,start=1960,frequency = 12)# Run the time series data
ts.plot(data)# Plot the data
```



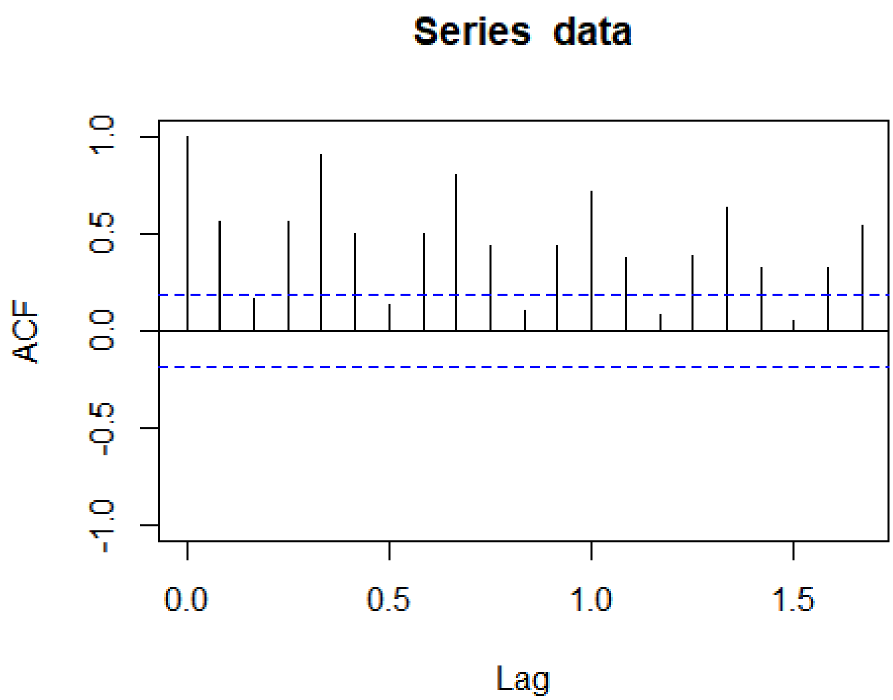
It is observed that the data is a non-stationary data with longterm increasing trend and seasonality component.

Test for stationarity

Here we use ACF Plot and ADF test to check for stationarity

ACF Plot(Graphical Test)

```
acf(data,ylim=c(-1,1))
```



Interpretation

Here, we observe that most of the lag values are above the threshold line. Therefore, the data is non-stationary

ADF Test (Statistical test)

```
library(tseries)

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

adf.test(data)

##
##  Augmented Dickey-Fuller Test
##
## data:  data
## Dickey-Fuller = -1.6079, Lag order = 4, p-value = 0.7393
## alternative hypothesis: stationary
```

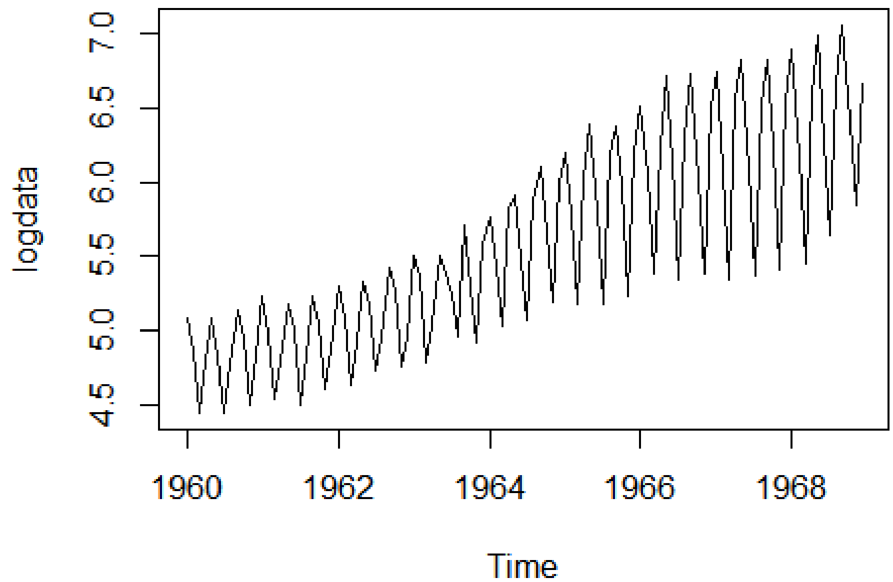
From the results, we observe that the p-value is greater than 0.7393.Therefore, the data is non-stationary

Eliminating trend and seasonal component

Method of Seasonal Differencing

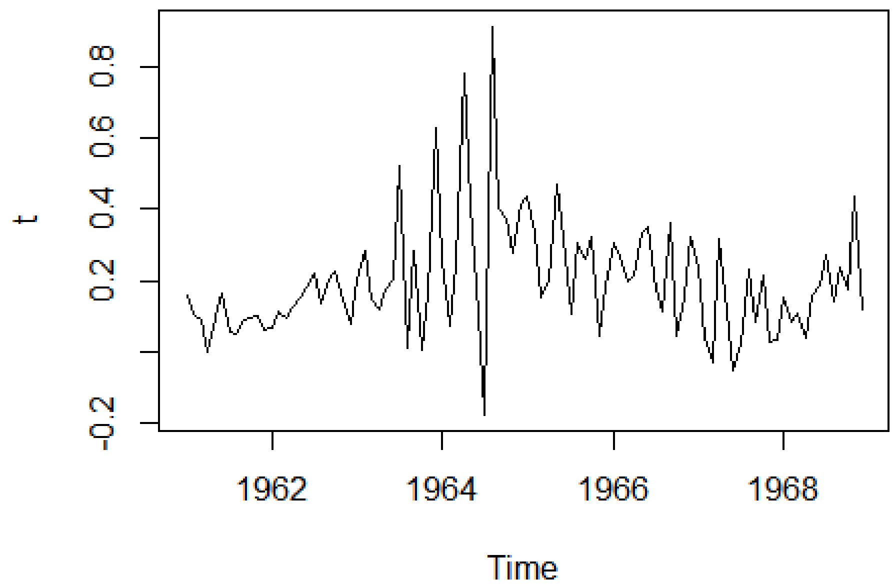
Since the data has both seasonal and trend component, we need to convert it to additive structure before seasonal differencing. The log of the data is created using the following function.

```
logdata=log(data)
ts.plot(logdata)
```

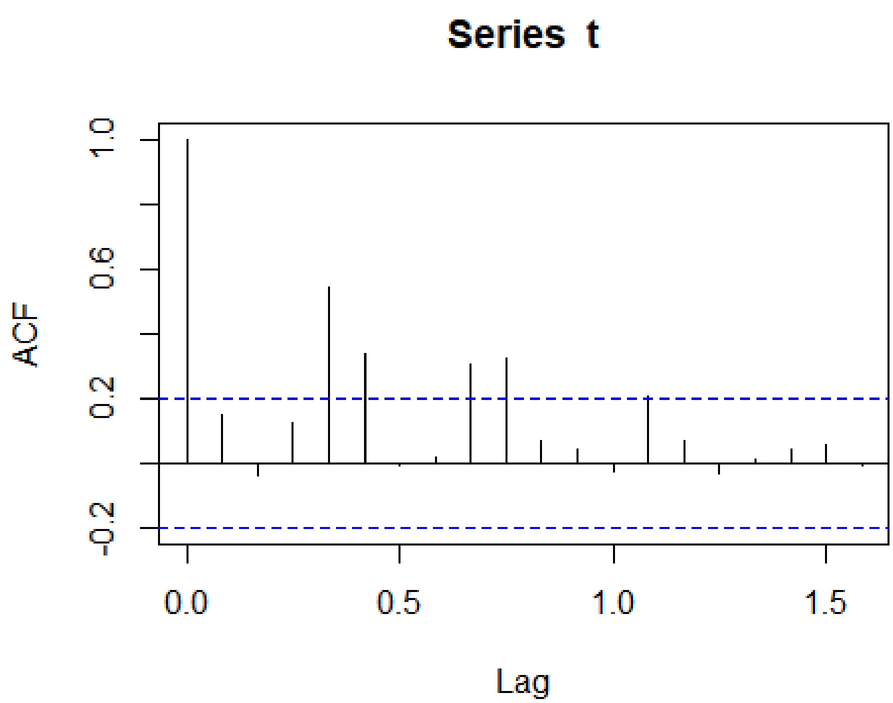


Now, we perform the method of seasonal differencing to eliminate the seasonality component.

```
t=diff(logdata,lag=12)
ts.plot(t)
```

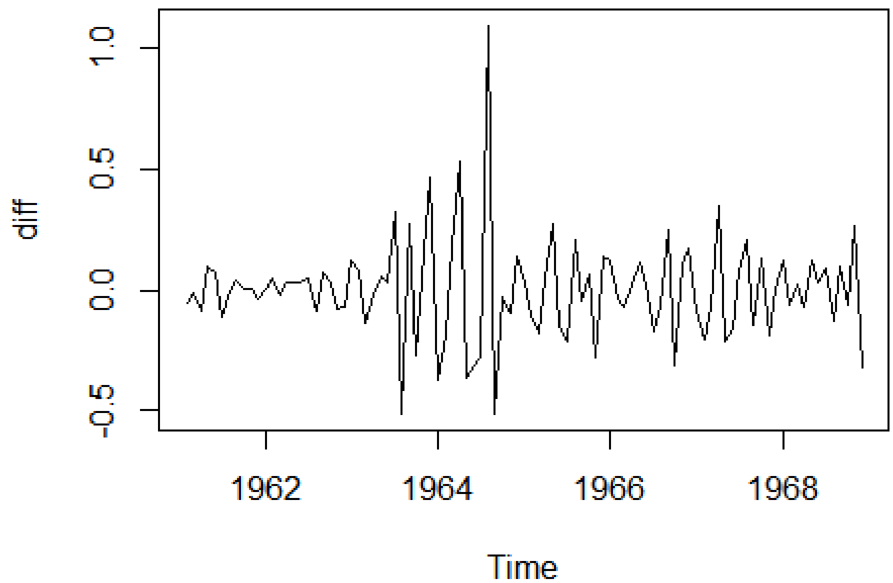



```
acf(t)
```



Here, after Eliminating the seasonal component, the model is reduced into zt = mt + et. No, we perform the method of differencing to eliminate trend component.

```
diff=diff(t)
ts.plot(diff)
```



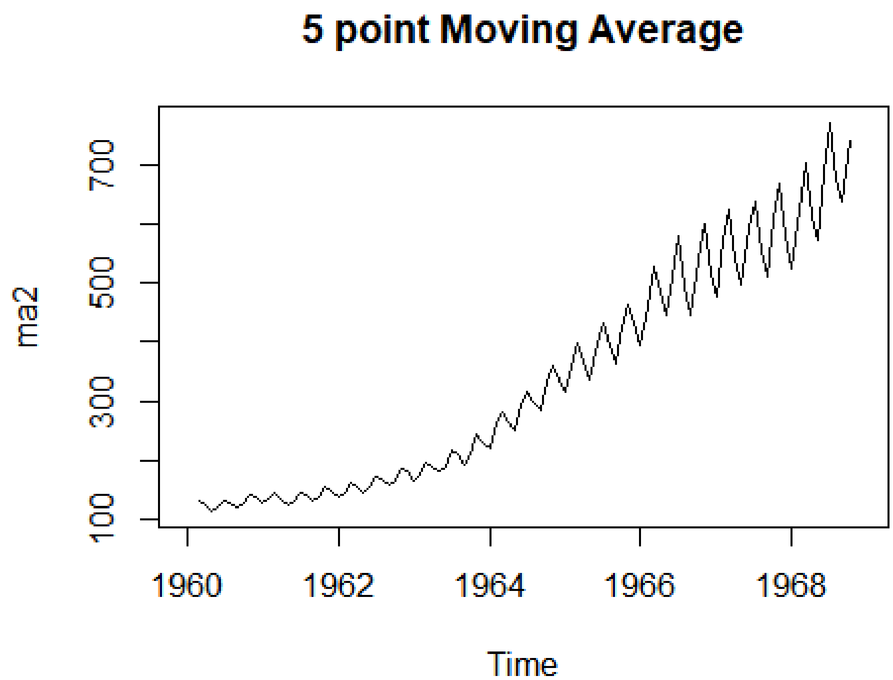

```
library(tseries)
adf.test(diff)

## Warning in adf.test(diff): p-value smaller than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  diff
## Dickey-Fuller = -7.7602, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

From the plots and test, it can be seen that after performing differencing operation to the seasonal differenced series, the data set becomes a stationary process.

Moving Average Method

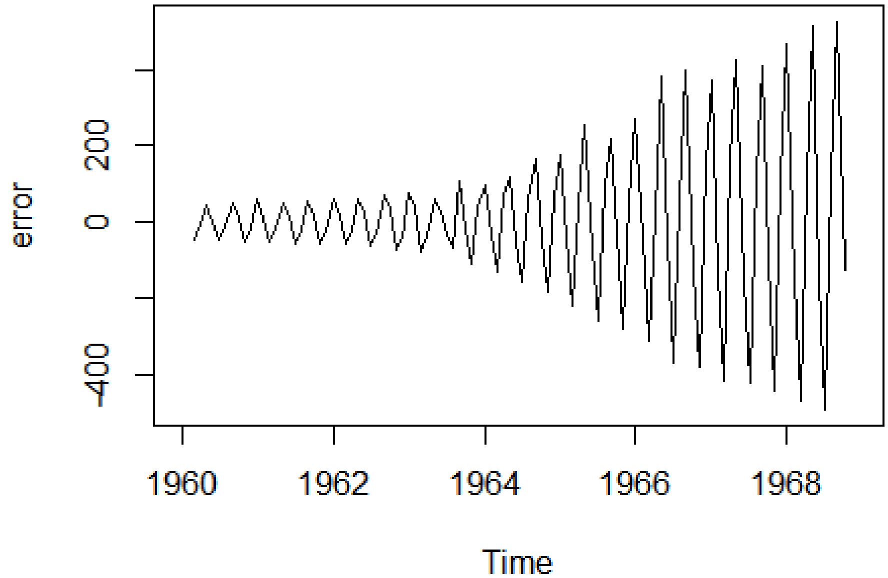
Here, we perform the moving average smoothing to estimate the trend component in the model zt = mt + et.

```
library(forecast)
ma2=ma(data, order = 5)
ts.plot(ma2,main="5 point Moving Average")
```



The extraction of stationary data from original data is done by subtracting the data from the estimated values.

```
error=data-ma2
ts.plot(error)
```

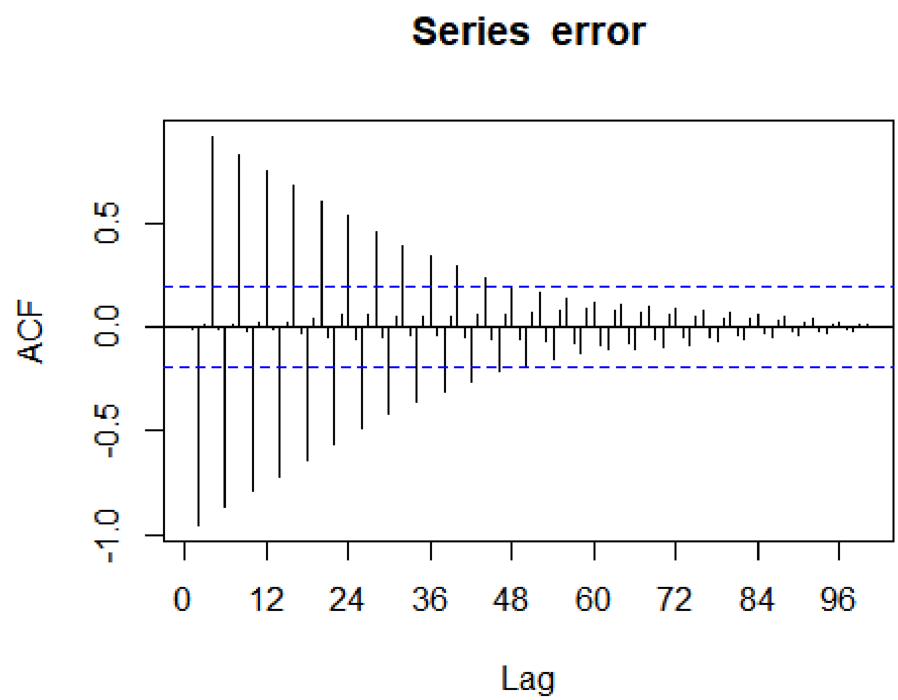


Here, we observe that the time series plot of the error component is stationary in nature.

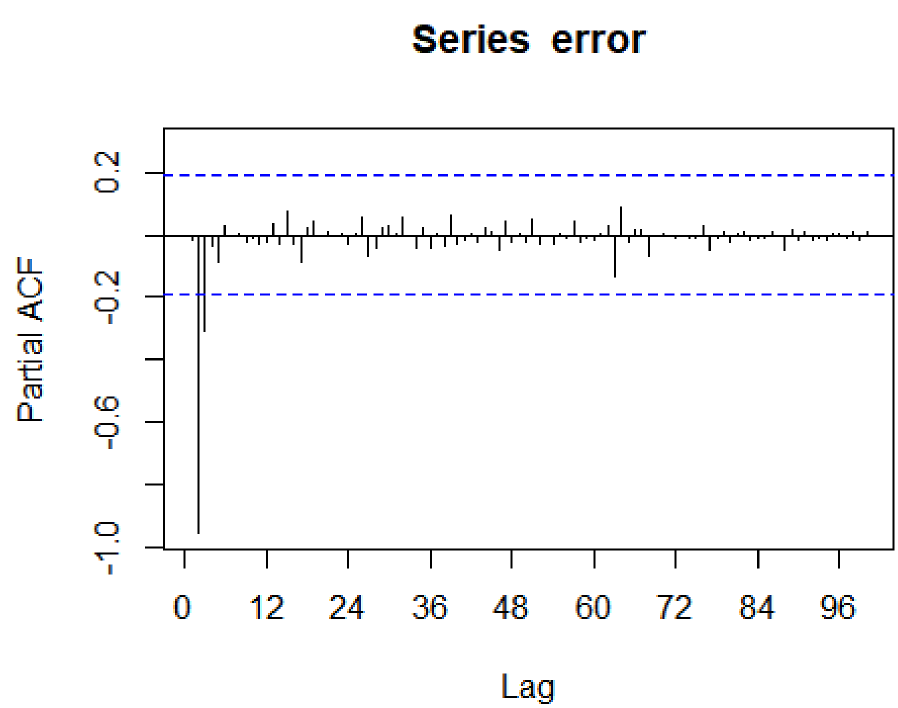
Thus, the method of differencing and method of Moving Average is used to convert the non-stationary data into stationary.

ACF and PACF of stationary data from Method of Moving Average

```
Acf(error,lag=100)
```

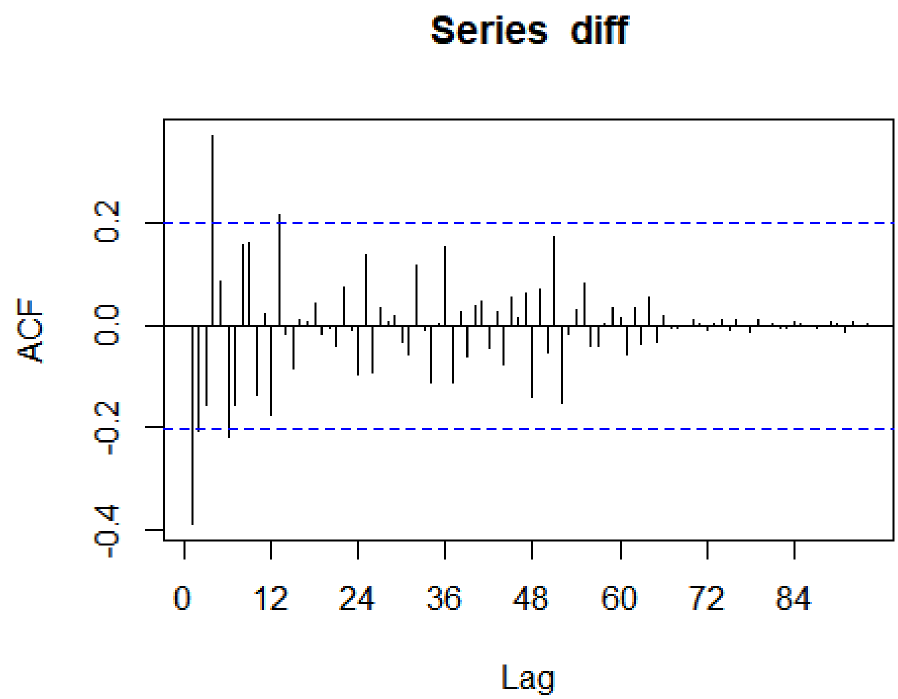



```
Pacf(error,lag=100)
```

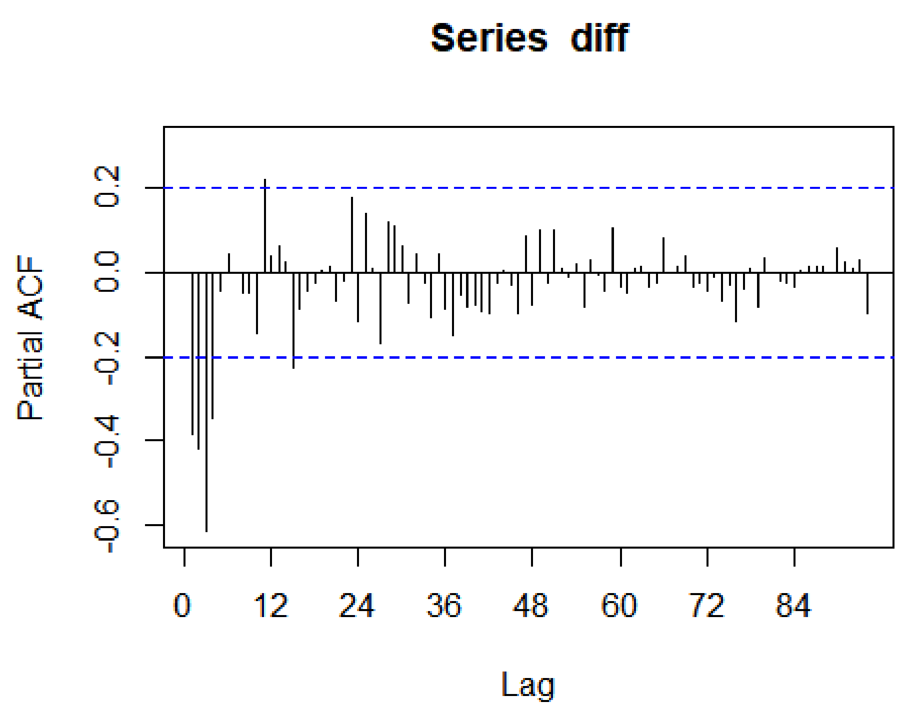


ACF and PACF of stationary data from Method of Differencing

```
Acf(diff,lag=100)
```




```
Pacf(diff,lag=100)
```



ARMA Model for data from Method of Moving Average

```
fit1 = auto.arima(error, seasonal="FALSE")
fit1

## Series: error
## ARIMA(0,0,0) with zero mean
##
## sigma^2 estimated as 37007:  log likelihood=-694.55
## AIC=1391.1   AICc=1391.14   BIC=1393.74
```

An ARIMA(0,0,0) model with zero mean is white noise, so it means that the errors are uncorrelated across time.

ARMA Model for data from Method of Differencing

```
fit2 = auto.arima(diff, seasonal="FALSE")
fit2

## Series: diff
## ARIMA(4,0,0) with zero mean
##
## Coefficients:
##           ar1      ar2      ar3      ar4
##       -1.0534  -1.0486  -0.9213  -0.3661
## s.e.   0.0953   0.1075   0.1054   0.0945
##
## sigma^2 estimated as 0.01726:  log likelihood=58.78
## AIC=-107.55   AICc=-106.88   BIC=-94.78
```

The best fitted ARMA model is ARIMA (4,0,0).Model can be written as -1.0534-1.04860.9213-0.3661

Comparing the AIC and BIC values for the ARMA models from Method of Differencing and Methof of Moving Average, we see that AIC and BIC for model from the Method of Differencing is less than the latter. Therefore, the fitted model from Method of Differencing would be preferred the most for the prediction purpose

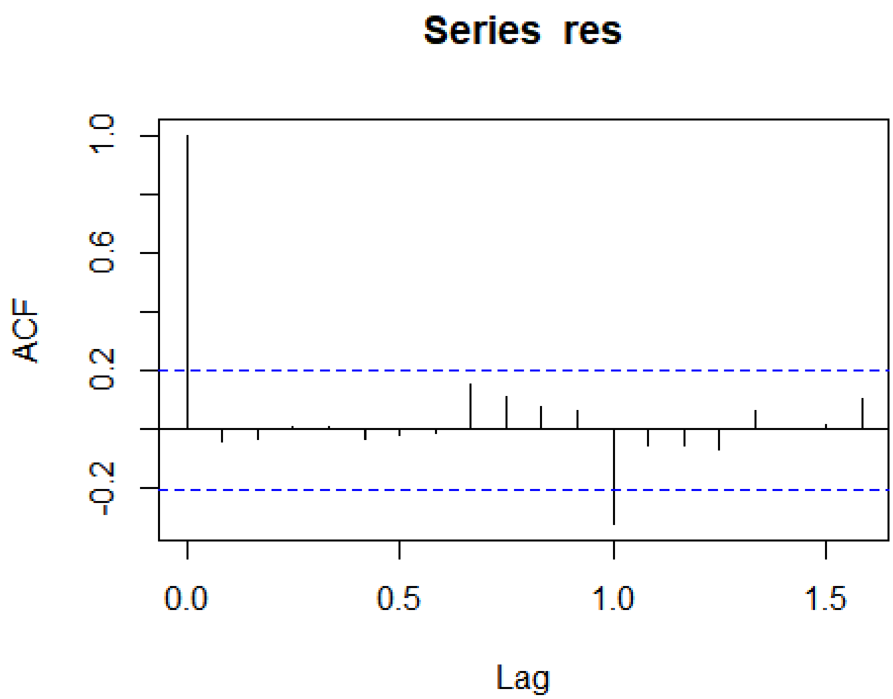
Testing the Assumptions

Residual Value

```
res = resid(fit2)
head(res)

##              Feb         Mar         Apr         May         Jun
Jul
## 1961 -0.03143515 -0.02320086 -0.08822238 -0.02131921  0.04253255 -
0.02467213
```

Assumption : Errors are Uncorrelated

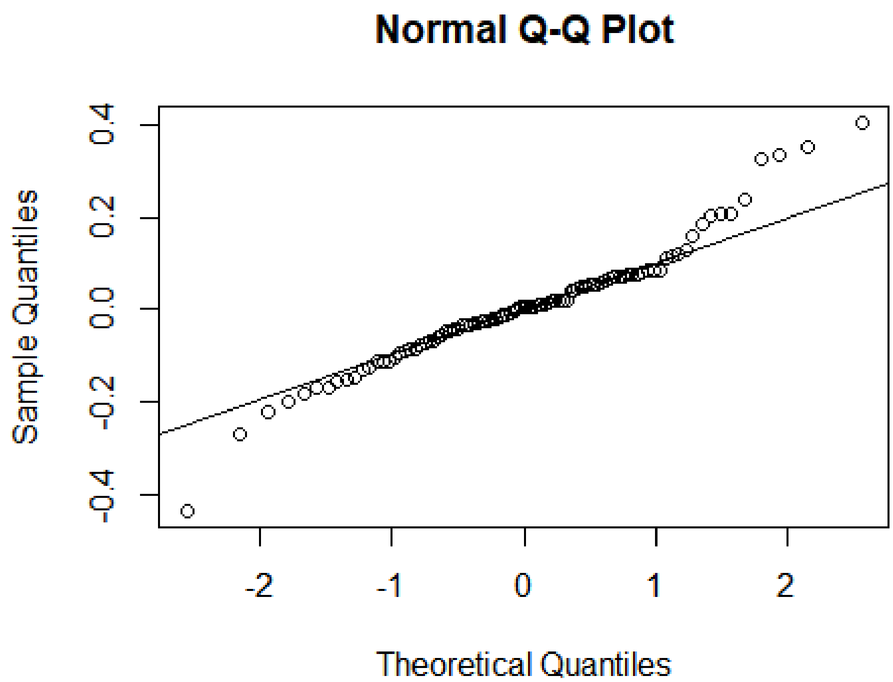
```
acf(res)
```



Here, most of the values are below the threshold line. Thus, we can conclude that the data errors are uncorrelated.

Assumption : Errors follow Normal Distribution

```
qqnorm(res);qqline(res)
```



Since most of the values lie on the line, the data is normally distributed.

Shapiro Wilk Test

```
shapiro.test(res)

##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.95683, p-value = 0.003292
```

The Shapiro-Wilk test also concludes that the data is normally distributed as p-value is less than 0.05

All assumptions are satisfied. Therefore, the model is the best fit

MMSE forecast

Stationarity Check

```
adf.test(data)

##
##  Augmented Dickey-Fuller Test
##
## data:  data
```

```
## Dickey-Fuller = -1.6079, Lag order = 4, p-value = 0.7393
## alternative hypothesis: stationary
```

Since the p-value is less than 0.05, the data is non-stationary

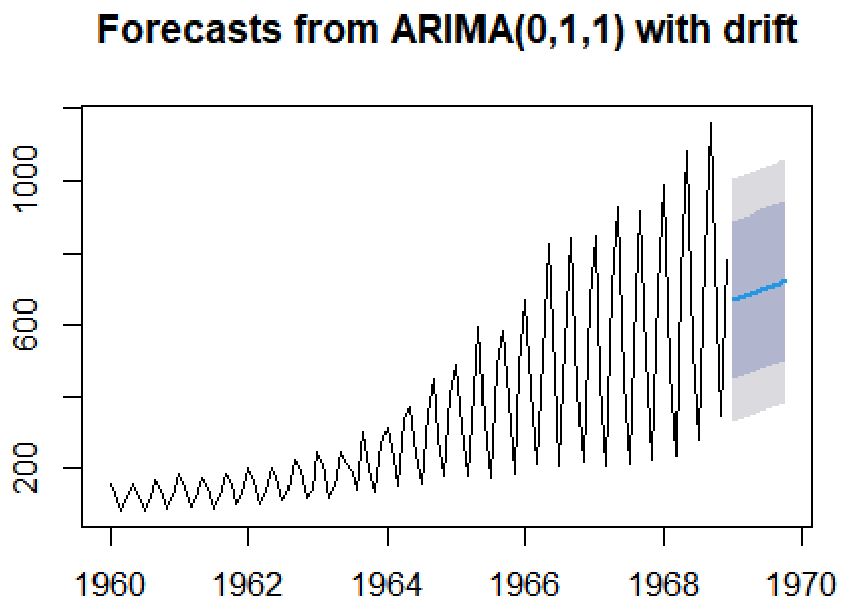
Fitting ARIMA Model

```
fit=auto.arima(data,seasonal="False")
fit

## Series: data
## ARIMA(0,1,1) with drift
##
## Coefficients:
##           ma1   drift
##       -0.9722  5.8899
## s.e.   0.0851  0.7852
##
## sigma^2 estimated as 29505:  log likelihood=-702.91
## AIC=1411.81   AICc=1412.04   BIC=1419.83
```

The best fitted arima model is ARIMA(0,1,1), which means that if we differenced the data set once, the model will be the simple AR(0,1) process.

Forecasting Values

```
forecast=forecast(fit,h=10)
forecast

##          Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## Jan 1969       669.2068 449.0625 889.3510 332.5252 1005.888
## Feb 1969       675.0967 454.8674 895.3259 338.2851 1011.908
## Mar 1969       680.9866 460.6723 901.3008 344.0450 1017.928
## Apr 1969       686.8765 466.4773 907.2757 349.8050 1023.948
## May 1969       692.7664 472.2823 913.2505 355.5650 1029.968
## Jun 1969       698.6563 478.0873 919.2254 361.3251 1035.988
## Jul 1969       704.5462 483.8923 925.2001 367.0852 1042.007
## Aug 1969       710.4361 489.6974 931.1749 372.8454 1048.027
## Sep 1969       716.3260 495.5025 937.1496 378.6056 1054.046
## Oct 1969       722.2160 501.3076 943.1243 384.3659 1060.066

plot(forecast)
```



The function has forecasted the 10 values for the future and plotted it. The blue part of the plot shows the forecasted values plot. We see that predicted values follow the general trend

Conclusion

We conclude that out of the methods used, Method of Differencing gave the best fit model as the AIC and BIC values were comparatively less. The best fit model was ARIMA(4,0,0) with 0 mean. We have also used the forecast function to predict the 0 future values. From the prediction we see that the values follow the general trend of the data.