PSTAT 274: Final Project (Study of CPI Time Series Data)

Mayuresh Anand

08/05/2020

ABSTRACT

In this report we are going to analyze and model time series data on consumer price index (CPILFESL). We analyze the data by looking at its plot to predict its tred and seasonalit and look at ACF/PACF to find what model could be best suited. To reduce variability we use Box-Jenkins approach. To ascertain the validity of our results we perform following on residuals of the fitted model Portmanteau tests, QQPlot, Shapiro-Wilk test and analyze the ACF/PACF

We choose amongs the models using AICc and Law of Parsimony and find that the given dataset fits better on ARIMA(8,1,7).

INTRODUCTION

We choose to study and model "Consumer Price Index for All Urban Consumers: All Items Less Food and Energy in U.S. City Average (CPILFESL)" dataset to forecast CPI index values.

This is a monthly dataset from from 1957 to current year i.e. 2020. But we choose to model the dataset from 1990-01-01 to 2010-12-01. This decision was made after we observed that dataset was not being able to be model accurately due to an outlier happening of 1980 great depression. But between these two time period the dataset is stable and hence the reason for this choice.

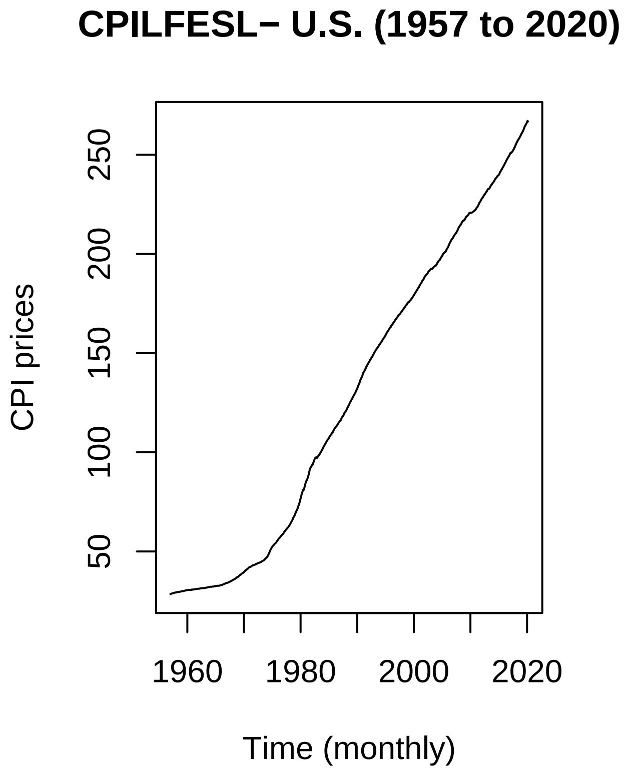
The Consumer Price Index (CPI) is a monthly measurement of U.S. prices for most household goods and services. The Bureau of Labor Statistics surveys the prices of 80,000 consumer items to create the index. It represents the prices of a cross-section of goods and services commonly bought by primarily urban household (representative of 87% of the U.S. population).

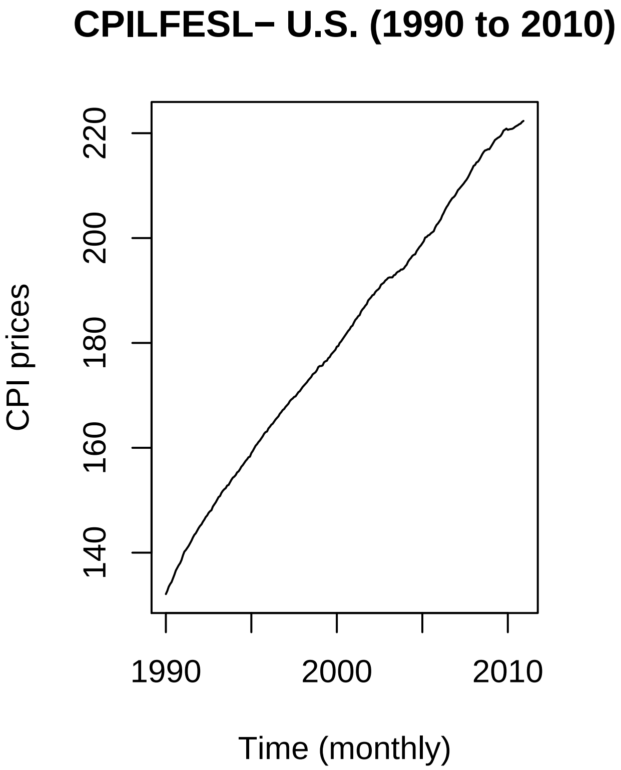
CPI is used to measure inflation or deflation in the economy. Inflation and deflation are a indicators of the performance of an economy and directly impacts standard of living of US population. For example, if inflation increases this means that the prices of goods are increasing and henceforth the affordability of products decreases unless the income of population also increases at a similar rate. It can be seen that if the inflation rate is high enough it can hurt economy. As products would cost more manufacturers produce less and this means that less manpower would be required. This would lead to condition that industry would have to lay off workers.

Hence, if one is able to predict the trend of CPI index then goverment and fed would be in a better position to make decisions governing the health of economy.

We have used R (statistical analysis tool) and Mozilla Firefox (web browser) while working on this project.

ANALYSIS OF TIME SERIES


CPILFESL− U.S. (1957 to 2020)

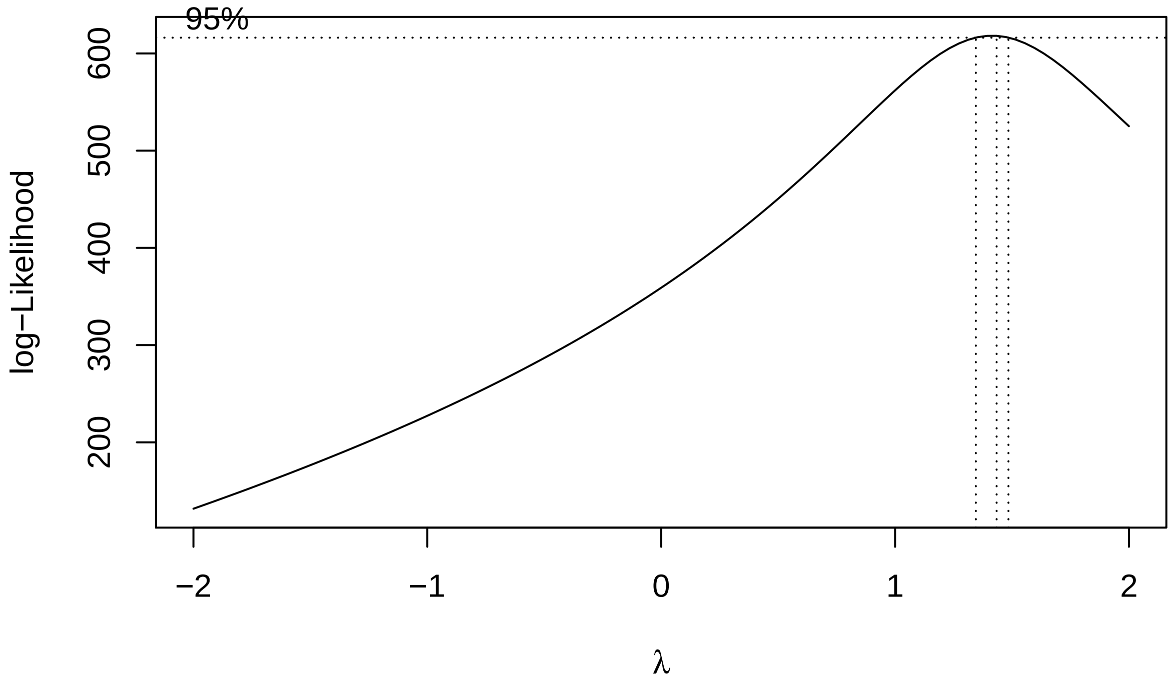

CPILFESL− U.S. (1990 to 2010)

Above we plot the times series dataset for CPILFESL. We can see that there is a change in nature of data from 1970 to 1990. We use this dataset to work on values from 1990(January) to 2010(December, inclusive), of which we shall reserve next two years i.e. 2011(January) and 2012(December inclusive) for testing the forecasted data.

OBSERVATIONS ON TIME SERIES

1. **TREND**: We find that this time series has increasing trend, most probbably linear.
2. **SEASONALITY**: There is no pattern of seasonality in the given time series
3. **CHANGES**: Changes were obeserved in original series from (1970 to 1990) therefore we choose to work on period later to that. I also think that variability of the data is changing. Hence, to confirm if the variability is changing or not and to stabilize it I use Box-Cox transform.

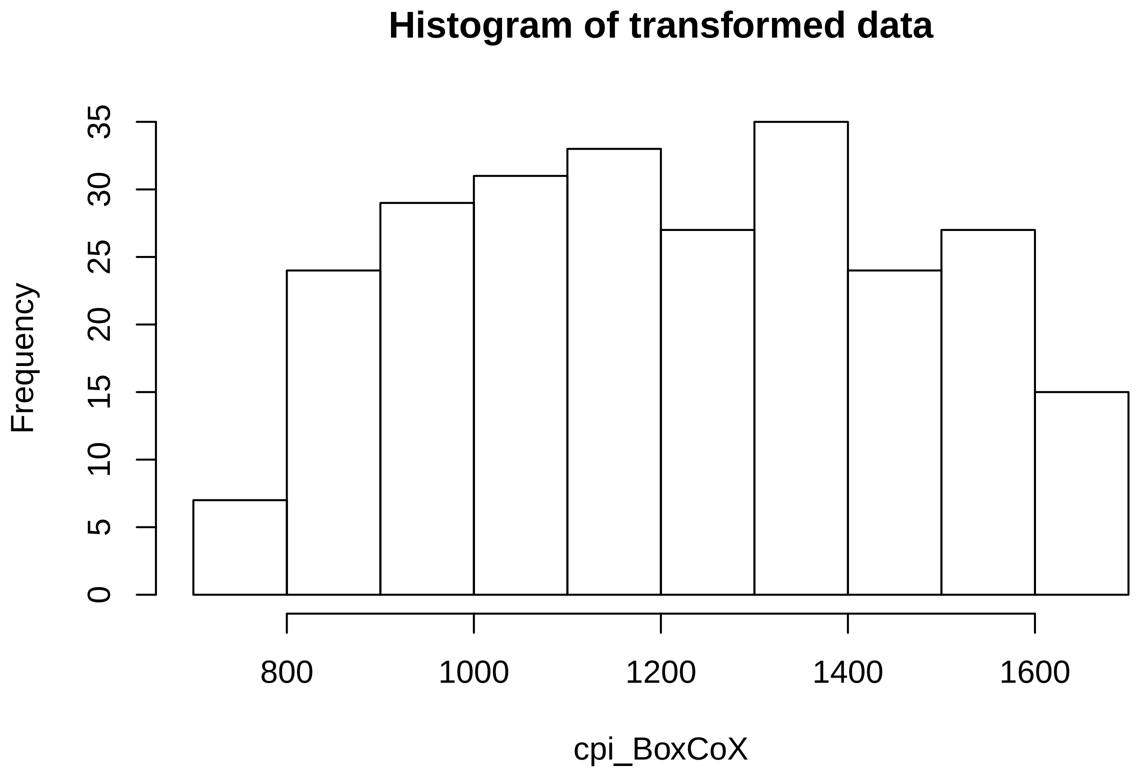
TRANSFORMATION OF DATA



As we can see here 0 is not included in the confidence interval therefore, I choose to tansform the data to lower its variability.

```
## [1] "Box-Cox transform gives lambda as below:"
```

```
## [1] 1.434343
```

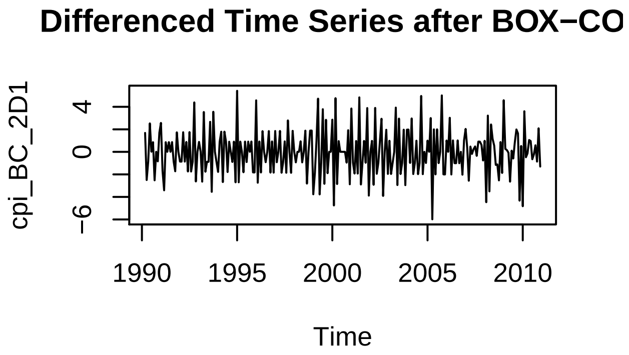


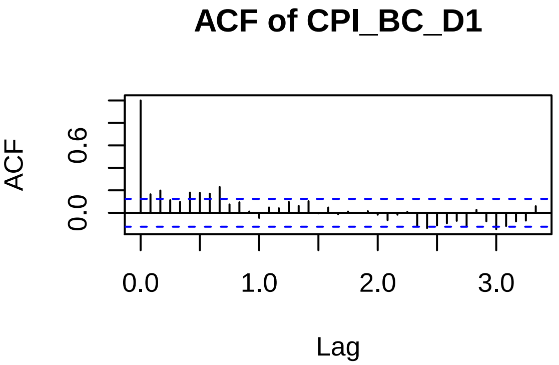
MAKING TIME SERIES STATIONARY

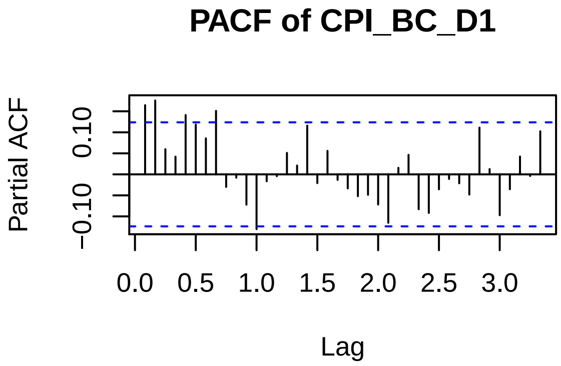
As it has no seasonality but trend, we are going to difference it two times successively at lag 1 and check the behavior of variance.

```
## [1] "Variance of the tansformed timeseries is:"

## [1] 59551.69

## [1] "Variance of first difference of tansformed timeseries is:"

## [1] 2.563912

## [1] "Variance of second difference of tansformed timeseries is:"

## [1] 4.28285
```

We see that after second differencing variance increases so we take data from only the first differencing. Now we analyze the behaviour of differenced data by plotting ACF/PACF and the plot of differenced data.







Looking at he ACF values outside CI are at lag 2,8 and PACF value outside CI are lag 8,12. Hence, we think this has $p = 8, 12$ and $q = 5, 8$. Thus, this could be $ARIMA(8, 1, 5)$, $ARIMA(8, 1, 8)$, $ARIMA(12, 1, 5)$, $ARIMA(12, 1, 8)$.

FITTING THE SERIES PREDICTED AND DIAGNOSTIC CHECKING

Below we are going to check the AICc value for each of the proposed models.

```
## [1] "AICC FOR ARIMA(8,1,5)"
```

```
## [1] 938.0116
```

```
## [1] "AICC FOR ARIMA(8,1,8)"
```

```
## [1] 943.7431
```

```
## [1] "AICC FOR ARIMA(12,1,5)"
```

```
## [1] 950.2218
```

```
## [1] "AICC FOR ARIMA(12,1,8)"
```

```
## [1] 949.5878
```

Looking at the AICc values we choose to go with ARIMA(8,1,5) as it has lowest AICc values. But then we find that this model doesn't produce causal and invertible model. Hence, we reject this and analyze ARIMA(8,1,8) model.

Note: Before picking these models I was working with dataset values upto 2020. But the model fit were not doing good at some years so I choose to decrease the dataset values. Earier models that were being fit were ARIMA(6,1,10), ARIMA(12,1,10), ARIMA (6,1,8) and ARIMA(12,1,8). These models passed all the Portmanteau BOX tests, histogram, white-noise, qqnorm tests but **failed on Shapiro-Wilk test**. Hence, I thought that I can decrease the data points and then try again.zz

PICKING MODEL AND FIXING COEFFICIENTS

Below is the details of model parameter and AICc Score for ARIMA(8,1,8) model we chose.

```
##
## Call:
## arima(x = cpi_BoxCoX, order = c(8, 1, 8), method = "ML")
##
## Coefficients:
##           ar1     ar2     ar3     ar4      ar5      ar6     ar7     ar8     ma1
##       -0.1426  0.3990  0.0722  0.1253  -0.1334  -0.4057  0.6407  0.4331  0.2468
## s.e.   0.3015  0.1426  0.1317  0.0913   0.0972   0.1133  0.1218  0.2399  0.3127
##           ma2      ma3      ma4     ma5     ma6      ma7      ma8
##       -0.2758  -0.0239  -0.0915  0.2515  0.6419  -0.5031  -0.2630
## s.e.   0.1417   0.1366   0.0950  0.0860  0.1405   0.1819   0.2504
##
## sigma^2 estimated as 2.1:  log likelihood = -453.71,  aic = 941.43
```

```
## [1] "Below is the AICc Score"
```

```
## [1] 943.7431
```

We check for the confidence intervals and fix the coeffecients inside the C.I. to be 0. Following is the are parameters fixed = c(NA,NA,0,0,0,NA,NA,NA, 0,0,0,0,NA,NA,NA,0), method="ML").

```
##
## Call:
## arima(x = cpi_BoxCoX, order = c(8, 1, 8), fixed = c(NA, NA, 0, 0, 0, NA, NA,
##     NA, 0, 0, 0, 0, NA, NA, NA, 0), method = "ML")
##
## Coefficients:
##          ar1     ar2  ar3  ar4  ar5      ar6     ar7     ar8  ma1  ma2  ma3
##       0.1078  0.1726    0    0    0  -0.1192  0.6452  0.1799    0    0    0
## s.e.  0.0622  0.0537    0    0    0   0.1974  0.1692  0.0807    0    0    0
##       ma4     ma5     ma6      ma7  ma8
##         0  0.1363  0.2968  -0.4931    0
## s.e.    0  0.0566  0.2015   0.1455    0
##
## sigma^2 estimated as 2.253:  log likelihood = -460.02,  aic = 938.04
```

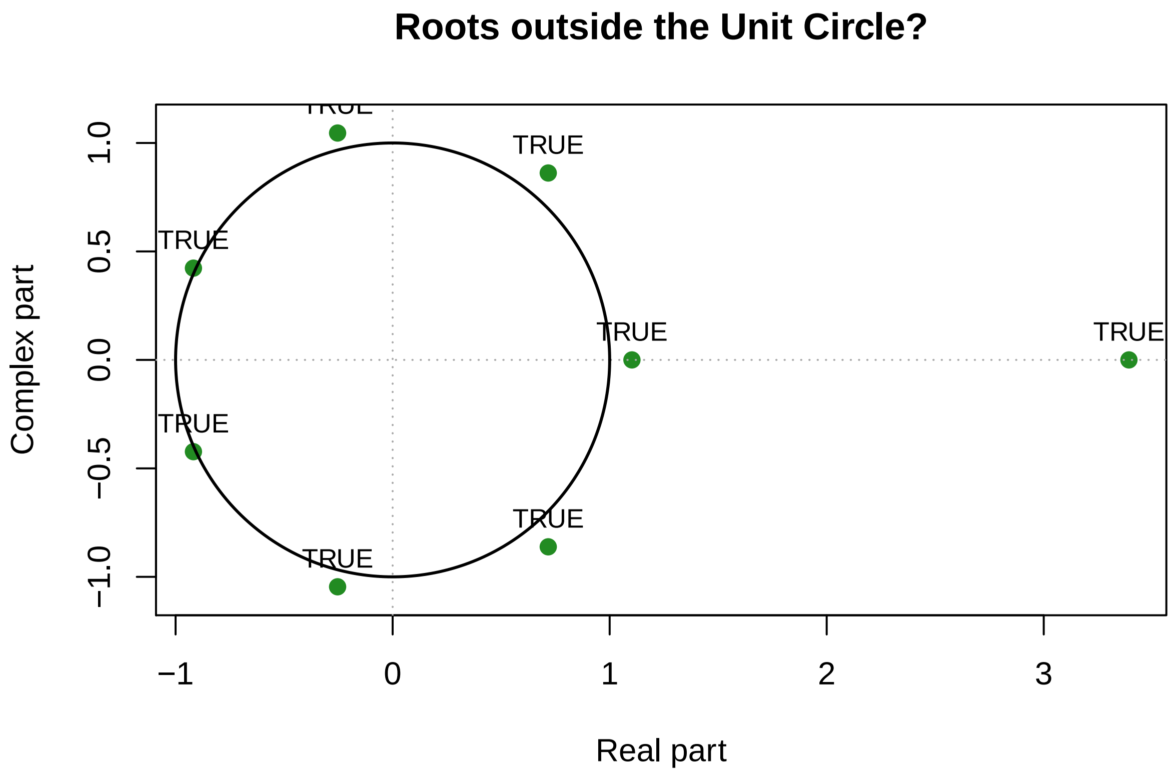
```
## [1] "Below is the AICc value of the data:"
```

```
## [1] 940.359
```

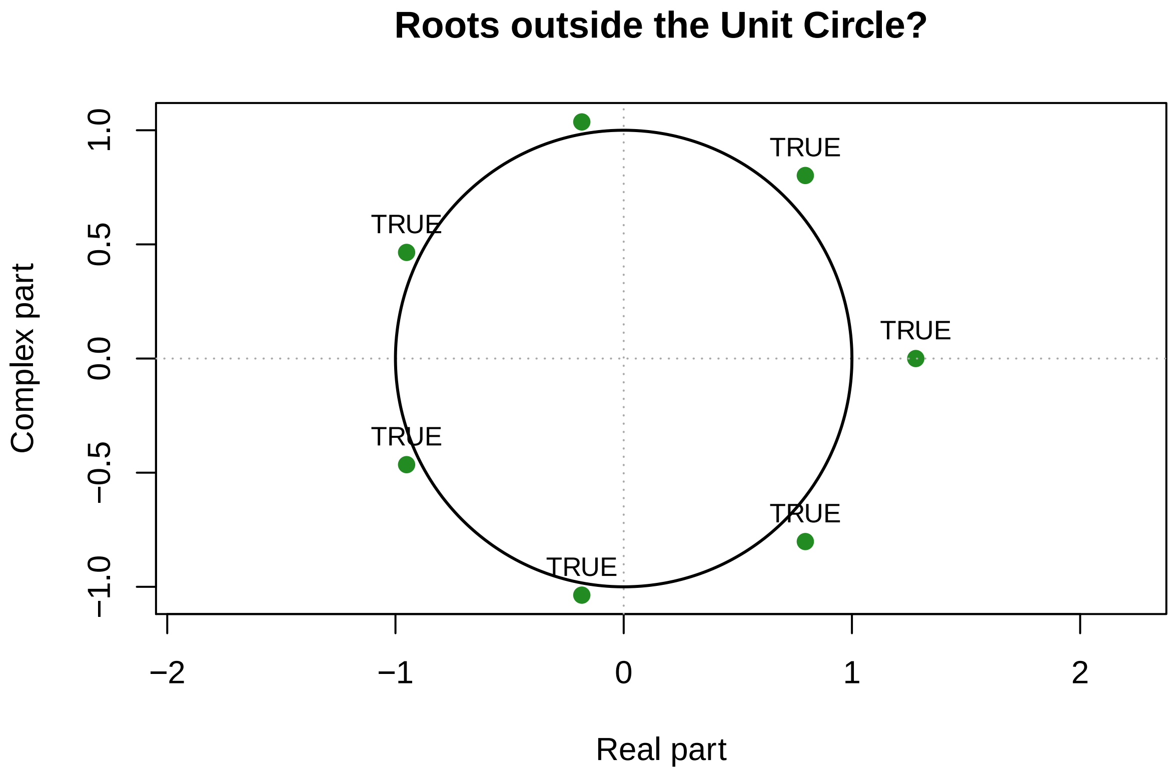
We see that the AICc values decreases after making coefficients falling inside the confidence interval to be 0.

CHECKING IF THE MODEL IS CAUSAL AND INVERTIBLE

```
##        real   complex outside
## 1  0.717001  0.861588    TRUE
## 2 -0.917857  0.422994    TRUE
## 3 -0.917857 -0.422994    TRUE
## 4  0.717001 -0.861588    TRUE
## 5 -0.253541  1.045768    TRUE
## 6 -0.253541 -1.045768    TRUE
## 7  1.102646  0.000000    TRUE
## 8  3.392585  0.000000    TRUE
## *Results are rounded to 6 digits.
```



```
##        real   complex outside
## 1  0.795866  0.801757    TRUE
## 2 -0.951180  0.464860    TRUE
## 3 -0.183577 -1.036427    TRUE
## 4  0.795866 -0.801757    TRUE
## 5 -0.183577  1.036427    TRUE
## 6 -0.951180 -0.464860    TRUE
## 7  1.279690  0.000000    TRUE
## *Results are rounded to 6 digits.
```



We see that the model we have chosen is **causal** and **invertible** as all the roots lies inside the unit circle for both MA and AR parts.

TRYING TO REDUCE PARAMETERS (Checking possible MA, AR and ARIMA models)

We want to check if more coefficients can be removed, so we iterate over various permutations of choices for remaining 8 parameters and check if that produces lower AICc values.

Below is the matrix of all possible 256 models and their AICc values:

```
##   [1] 1076.4551  998.4402 1077.1153  980.9109 1078.4549 1000.1874 1079.1083
##   [8]  982.7959 1078.1212 1000.0566 1078.7190  982.7494 1080.0123 1002.0137
##  [15] 1080.6421  984.7268 1074.3591  996.3088 1075.0157  978.7579 1076.3383
##  [22]  998.2001 1077.0096  980.7262 1076.0719  998.0176 1076.7878  980.1183
##  [29] 1078.0390  999.4187 1078.7395  981.5154 1077.8316  999.9181 1078.3206
##  [36]  982.9015 1079.6001 1001.6692 1079.9945  984.7845 1079.5002 1001.7561
##  [43] 1080.0041  984.7458 1081.4706 1003.6465 1081.9248  986.7213 1075.7822
##  [50]  997.8461 1076.4040  980.5724 1077.7055  999.7414 1078.2970  982.5401
##  [57] 1077.5017  999.1069 1078.1344  981.7409 1078.8871  999.9904 1079.3794
##  [64]  982.9707 1078.4549 1000.3833 1079.1102  982.1179 1080.4549 1002.0728
##  [71] 1077.5315  983.5413 1080.0264 1001.9912 1080.5609  983.9642 1081.8955
##  [78] 1003.4936 1079.5306  979.5727 1076.3522  997.6729 1077.0152  979.9073
##  [85] 1078.1774  998.9655 1078.8565  980.8306 1078.0106  999.2369 1078.7100
##  [92]  981.2353 1079.8828 1000.1687 1080.6005  981.0564 1079.4915 1001.6898
##  [99] 1080.0844  983.9718 1080.5137 1003.6630 1079.4580  984.8994 1081.4123
## [106] 1003.5556 1081.9661  985.8430 1082.4782 1004.2261 1081.4279  977.3642
## [113] 1077.5827  998.7293 1078.1697  981.3134 1078.5743  997.4549 1079.2675
## [120]  980.6518 1078.3706  999.2906 1078.8558  982.2781 1078.9800  998.4309
## [127] 1079.6560  979.3392 1078.0966 1000.1774 1078.8360  981.9283 1080.0087
## [134] 1001.3184 1080.8040  982.5646 1079.7819  998.7877 1080.0665  980.2460
## [141] 1081.5859 1000.4766 1081.6275  981.4694 1076.1937  998.1185 1076.8606
## [148]  979.9714 1078.1899  999.6776 1078.8462  981.2021 1077.6007  999.6197
## [155] 1078.6264  981.9377 1079.4464 1000.7912 1080.4767  983.1768 1079.5476
## [162] 1000.8951 1080.1731  983.6334 1081.5056 1000.8173 1081.9934  983.4601
## [169] 1080.5817  998.8700 1080.0309  980.4818 1082.5651  999.6772 1081.9509
## [176]  980.6953 1077.5536  999.0437 1078.1322  981.2361 1079.0081  999.8174
## [183] 1079.4241  981.7085 1079.3608 1000.8466 1080.1321  982.4805 1080.7472
## [190] 1001.6764 1081.3781  982.6866 1080.0234 1002.0941 1080.7459  983.0107
## [197] 1081.8870 1001.0126 1079.4964  983.1891 1082.0253 1000.3850 1081.5429
## [204]  981.1944 1083.6459 1001.0687       Inf  981.5708 1078.1777  999.4033
## [211] 1078.8284  981.0045 1080.0363  999.9342 1080.7030  982.8176 1079.4050
## [218] 1000.6693 1080.4303  982.9234 1081.2909 1002.1577 1082.3062  982.8859
## [225] 1081.4364 1002.2514 1082.0669  984.1396 1082.4926 1000.9655 1081.4579
## [232]  984.1270 1082.5510  999.2482 1081.9948  980.0787 1083.5369  999.7130
## [239]       Inf  978.7593 1078.5938  999.1751 1079.0079  981.2846 1079.5448
## [246]  998.8224 1080.0751  982.5936 1080.1898 1001.1439 1080.8192  982.0785
## [253] 1080.2754 1000.4240 1081.0950  980.4578
```

We find that none of the model choices produce lower value of AICc than the one which we fixed using the confidence interval values. Hence, we move ahead with the model that we have and needs no updates.

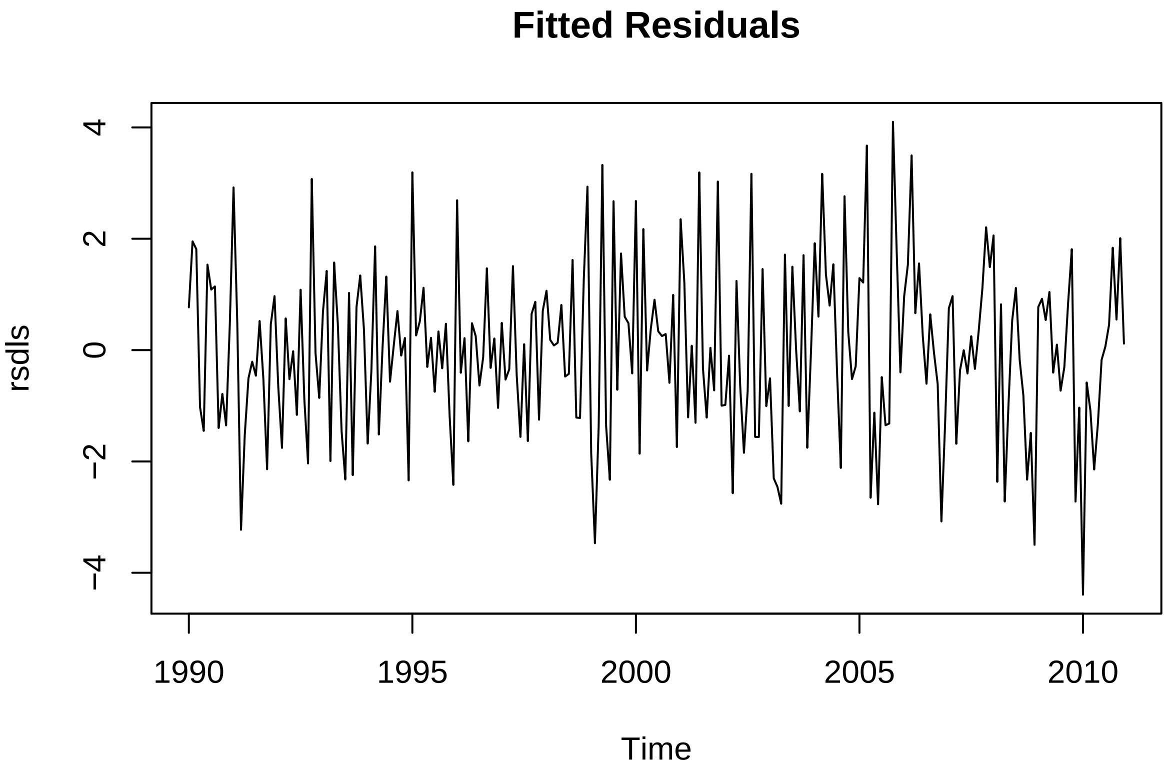
ALGEBRAIC EQUATION OF THE MODEL

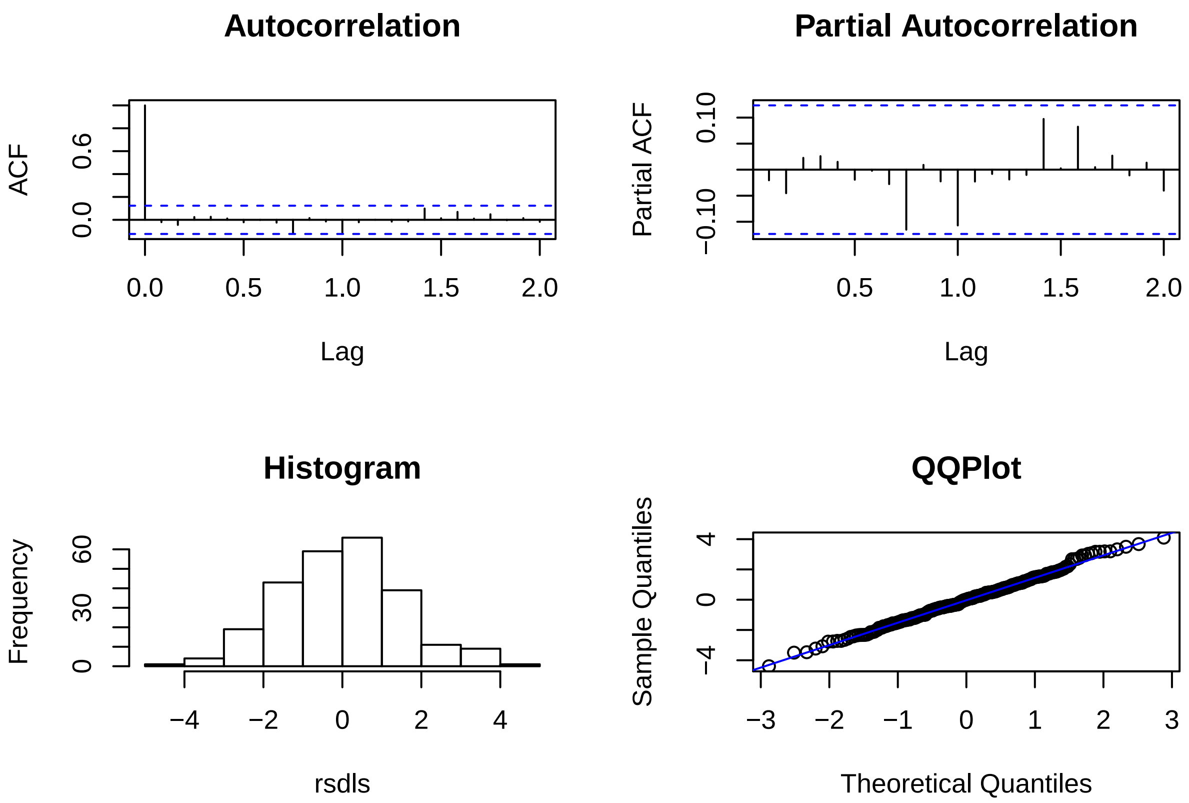
$X_t = 0.1078X_{t-1} + 0.1726X_{t-2} - 0.1192X_{t-6} + 0.6452X_{t-7} + 0.1799X_{t-8} + Z_t - 0.1363Z_{t-5} + 0.2968Z_{t-6} - 0.4931Z_{t-7}$

So, we have ARIMA(8,1,7) model finally which is **causal** and **invertible**.

PLOT OF RESIDUALS AND DIAGNOSTICS

Below is the plot for the residuals of the chosen model. We find that plot looks like that of white noise.


Fitted Residuals
rsdls
4
2
0
−2
−4
1990
1995
2000
2005
2010
Time



SHAPIRO-WILK NORMALITY TEST

```r
shapiro.test(rsdls)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rsdls
## W = 0.99604, p-value = 0.7753
```

BOX-LJUNG TEST

```r
Box.test(rsdls, type = "Ljung-Box", lag = 16, fitdf = 8)
```

```
##
##  Box-Ljung test
##
## data:  rsdls
## X-squared = 8.3886, df = 8, p-value = 0.3965
```

BOX-PIERCE TEST

```
Box.test(rsdls, type = "Box-Pierce", lag = 16, fitdf = 8)
```

```
##
##  Box-Pierce test
##
## data:  rsdls
## X-squared = 8.0075, df = 8, p-value = 0.4327
```

McLeod-Li SQUARE TEST

```
Box.test(rsdls^2, type = "Box-Pierce", lag = 16, fitdf = 0)
```

```
##
##  Box-Pierce test
##
## data:  rsdls^2
## X-squared = 18.672, df = 16, p-value = 0.2861
```

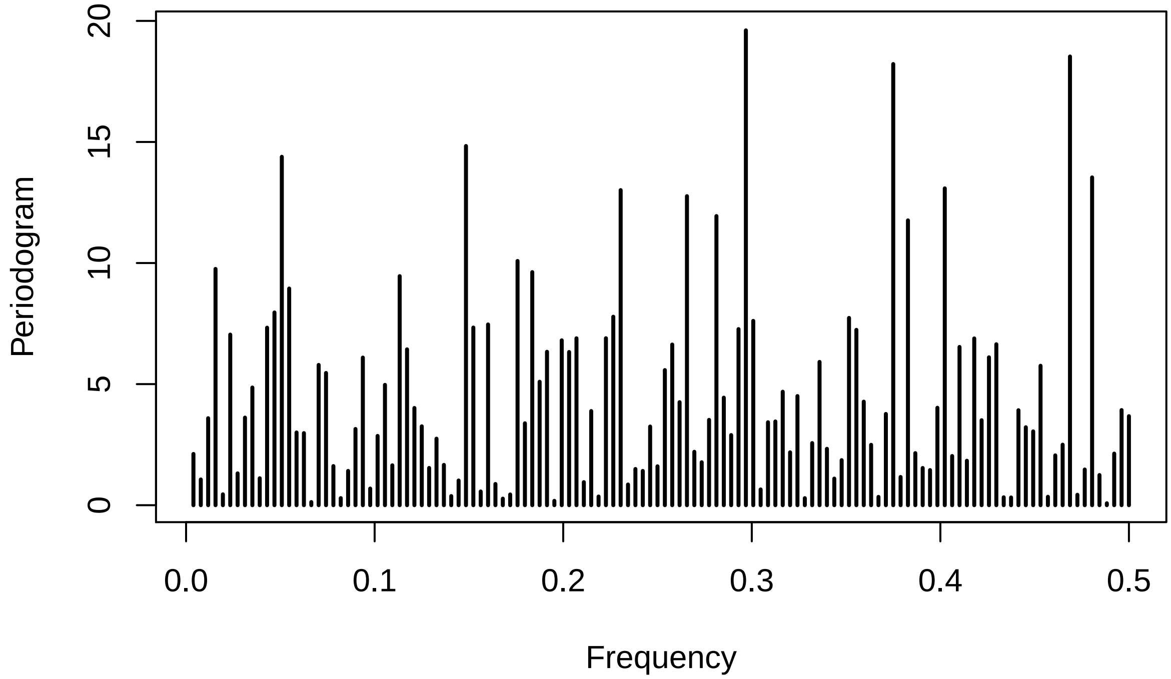
As for all the tests p-value is greater than 0.05 we cannot reject the hypothesis that this is a white noise and hence we are going to **ACCEPT** this model.

So, we have fitted models ARIMA(8,1,7) which has lesser number of parameters and AICc value than that of ARIMA(8,1,8) choosen earlier.

SPECTRAL ANALYSIS

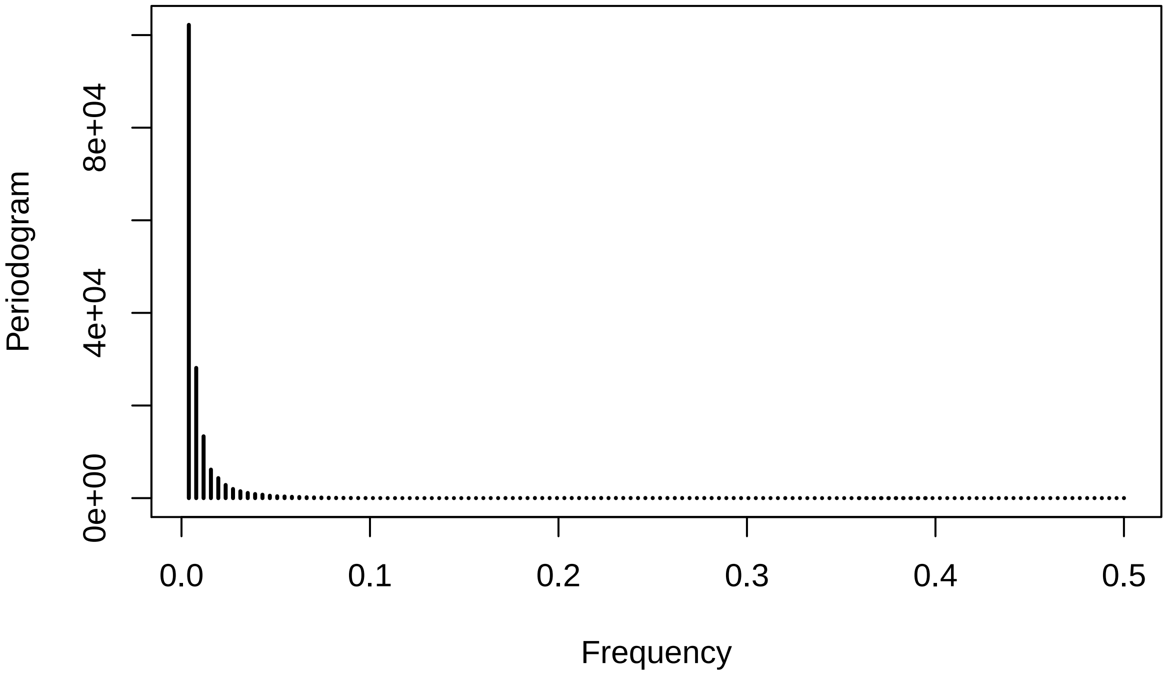
PERIODOGRAM

This is the periodogram on the residuals for the data.



We find that there is no dominant frequencies among the plotted frequencies.

Below is the periodogram for the orignal data

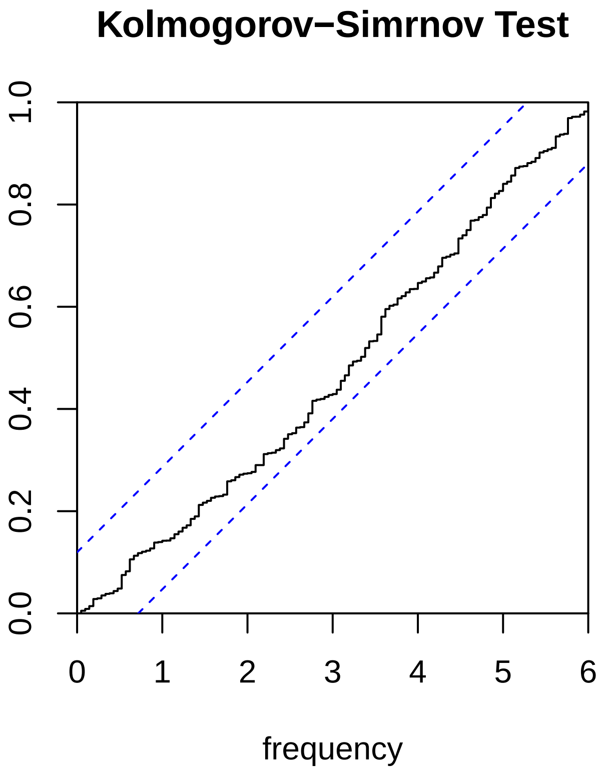


So we can see that original data has no period and this solifies our original observation.

FISHER's TEST

```
## [1] 0.8300579
```

KOLMOGOROV-SIMRNOV TEST



In the above tests involving spectral analysis we see that p-value for Fisher's test is greater than 0.05 hence we cannot reject the hypothesis of this being white noise. Also, values for Kolmogorov-Smirnov test lies withing the confidence interval.

FORECASTING

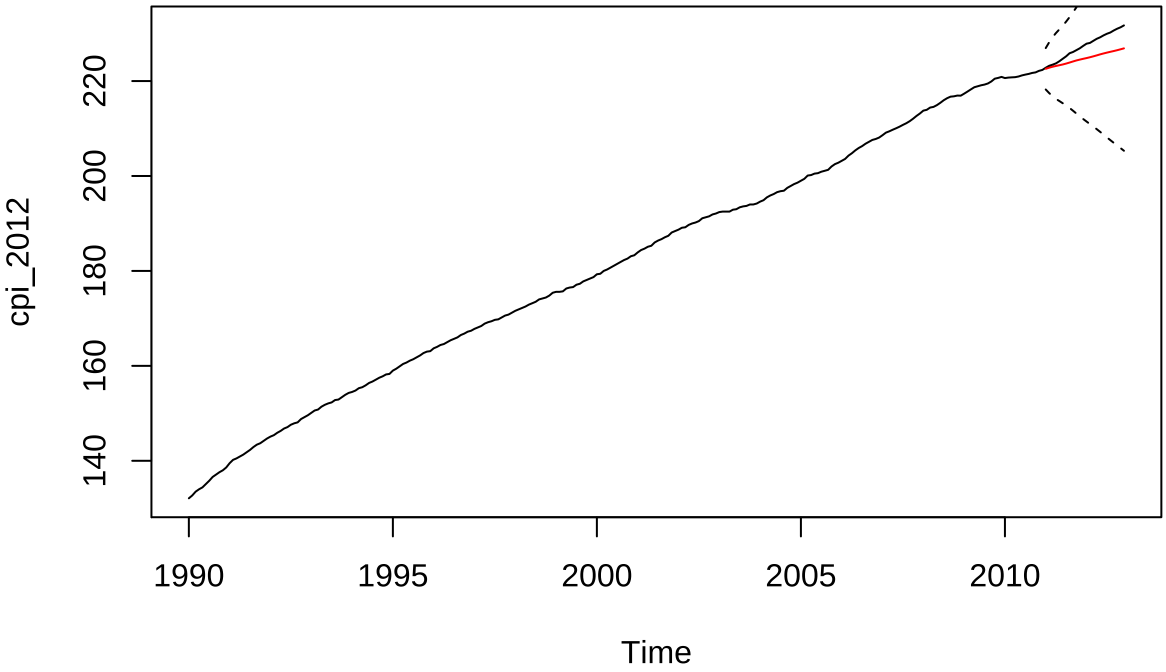
```
forecast_series <-predict(fit, n.ahead = 24)
values = (((forecast_series$pred)*lambda+1)^(1/lambda))
errors = (((forecast_series$se)*lambda+1)^(1/lambda))
#errors = forecast_series$se
print(values)
```

```
##             Jan      Feb      Mar      Apr      May      Jun      Jul      Aug
## 2011 222.5891 222.8161 223.0211 223.1794 223.3393 223.4961 223.6879 223.9018
## 2012 224.8450 225.0171 225.2083 225.4143 225.6259 225.8192 225.9993 226.1625
##             Sep      Oct      Nov      Dec
## 2011 224.1238 224.3384 224.5195 224.6887
## 2012 226.3273 226.5026 226.6917 226.8925
```

```
print(errors)
```

```
##              Jan       Feb       Mar       Apr       May       Jun       Jul
## 2011  2.226846  2.725536  3.177392  3.538009  3.848066  4.167987  4.515073
## 2012  6.823002  7.183445  7.547413  7.932187  8.330127  8.729997  9.117879
##            Aug       Sep       Oct       Nov       Dec
## 2011  4.878062  5.285932  5.700259  6.097923  6.465946
## 2012  9.495328  9.869140 10.240563 10.620051 11.008883
```

```
cpi_2012 = ts(data[,2], start=c(1990,1), end=c(2012,12),frequency = 12)
ts.plot(cpi_2012)
lines(values,lty=1,col="red")
lines(values+1.96*errors,lty=2)
lines(values-1.96*errors,lty=2)
```



APPENDIX [CODE SECTION]

```
#Readding the data from comma seperated value file CPILFESL.csv

data_original <- read.table("CPILFESL.csv", header=TRUE, sep=",")
data <- read.table("CPILFESL1.csv", header=TRUE, sep=",")

#Load the dataset with the start date and frequency
cpi_original = ts(data_original[,2], start=c(1957,1), end=c(2020,3),
                  frequency = 12)
cpi = ts(data[,2], start=c(1990,1), end=c(2010,12),frequency = 12)

#Plot the time series data
```

```r
op = par(mfrow=c(1,2))
ts.plot(cpi_original, main="CPILFESL- U.S. (1957 to 2020)",
        xlab="Time (monthly)", ylab="CPI prices")

ts.plot(cpi, main="CPILFESL- U.S. (1990 to 2010)", xlab="Time (monthly)",
        ylab="CPI prices")
par(op)

# BOX-COX Transformations
lambda = bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
print("Box-Cox transform gives lambda as below:")
print(lambda)
cpi_BoxCoX = (1/lambda)*(cpi^lambda-1)

#Checking Variances of the time series
options(warn=-1)
print("Variance of the tansformed timeseries is:")
print(var(cpi_BoxCoX))
print("Variance of first difference of tansformed timeseries is:")
cpi_BC_D1 = diff(cpi_BoxCoX, lag = 1)
print(var(cpi_BC_D1))
cpi_BC_2D1 = diff(cpi_BC_D1, lag = 1)
print("Variance of second difference of tansformed timeseries is:")
print(var(cpi_BC_2D1))

# BOX COX of differenced timeseries
plot(cpi_BC_2D1, main="Differenced Time Series after BOX-COX")
op = par(mfrow=c(1,2))
acf(cpi_BC_D1, lag.max = 40,main="ACF of CPI_BC_D1")
pacf(cpi_BC_D1, lag.max = 40,main="PACF of CPI_BC_D1")

#Finding AICc values for the chosen model
library(qpcR)
print("AICC FOR ARIMA(8,1,5)")
print(AICc(arima(cpi_BoxCoX, order=c(8,1,5), method="ML")))
print("AICC FOR ARIMA(8,1,8)")
print(AICc(arima(cpi_BoxCoX, order=c(8,1,8), method="ML")))
print("AICC FOR ARIMA(12,1,5)")
print(AICc(arima(cpi_BoxCoX, order=c(12,1,5), method="ML")))
print("AICC FOR ARIMA(12,1,8)")
print(AICc(arima(cpi_BoxCoX, order=c(12,1,8), method="ML")))

# We want to fix CPI of BOX COX
fit <- arima(cpi_BoxCoX, order=c(8,1,8), method="ML")
rsdls = residuals(fit)
fit
print("Below is the AICc Score")
AICc(fit)

# We want to fix CPI of BOX COX FOR ESIMATED PARAMETES
fit <- arima(cpi_BoxCoX, order=c(8,1,8),
      fixed = c(NA,NA,0,0,0,NA,NA,NA,  0,0,0,0,NA,NA,NA,0 ), method="ML")
rsdls = residuals(fit)
```

```r
fit
print("Below is the AICc value of the data:")
AICc(fit)

# Checking if the model roots lies in unit circle or not
source("plot.roots.R")
plot.roots(NULL,polyroot(c(1 , 0.1078 , 0.1726  ,  0  ,  0  ,  0  , -0.1192 , 0.6452  , 0.1799)),
           main="Roots of AR part")
plot.roots(NULL,polyroot(c(1, 0, 0, 0, 0, 0.1363, 0.2968, -0.4931, 0)),
           main="Roots of MA part")

# Iterating over all the possible choices of the model parameters
parameters <- 8
w <- expand.grid(rep(list(0:1),parameters))
for(i in 1:(2^parameters)){
  for(j in 1:parameters){
    if(w[i,j]==1){
      w[i,j] = NA
    }
  }
}

calcAICc <- c(rep(Inf,2^parameters))

for(i in 1:2^8){
  x <- i
  w[x,]
  ar1 <- w[x,1]
  ar2 <- w[x,2]
  ar3 <- 0
  ar4 <- 0
  ar5 <- 0
  ar6 <- w[x,3]
  ar7 <- w[x,4]
  ar8 <- w[x,5]
  ma1 <- 0
  ma2 <- 0
  ma3 <- 0
  ma4 <- 0
  ma5 <- w[x,6]
  ma6 <- w[x,7]
  ma7 <- w[x,8]
  ma8 <- 0

  try(
    {
      calcAICc[i]<- AICc(arima(cpi_BC_D1, order=c(8,1,8),
      fixed = c(ar1,ar2,ar3,ar4,ar5,ar6,ar7,ar8, ma1,ma2,ma3,ma4,ma5,ma6,ma7,ma8),
      method="ML"))
    }
    , silent = TRUE
   )
}
```

```r
print(calcAICc)
#minAICc <- which(calcAICc < 943.7431)
#print(minAICc)
#print("ar1 ar3 ar4  ar5 ar8  ma1  ma3  ma4  ma8")
#for(i in 1:length(minAICc)){
#  print(w[minAICc[i],])
#  print(calcAICc[minAICc[i]])
#}

# We want to fix CPI of BOX COX FOR ESIMATED PARAMETES
fit <- arima(cpi_BoxCoX, order=c(8,1,8),
      fixed = c(NA,NA,0,0,0,0,NA,NA,NA,  0,0,0,0,NA,NA,NA,0 ), method="ML")
rsdls = residuals(fit)
fit
print("Below is the AICc value of the data:")
AICc(fit)

# Plot of residual
plot(rsdls, main="Fitted Residuals")

# Plot of ACF, PACF, HISTOGRAM and QQPLOT
op <- par(mfrow=c(2,2))
acf(rsdls,main = "Autocorrelation")
pacf(rsdls,main = "Partial Autocorrelation")
hist(rsdls,main = "Histogram")
qqnorm(rsdls,main = "QQPlot")
qqline(rsdls,col="blue")
par(op)

# SHAPIRO WILK NORMALITY TEST
shapiro.test(rsdls)

# Ljung-Box test
Box.test(rsdls, type = "Ljung-Box", lag = 16, fitdf = 8)

# Box-Pierce test
Box.test(rsdls, type = "Box-Pierce", lag = 16, fitdf = 8)

# McLeod-Li test
Box.test(rsdls^2, type = "Box-Pierce", lag = 16, fitdf = 0)

# SPECTRAL ANALYSIS

# P

# FISHER's TEST
library("GeneCycle")
fisher.g.test(rsdls)

## KOLMOGOROV-SIMRNOV TEST
cpgram(rsdls,main="Kolmogorov-Simrnov Test")

# Forecasting ahead
```

```
forecast_series <-predict(fit, n.ahead = 24)
values = (((forecast_series$pred)*lambda+1)^(1/lambda))
errors = (((forecast_series$se)*lambda+1)^(1/lambda))
#errors = forecast_series$se
print(values)
print(errors)
cpi_2012 = ts(data[,2], start=c(1990,1), end=c(2012,12),frequency = 12)
ts.plot(cpi_2012)
lines(values,lty=1,col="red")
lines(values+1.96*errors,lty=2)
lines(values-1.96*errors,lty=2)
```