timeseries(arima model)

2023-03-09

#loading file

library(readxl)  
gdp <- read\_excel("/Users/nirajkc/Desktop/gdp.xlsx")  
View(gdp)

class(gdp)

## [1] "tbl\_df" "tbl" "data.frame"

To check class of dataset, R is not recognizing the time series, Changing into timeseries with ts(), start with min date and max date, frequency is 4 as it is quaterly data

gdptime=ts(gdp$GDP, start = min(gdp$DATE), end = max(gdp$DATE), frequency = 4)

###to check class now###

class(gdptime)

## [1] "ts"

##package instillation for arima model

install.packages(“tseries”) install.packages(“forecast”)

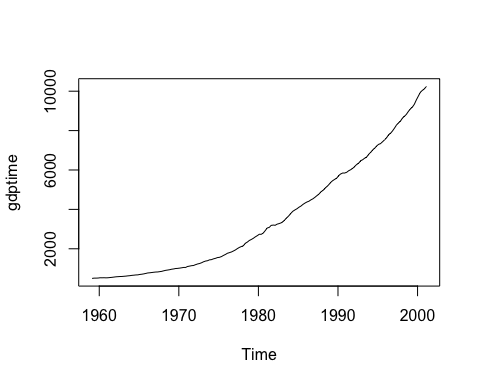
library(forecast)

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

library(tseries)

plotting line graph

plot(gdptime)



The ARIMA (Autoregressive Integrated Moving Average) model has some assumptions that should be met in order to obtain accurate and reliable results. Here are some of the key assumptions of the ARIMA model: Stationarity: The time series should be stationary, which means that its statistical properties should not change over time. This includes a constant mean, constant variance, and constant autocorrelation structure. If the time series is non-stationary, it can be made stationary by applying differencing or other transformations.

Linearity: The relationship between the time series and its past values should be linear. This means that the model assumes that the effects of past values on the current value are additive and proportional to the time lag.

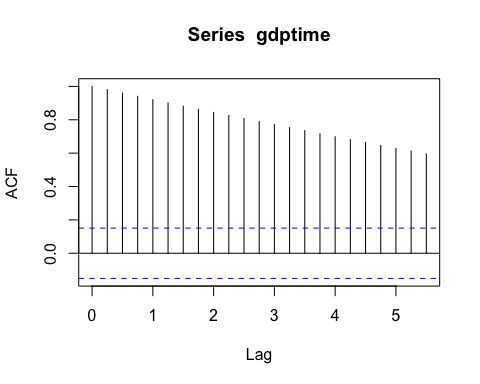
Normality: The residuals of the ARIMA model should be normally distributed with mean zero and constant variance. This assumption can be checked using diagnostic plots and statistical tests.

Independence: The residuals should be independent of each other, meaning that there should be no correlation between the residuals at different time lags.

Parsimony: The model should be as simple as possible while still capturing the essential features of the time series. This means that the model should not be overparameterized or include unnecessary variables.

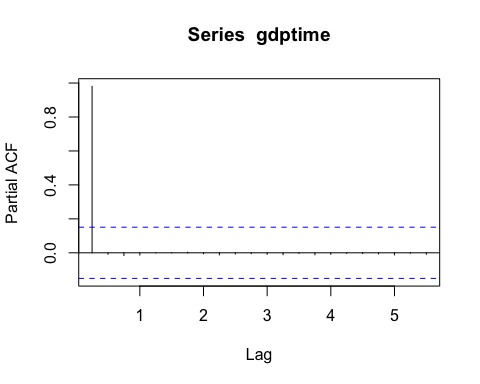
###First checking auto correlation spikes line is crossing blue line, so not Stationarity.

acf(gdptime)

 spikes line is crossing blue line, so not Stationarity.

partial auto correlation, have no issue, lines not crossing blue lines

pacf(gdptime)

 Augmented Dickey-Fuller (ADF) test if the p-value is greater than the level of significance, we fail to reject the null hypothesis and conclude that the time series is non-stationary.

adf.test(gdptime)

## Warning in adf.test(gdptime): p-value greater than printed p-value

##   
## Augmented Dickey-Fuller Test  
##   
## data: gdptime  
## Dickey-Fuller = 0.95078, Lag order = 5, p-value = 0.99  
## alternative hypothesis: stationary

Now we need to make data stationary. auto.arima() function will take care of that.

gdpmodel=auto.arima(gdptime,ic="aic", trace=TRUE)

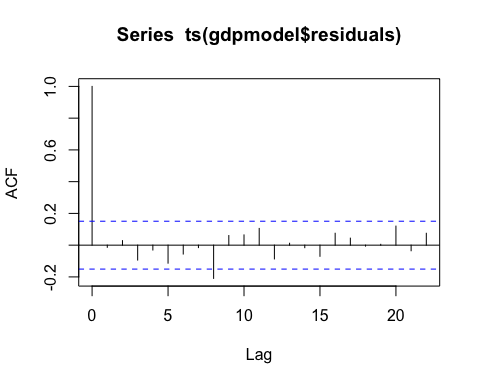
##   
## Fitting models using approximations to speed things up...  
##   
## ARIMA(2,2,2)(1,0,1)[4] : 1581.279  
## ARIMA(0,2,0) : 1624.378  
## ARIMA(1,2,0)(1,0,0)[4] : 1606.306  
## ARIMA(0,2,1)(0,0,1)[4] : 1574.484  
## ARIMA(0,2,1) : 1573.729  
## ARIMA(0,2,1)(1,0,0)[4] : 1578.73  
## ARIMA(0,2,1)(1,0,1)[4] : 1578.292  
## ARIMA(1,2,1) : 1569.636  
## ARIMA(1,2,1)(1,0,0)[4] : 1575.176  
## ARIMA(1,2,1)(0,0,1)[4] : 1571.602  
## ARIMA(1,2,1)(1,0,1)[4] : Inf  
## ARIMA(1,2,0) : 1600.506  
## ARIMA(2,2,1) : 1572.312  
## ARIMA(1,2,2) : 1571.632  
## ARIMA(0,2,2) : 1569.263  
## ARIMA(0,2,2)(1,0,0)[4] : 1575.068  
## ARIMA(0,2,2)(0,0,1)[4] : 1571.263  
## ARIMA(0,2,2)(1,0,1)[4] : 1576.052  
## ARIMA(0,2,3) : 1570.254  
## ARIMA(1,2,3) : 1572.676  
##   
## Now re-fitting the best model(s) without approximations...  
##   
## ARIMA(0,2,2) : 1584.116  
##   
## Best model: ARIMA(0,2,2)

gdpmodel

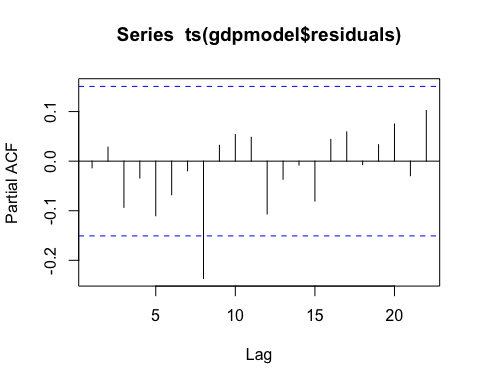
## Series: gdptime   
## ARIMA(0,2,2)   
##   
## Coefficients:  
## ma1 ma2  
## -0.6526 -0.1756  
## s.e. 0.0713 0.0675  
##   
## sigma^2 = 748.2: log likelihood = -789.06  
## AIC=1584.12 AICc=1584.26 BIC=1593.47

Now again check auto correction.

acf(ts(gdpmodel$residuals))



pacf(ts(gdpmodel$residuals))



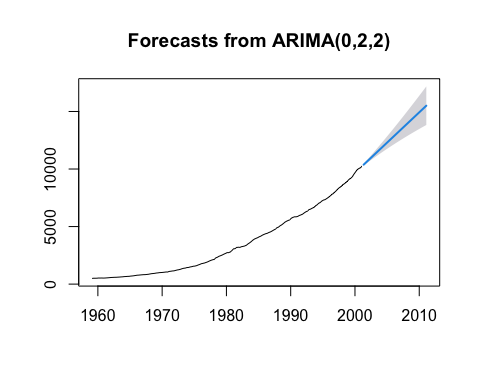
##Now doing forecast, confidence interval=95%, forecast 10 years for 4 quaters

mygdpforecast=forecast(gdpmodel, level = c(95), h=10\*4)  
mygdpforecast

## Point Forecast Lo 95 Hi 95  
## 2001.35 10376.53 10322.92 10430.15  
## 2001.60 10507.98 10418.02 10597.94  
## 2001.85 10639.42 10518.06 10760.78  
## 2002.10 10770.87 10619.38 10922.36  
## 2002.35 10902.31 10720.85 11083.77  
## 2002.60 11033.75 10822.02 11245.48  
## 2002.85 11165.20 10922.65 11407.74  
## 2003.10 11296.64 11022.61 11570.67  
## 2003.35 11428.09 11121.83 11734.34  
## 2003.60 11559.53 11220.27 11898.79  
## 2003.85 11690.97 11317.90 12064.04  
## 2004.10 11822.42 11414.73 12230.10  
## 2004.35 11953.86 11510.75 12396.97  
## 2004.60 12085.31 11605.96 12564.65  
## 2004.85 12216.75 11700.36 12733.14  
## 2005.10 12348.19 11793.97 12902.42  
## 2005.35 12479.64 11886.79 13072.48  
## 2005.60 12611.08 11978.84 13243.32  
## 2005.85 12742.52 12070.12 13414.93  
## 2006.10 12873.97 12160.65 13587.29  
## 2006.35 13005.41 12250.43 13760.40  
## 2006.60 13136.86 12339.47 13934.24  
## 2006.85 13268.30 12427.79 14108.81  
## 2007.10 13399.74 12515.39 14284.09  
## 2007.35 13531.19 12602.29 14460.08  
## 2007.60 13662.63 12688.50 14636.77  
## 2007.85 13794.08 12774.01 14814.14  
## 2008.10 13925.52 12858.85 14992.19  
## 2008.35 14056.96 12943.02 15170.90  
## 2008.60 14188.41 13026.54 15350.28  
## 2008.85 14319.85 13109.39 15530.31  
## 2009.10 14451.29 13191.61 15710.98  
## 2009.35 14582.74 13273.19 15892.29  
## 2009.60 14714.18 13354.14 16074.22  
## 2009.85 14845.63 13434.47 16256.78  
## 2010.10 14977.07 13514.19 16439.95  
## 2010.35 15108.51 13593.30 16623.73  
## 2010.60 15239.96 13671.80 16808.11  
## 2010.85 15371.40 13749.72 16993.09  
## 2011.10 15502.85 13827.04 17178.65

##plot

plot(mygdpforecast)

 ##Validate forecast, p-value is checked with diff lag values in Box.test When performing the Ljung-Box test (as implemented in Box.test), the null hypothesis is that there is no autocorrelation in the residuals of a time series model, up to a specified lag.

If the p-value from the Ljung-Box test is greater than the chosen level of significance (usually 0.05)

Box.test(mygdpforecast$resid, lag=5, type="Ljung-Box")

##   
## Box-Ljung test  
##   
## data: mygdpforecast$resid  
## X-squared = 4.1364, df = 5, p-value = 0.5299

Box.test(mygdpforecast$resid, lag=15, type="Ljung-Box")

##   
## Box-Ljung test  
##   
## data: mygdpforecast$resid  
## X-squared = 18.444, df = 15, p-value = 0.24

Box.test(mygdpforecast$resid, lag=30, type="Ljung-Box")

##   
## Box-Ljung test  
##   
## data: mygdpforecast$resid  
## X-squared = 36.94, df = 30, p-value = 0.1789

##Model performed well ##The END