TIME SERIES ANALYSIS OF U.S. ANNUAL STOCK PRICE DATA (1871-1970)

**INTRODUCTION**

***Problem Description:*** Here in this problem we are interested in,

Choosing a non seasonal time series data set and answer the following questions

1. Fit a suitable ARMA model for describing the patterns in the model and draw our conclusions.
2. Further we also want to perform the residual analysis to check if the model provides adequate information about the data.

***Objective:*** Here our main objective is to fit a suitable ARMA model for the give US Stock price dataset and further we want to perform the residual analysis to check if the model provides adequate information about the data.

***Moving average process:*** In time series analysis, the moving-average model, also known as moving-average process, is a common approach for modeling univariate time series. The moving-average model specifies that the output variable depends linearly on the current and various past values of a stochastic term.

*#Setting and getting the current working directory.*  
**setwd**("E:/M.Sc/SEM III/TIME\_SERIES\_ANALYSIS(MST371)/Practical Labs")  
**getwd**()

## [1] "E:/M.Sc/SEM III/TIME\_SERIES\_ANALYSIS(MST371)/Practical Labs"

***Data Description:*** Below is the Annual U.S. common stock price recorded in the period 1871 - 1970.

The data set consist of yearly record of U.S. common stock price from the year 1871 - 1970.

The dataset consist of 99 records and two columns i.e.

***Year***, denoted by ***t***

***Stock Price***, denoted by ***zt***

*#Loading the 'readxl' package required to load the dataset from excel.*  
**library**(readxl)  
  
*#Loading the US stock price dataset.*  
StockUK <- **read\_excel**("E:/M.Sc/SEM III/TIME\_SERIES\_ANALYSIS(MST371)/StockUK.xlsx")  
  
*#Obtaining the first few records of the dataset.*  
**head**(StockUK)

## # A tibble: 6 x 2  
## Year Stock\_Price  
## <dbl> <dbl>  
## 1 1871 5.03  
## 2 1872 4.8   
## 3 1873 4.57  
## 4 1874 4.45  
## 5 1875 4.06  
## 6 1876 3.14

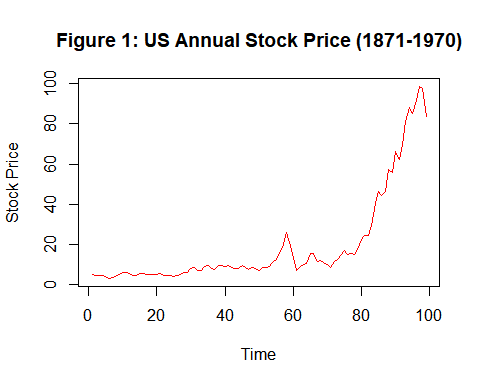
Thus, the stock price dataset is loaded in R.

**ANALYSIS**

*#Storing the stock price data in a seperate variable 'stockP'*  
stockP<-StockUK**$**Stock\_Price

Now, we proceed to perform the exploratory data analysis of the time series data to understand the behaviour of the data.

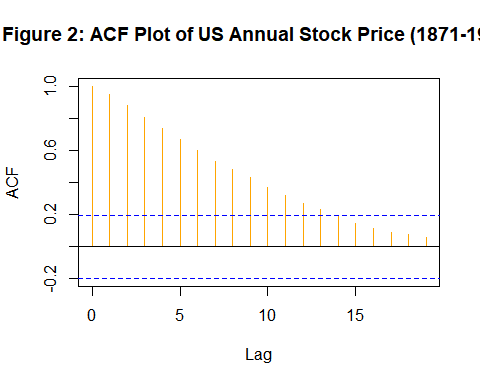
*#Obtaining timeseries plot for UK Stock price data.*   
**ts.plot**(stockP, main = "Figure 1: US Annual Stock Price (1871-1970)", xlab = "Time", ylab = "Stock Price", col = "red")



***Interpretation:*** Thus, we observe from the above time series plot (Figure 1) that there exist a trend component in the dataset.

Now, we proceed to examine the stationarity of the time series data using acf plot and augmented dickey feller test.

*#Obtaining the ACF plot of the above time series data.*  
**acf**(stockP, main = "Figure 2: ACF Plot of US Annual Stock Price (1871-1970)", col = "orange")



***Interpretation:*** From the above ACF plot (Figure 2) we observe that most of the lag values are significant thus we can conclude that the UK stock price time series data is not stationary.

*#loading the package 'tseries'*  
**library**(tseries)

## Warning: package 'tseries' was built under R version 4.0.5

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

*#Checking for the stationarity of the dataset.*  
**adf.test**(stockP)

##   
## Augmented Dickey-Fuller Test  
##   
## data: stockP  
## Dickey-Fuller = -1.1455, Lag order = 4, p-value = 0.9115  
## alternative hypothesis: stationary

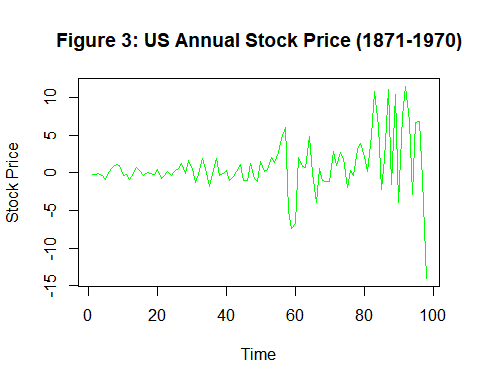
***Interpretation:*** Thus, at 95 % level of significance, from the statistical test, augmented dickey feller test we observe that the p value obtained for the dataset is 0.9115 which is greater than 0.05 thus we conclude the the above time series data is non-stationary.

Now, since the data is non-stationary we try to extract stationary component of the dataset before fitting a suitable model.

Since we only have a trend component in our dataset therefore we detrend the time series data using the method of differencing.

*#Detrending the dataset to extract the stationary component from the dataset.*  
data=**diff**(stockP)

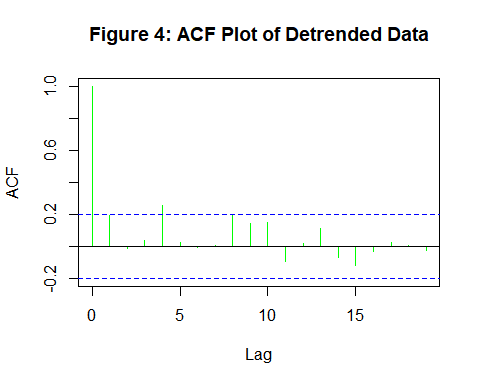
*#Obtaining the time series plot of detrended data.*  
**ts.plot**(data, main = "Figure 3: US Annual Stock Price (1871-1970)", xlab = "Time", ylab = "Stock Price", col = "green")



***Interpretation:*** Thus, from the time series plot (Figure 3) we observe that the trend component is removed and the data seems to be stationary now.

However, we crosscheck the stationarity of the dataset using the ACF plot and with the help of adf test.

*#Obtaining the ACF plot of the above detrended time series data.*  
**acf**(data, main = "Figure 4: ACF Plot of Detrended Data", col = "green")



*#loading the package 'tseries'*  
**library**(tseries)  
  
*#Checking for the stationarity of the dataset.*  
**adf.test**(data)

##   
## Augmented Dickey-Fuller Test  
##   
## data: data  
## Dickey-Fuller = -3.5453, Lag order = 4, p-value = 0.04179  
## alternative hypothesis: stationary

***Interpretation:*** Thus from the ACF plot in Figure 4 we observe that lag 4 is crossing the threshold line and also from the augmented dickey-fuller test we observe that the p value associated with the adf test is 0.04179 < 0.05 thus we reject the null hypothesis and conclude that the detrended dataset is stationary.

Now, since we extracted the stationary component of the time series data we proceed to fit the suitable model.

*#Loading the package 'forecast' required for fitting the suitable ARMA model.*  
**library**(forecast)

## Warning: package 'forecast' was built under R version 4.0.5

*#Fitting the suitable ARMA model using stationary data.*  
fit=**auto.arima**(data, seasonal=FALSE)  
**summary**(fit)

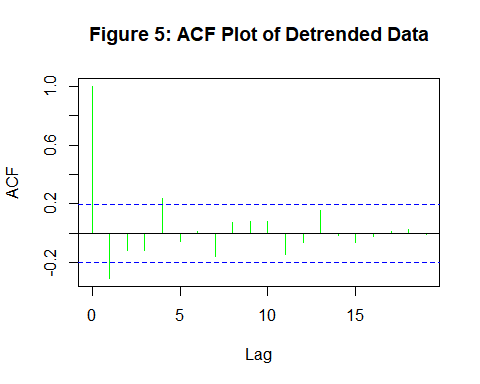
## Series: data   
## ARIMA(0,1,2)   
##   
## Coefficients:  
## ma1 ma2  
## -0.7464 -0.1924  
## s.e. 0.1103 0.1128  
##   
## sigma^2 estimated as 12.49: log likelihood=-260.09  
## AIC=526.19 AICc=526.45 BIC=533.91  
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.3000958 3.479911 2.152669 114.9318 135.0712 0.8116332  
## ACF1  
## Training set -0.02675725

***Interpretation:*** Thus, from the above summary it is observed that the model obtained using auto.arima command is MA(2) with 1 differencing with AIC = 526.19. Therefore, we difference the data once again to get a better form of a stationary data and try to fit the suitable model.

Now we again try to detrend the data using 2nd order differencing to extract the stationary component of the data.

*#Detrending the dataset again to extract the stationary component from the dataset.*  
data1=**diff**(data)

*#Obtaining the ACF plot of the above detrended time series data.*  
**acf**(data1, main = "Figure 5: ACF Plot of Detrended Data", col = "green")



*#loading the package 'tseries'*  
**library**(tseries)  
  
*#Checking for the stationarity of the dataset.*  
**adf.test**(data1)

## Warning in adf.test(data1): p-value smaller than printed p-value

##   
## Augmented Dickey-Fuller Test  
##   
## data: data1  
## Dickey-Fuller = -5.5647, Lag order = 4, p-value = 0.01  
## alternative hypothesis: stationary

***Interpretation:*** Thus from the ACF plot in Figure 5 we observe that lag 1 & 4 is crossing the threshold line whereas also from the augmented dickey-fuller test we observe that the p value associated with the adf test is less than 0.01 < 0.05 thus we reject the null hypothesis and conclude that the detrended dataset is stationary.

Now, since we extracted the best form of stationary component of the time series data we proceed to fit the suitable model.

*#Fitting the suitable ARMA model using stationary data.*  
fit1=**auto.arima**(data1, seasonal=FALSE)  
**summary**(fit1)

## Series: data1   
## ARIMA(0,0,2) with zero mean   
##   
## Coefficients:  
## ma1 ma2  
## -0.7464 -0.1924  
## s.e. 0.1103 0.1128  
##   
## sigma^2 estimated as 12.49: log likelihood=-260.09  
## AIC=526.19 AICc=526.45 BIC=533.91  
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE ACF1  
## Training set 0.303192 3.497803 2.174859 -27.94577 249.5656 0.531642 -0.02687996

***Interpretation:*** Thus, from the above summary it is observed that the model obtained using auto.arima command is MA(2) with AIC = 526.19. Thus, auto.arima gave the best suitable model for the data which is a MA(2) model.

**RESIDUAL ANALYSIS**

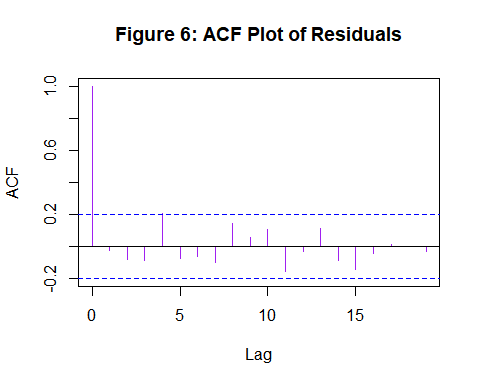
***ASSUMPTIONS***

1. Residuals are uncorrelated random variables.
2. Residuals are normally distributed random variables.

*#Obtaining the residuals.*  
res <- **resid**(fit1)

***Assumption 1: Residuals are uncorrelated random variables.***

*#Obtaining the acf plot for residuals.*  
**acf**(res, main = "Figure 6: ACF Plot of Residuals", col = "purple")



***Interpretation:*** Thus, from the acf plot in Figure 5 we observe that almost all the lags lies inside the threshold line therefore we can say that the residuals are uncorrelated random variables.

To confirm the same we perform statistical test.

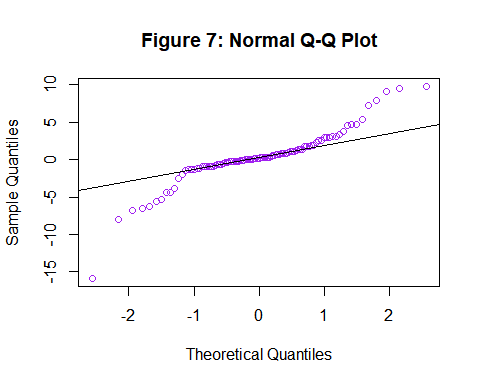
*#Computing Box-Pierce and Ljung-Box Tests to check if the residuals are uncorrelated.*  
**Box.test**(res, lag=10, fitdf = 2)

##   
## Box-Pierce test  
##   
## data: res  
## X-squared = 11.122, df = 8, p-value = 0.1949

***Interpretation:*** Since from the above test we observe that p value is 0.0.2783 which is greater than 0.05 thus we accept the null hypothesis and conclude that the residuals are uncorrelated random variables.

***Assumptions 2: Residuals are normally distributed random variables.***

*#Checking for normality with the help of a qq plot.*  
**qqnorm**(res, main = "Figure 7: Normal Q-Q Plot", col = "purple")  
**qqline**(res)



***Interpretation:*** From the Q-Q plot in figure 7 we observe that the points in the graph do not form a straight line thus we can say that the residuals are not normally distributed random variables.

*#Checking for normality with the help of shapiro wilks test.*  
**shapiro.test**(res)

##   
## Shapiro-Wilk normality test  
##   
## data: res  
## W = 0.87803, p-value = 2.111e-07

***Interpretation:*** Since from the above test we observe that p value is 2.111e-07 which is less than 0.05 thus we reject the null hypothesis and conclude that the residuals are not normally distributed random variables.

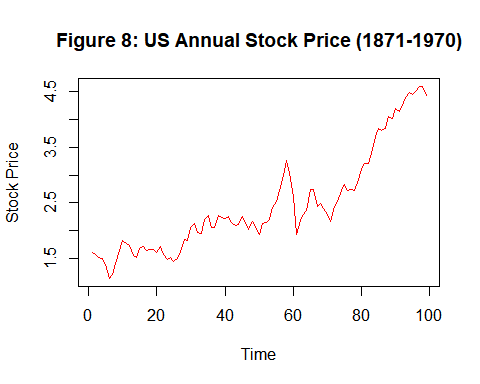
Since the residuals are not normally distributed now we need to perform transformation of the original data and perform the analysis again.

Transforming the data using log transformation.

*#Transforming the stock price data using log transformation.*  
log\_stockP<-**log**(stockP)

Now, we proceed to perform the exploratory data analysis of the transformed time series data to understand the behaviour of the data.

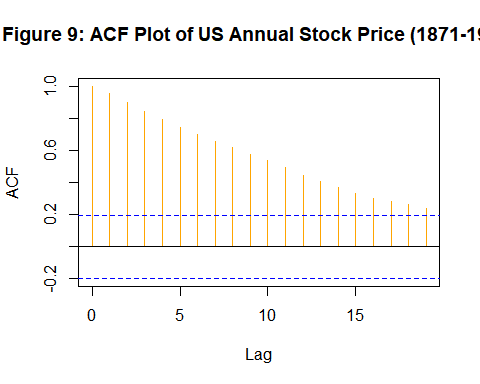
*#Obtaining timeseries plot for transformed UK Stock price data.*   
**ts.plot**(log\_stockP, main = "Figure 8: US Annual Stock Price (1871-1970)", xlab = "Time", ylab = "Stock Price", col = "red")



***Interpretation:*** Thus, we observe from the above time series plot (Figure 8) that there exist a trend component in the dataset.

Now, we proceed to examine the stationarity of the transformed time series data using acf plot and augmented dickey feller test.

*#Obtaining the ACF plot of the above transformed time series data.*  
**acf**(log\_stockP, main = "Figure 9: ACF Plot of US Annual Stock Price (1871-1970)", col = "orange")



***Interpretation:*** From the above ACF plot (Figure 9) we observe that most of the lag values are significant thus we can conclude that the transformed UK stock price time series data is not stationary.

*#Checking for the stationarity of the dataset.*  
**adf.test**(log\_stockP)

##   
## Augmented Dickey-Fuller Test  
##   
## data: log\_stockP  
## Dickey-Fuller = -1.5406, Lag order = 4, p-value = 0.7668  
## alternative hypothesis: stationary

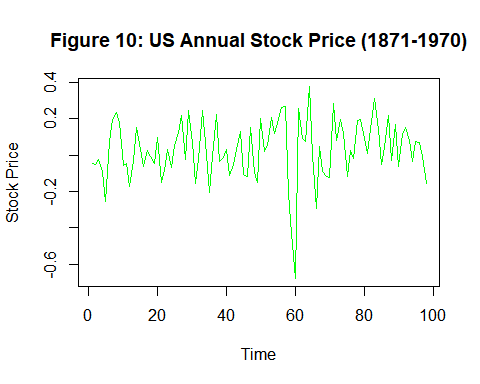
***Interpretation:*** Thus, at 95 % level of significance, from the statistical test, augmented dickey feller test we observe that the p value obtained for the dataset is 0.7668 which is greater than 0.05 thus we conclude the the above transformed time series data is non-stationary.

Now, since the data is non-stationary we try to extract stationary component of the dataset before fitting a suitable model.

Since we only have a trend component in our dataset therefore we detrend the time series data using the method of differencing.

*#Detrending the dataset to extract the stationary component from the dataset.*  
data2=**diff**(log\_stockP)

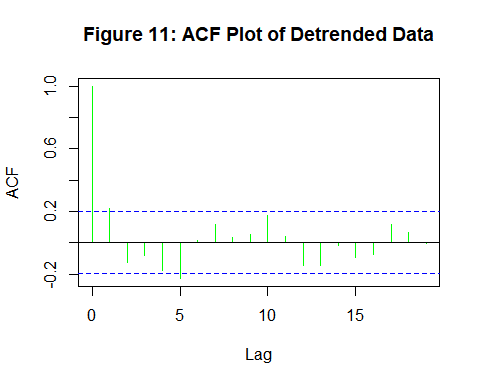
*#Obtaining the time series plot of detrended data.*  
**ts.plot**(data2, main = "Figure 10: US Annual Stock Price (1871-1970)", xlab = "Time", ylab = "Stock Price", col = "green")



***Interpretation:*** Thus, from the time series plot (Figure 10) we observe that the trend component is removed and the data seems to be stationary now.

However, we crosscheck the stationarity of the dataset using the ACF plot and with the help of adf test.

*#Obtaining the ACF plot of the above detrended time series data.*  
**acf**(data2, main = "Figure 11: ACF Plot of Detrended Data", col = "green")



*#Checking for the stationarity of the dataset.*  
**adf.test**(data2)

## Warning in adf.test(data2): p-value smaller than printed p-value

##   
## Augmented Dickey-Fuller Test  
##   
## data: data2  
## Dickey-Fuller = -6.3455, Lag order = 4, p-value = 0.01  
## alternative hypothesis: stationary

***Interpretation:*** Thus from the ACF plot in Figure 4 we observe that lag 1 and 5 is crossing the threshold line whereas from the augmented dickey-fuller test we observe that the p value associated with the adf test is 0.01 < 0.05 thus we reject the null hypothesis and conclude that the detrended dataset is stationary.

Now, since we extracted the stationary component of the time series data we proceed to fit the suitable model.

*#Fitting the suitable ARMA model using stationary data.*  
fit2=**auto.arima**(data2, seasonal=FALSE)  
**summary**(fit2)

## Series: data2   
## ARIMA(0,0,1) with zero mean   
##   
## Coefficients:  
## ma1  
## 0.3273  
## s.e. 0.1040  
##   
## sigma^2 estimated as 0.02482: log likelihood=42.5  
## AIC=-80.99 AICc=-80.87 BIC=-75.82  
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.02126968 0.1567407 0.1228001 115.5017 150.821 0.7926812  
## ACF1  
## Training set -0.04592277

***Interpretation:*** Thus, from the above summary it is observed that the model obtained using auto.arima command is MA(1) with AIC = -80.99 for the transformed data. Thus the best suitable model to the dataset is MA(1) model.

**RESIDUAL ANALYSIS**

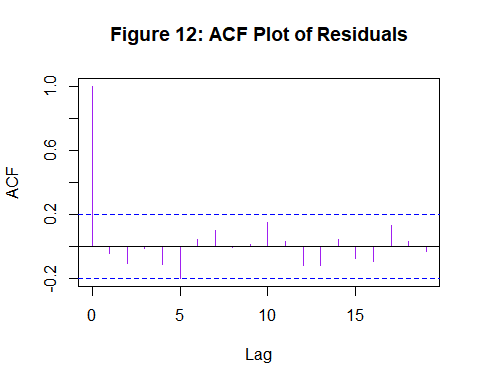
***ASSUMPTIONS***

1. Residuals are uncorrelated random variables.
2. Residuals are normally distributed random variables.

*#Obtaining the residuals.*  
res1 <- **resid**(fit2)

***Assumption 1: Residuals are uncorrelated random variables.***

*#Obtaining the acf plot for residuals.*  
**acf**(res1, main = "Figure 12: ACF Plot of Residuals", col = "purple")



***Interpretation:*** Thus, from the acf plot in Figure 12 we observe that almost all the lags lies inside the threshold line therefore we can say that the residuals are uncorrelated random variables.

To confirm the same we perform statistical test.

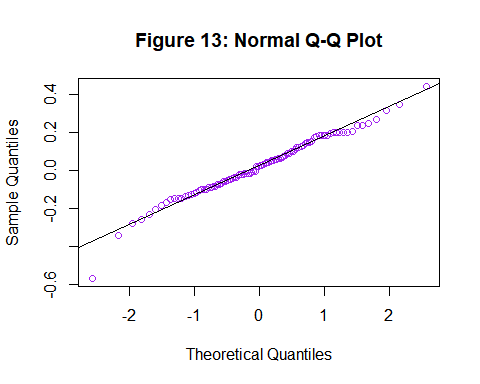
*#Computing Box-Pierce and Ljung-Box Tests to check if the residuals are uncorrelated.*  
**Box.test**(res1, lag=10, fitdf = 1)

##   
## Box-Pierce test  
##   
## data: res1  
## X-squared = 9.8148, df = 9, p-value = 0.3657

***Interpretation:*** Since from the above test we observe that p value is 0.3657 which is greater than 0.05 thus we accept the null hypothesis and conclude that the residuals are uncorrelated random variables.

***Assumptions 2: Residuals are normally distributed random variables.***

*#Checking for normality with the help of a qq plot.*  
**qqnorm**(res1, main = "Figure 13: Normal Q-Q Plot", col = "purple")  
**qqline**(res1)



***Interpretation:*** From the Q-Q plot in figure 13 we observe that the points in the graph almost form a straight line thus we can say that the residuals are normally distributed random variables.

*#Checking for normality with the help of shapiro wilks test.*  
**shapiro.test**(res1)

##   
## Shapiro-Wilk normality test  
##   
## data: res1  
## W = 0.98083, p-value = 0.1634

***Interpretation:*** Since from the above test we observe that p value is 0.1634 which is greater than 0.05 thus we accept the null hypothesis and conclude that the residuals are normally distributed random variables.

**CONCLUSION**

From the above analysis, we observed that the best fitted model to the above dataset is moving average model of order 1 MA (1) with AIC = -80.99 . On performing the residual analysis we observed that both the assumptions are made about the model is satisfied i.e. the residuals are uncorrelated and normally distributed random variables which is evident to the fact that the model provide adequate information about data.