**Time series analysis of stock market data**

**Name :** Saikat Ghosh

**Roll No :** 2403

**Registration No. :** A01-1112-0695-16

**Supervisor’s Name :** Prof. Madhura Dasgupta

**Introduction**

The stock or share or equity is a type of security that signifies a proportion of ownership in a corporation. The stock entitles stockholder to that proportion of corporation’s assets and earnings. The stock market is any exchange that allows people to buy and sell stocks and companies to issue stocks. A stock represents the company’s equity and shares the pieces of the company. The different components of a stock market data is recorded against time for the purpose of analyzing them. So, a great way of analyzing stock market data is by using the methods of time series analysis. Observations on a phenomenon which is moving through time generate an ordered set known as time series. A time series is said to be continuous when observations are made continuously over time and it is said to be discrete when observations are recorded at intervals, which usually are equally spaced.

**Data**

The given data is on Apple company stock prices for 10 years. The five variables in the data are open, low, high, close and volume. The prices are recorded weekly for every month of each year. The data is a sequence taken at successive equally spaced points in time. Thus it is a time series data of discrete-time and continuous-space.



**Objective**

The data exhibits non-stationary components such as seasonality, trend, cyclical, and irregular variations which make it difficult for us to forecast the future stock prices. The non-stationary components such as seasonality, cyclical and irregular variations are to be removed from every component of the data. The stationary data is to be analyzed in order to obtain meaningful statistics and other characteristics. The data is to be fitted into a mathematical model to generate forecasts. The irregular variations are to be explained in terms of probability models such as moving-average or auto-regressive models.

**Theory**

Ⅰ) The main components of a time series

a) Seasonality

The periodic movement which occurs within a year due to seasonal effects is known as seasonality.

b) Cyclical Variations

Cyclical movements are fluctuations which differ from seasonal movements, they are of longer duration than a year and also they do not ordinarily exhibit regular periodicity.

c) Trend

Trend is the definitive component of a time series, in layman terms it may be defined as ‘long term change in the mean level’. Over a long period, a time series is very likely to show a definite tendency which may or may be constant in direction and rate, this is known as trend.

d) Irregular Variations

After removing the other components of a time series the residuals are known as irregular variations which may or may not be random in nature.

A time series which exhibits the above components is termed as non-stationary time series. Probability theory of time series is concerned with stationary time series, so the non-stationary components are removed from the time series. A time series is said to be stationary if there is no systematic change in mean, variance and if strictly periodic variations have been removed.

The decomposition of time series is represented either in an additive or in multiplicative form based on the given data. Looking at our time series plot if we see that the seasonal variations increase with increase in trend, i.e. it increases over time then we fit a multiplicative model.

a) Multiplicative Model

Xt = TtStItCt

b) Additive Model

Xt = Tt+St+It+Ct

Where,

Xt is the observed time series.

Tt is the trend component.

St is the seasonality component.

It is the irregular component.

Ct is the cyclical component.

Ⅱ) Some descriptive methods used to analyze the data

a) Method of moving averages

The method of differencing is a simple method which cannot be used when the series contains both seasonality and trend, in that case the method of moving averages is used. The moving average of period k of a time series gives us a series of arithmetic means, each of k consecutive observations. We start with first k observations and this process is carried on until we reach the last k observations. Each of these means is centered against the time which is the mid-point of the time interval included in the calculation of moving average. This method is used to get an estimate of the trend.

b) Method of mathematical curve fitting

The movement of trend is considered to be fairly smooth over long periods of time. This means we can represent the trend component by a polynomial in time element t. This method of fitting a mathematical curve to the deterministic part of the time series allows us to forecast future values.

c) Autocorrelation

Given N observations x1, x2,....., xN on a discrete time series we can form N-1 pairs of observations. Regarding the first observation in each pair as one variable and the second observation as a second variable, the correlation coefficient between xt and xt+1 is given by,

r1 = N-1Ʃt=1[(xt-x(1))(xt-x(2))] / √[{N-1Ʃt=1(xt- x(1))2}{N-1Ʃt=1(xt- x(2))2}]

Where,

r1 is known as sample autocorrelation coefficient at lag 1.

x(1) is the mean of first N-1 observations.

x(2) is the mean of last N-1 observations.

Since, x(1) ≈ x(2), the above formula can be written as,

r1 = N-1Ʃt=1[(xt-x\_bar)(xt+1-x\_bar)] / NƩt=1(xt- x\_bar)2

Similarly the sample autocorrelation coefficient at lag k is given by,

rk = N-kƩt=1[(xt-x\_bar)(xt+k-x\_bar)] / NƩt=1(xt- x\_bar)2

Where,

x\_bar is the overall mean.

A set of autocorrelation coefficients is interpreted using a graph called correlogram which is rk plotted against k.

Ⅲ) Probability models for time series

A statistical phenomenon that evolves in time according to probability laws is called a stochastic process. A stochastic process is said to be strictly stationary if its properties are unaffected by a change of time origin. This why descriptive methods are used to make a time-series stationary. The residual series obtained is explained in terms of probability models.

a) Moving average process

Suppose {Zt} is a purely random process with mean zero and variance σz2. Then a process {Xt} is said to be a moving average process of order q if,

Xt = β0Zt + β1Zt-1 + .... + βqZt-q

Where {βi} are constants.

E(Xt) = 0

V(Xt) = σz2 qƩi=0 βi2

Since, Z’s are independent we have,

γ(k) = Cov(Xt, Xt+k)

= Cov(β0Zt + β1Zt-1 + .... + βqZt-q)

= 0 if k > q

σz2 q-kƩi=0 βiβi+k  if k = 0,1,......,q

γ(-k) if k < 0

The autocorrelation function of a MA(q) process is given by,

ρ(k) = 1 if k = 0

q-kƩi=0βiβi+k / qƩi=0 βi2 if k = 1,....,q

0 if k > q

ρ(-k) if k < 0

b) Autoregressive process

Suppose {Zt} is a purely random process with mean zero and variance σz2. Then a process {Xt} is said to be autoregressive of order p if,

Xt = α1Xt-1 + .... + αpXt-p + Zt

E(Xt) = 0

Var(Xt) = σz2 / (1-α2) [Variance is finite provided |α|<1 ]

γ(k) = Cov(Xt,Xt+k)

= σz2 ꝏƩi=0 αiαk+i for k>=0

= αkσx2

The autocorrelation function of an AR(p) process is given by,

ρ(k) = 1 if k = 0

= α|k| if k = +1,-1,+2,-2,.....,+p,-p

= 0 if k > p

Ⅳ) Forecasting

Single Exponential Smoothing

This forecasting procedure should only be used in its basic form for non-seasonal time series showing no systematic trend.

Given a non-seasonal time series with no systematic trend x1,x2,......,xN, the weighted sum of past observations can be taken as a estimate of xN+1.

x(N+1,1)’ = c0xN + c1xN-1 + ......

Where {ci} are weights.

We take ci = α(1-α)i , i=0,1,2,......

Where α is constant such that 0<α<1.

Therefore,

x(N+1,1)’ = αxN + α(1-α)xN-1 + α(1-α)2xN-1 +......

= αxN + (1-α) x(N-1,1)’

The accuracy of the forecasting method used is determined by some measures,

a) MAPE

This is known as mean absolute percentage error also known as mean absolute percentage deviation.

MAPE = 1/n nƩt=1|(At-Ft)/At|

b) MAD

This is known as mean absolute deviation.

MAD = 1/n nƩt=1|At-Ft|

c) MSD

This is known as mean squared deviation.

MSD = 1/n nƩt=1(At-Ft)2

Where,

At is the actual value.

Ft is the forecasted value.

t is the time over which values are recorded.

n is the total number of observations.

**Methodology**

A time series is analyzed for different purposes like for understanding the data better, finding meaningful statistical results and generating future results which is known as forecasting. The main aim of analyzing stock market data is to forecast future values. Which brings us to the first objective of analyzing the data: understanding the data. For example, if we see the time series plot for the variable Volume it is evident that a large number of stocks were purchased in the year 2011 and if we look at the plot for the price variables we can see that the stock prices in that year were very low; which is again the reason behind such a large number stocks being purchased. Such irregular variations are known as cyclical variations. Also if we observe the time series plot for all five variables it is observed that in general the seasonal variations increase with increase in trend, therefore we will fit a multiplicative model to the data.

Now for every variable we will follow the given steps,

Xt = TtStItCt

Where,

Xt is the observed time series.

Tt is the trend component.

St is the seasonality component.

Ct is the cyclical component.

It is the irregular component.

t is the time over which the series is observed.

Now we will use 12-point moving average method to remove the seasonality component. Since this data contains trend along with seasonality therefore we cannot use differencing method.

Sm(xt) = ( 1/2xt-6 + xt-5 + xt-4 + .... + xt+5 + 1/2xt+6 ) / 12

Clearly Sm(xt) is an estimate of trend.

We can get the detrended data as,

Yt = Xt / Sm(xt) = StIt [Since, Sm(xt)=Tt\*]

Now we will calculate the G.M of the monthly data for all the given years.

Hence, we will get 12 estimates of St. (say Si\*,i=1(1)12)

Now we calculate the adjustment factor,

S = G.M of Si\* , i=1(1)12

Then we divide the seasonal estimate by adjustment factor and get the seasonal index,

Si\*\* = Si\*/S , i=1(1)12

Now, the deseasonalised data is,

Zt = Xt/Si\*\* = TtIt

This deseasonalised data only has trend and we try to estimate the trend by fitting a mathematical curve. Looking at the time series plot for all five variables we decide to fit a quadratic trend curve to each of them.

We also generate some forecasts while analyzing the trend based on the mathematical model.

We will calculate the ACF and PACF using the data which has been detrended and deseasonalised.

We will generate a correlogram using the ACFs.

We will fit an auto-regressive model and a moving-average model whose parameters will be estimated based on the correlogram.

We will use single exponential smoothing method for the purpose of forecasting.

**Results**

Ⅰ) The trend curves

a) Open

Tt1 = -264 + 63.03t – 0.4956t2

MAPE = 57

MAD = 452

MSD = 300561

Monthly forecasts generated,

|  |  |
| --- | --- |
| Month | Forecast |
| Jan | 1021.48 |
| Feb | 1023.26 |
| Mar | 1025.04 |
| Apr | 1026.83 |
| May | 1028.62 |
| Jun | 1030.41 |
| Jul | 1032.21 |
| Aug | 1034.01 |
| Sep | 1035.81 |
| Oct | 1037.61 |
| Nov | 1039.42 |
| Dec | 1041.23 |
|  |  |

b) High

Tt2 = -252 + 64.14t – 0.5035t2

MAPE = 57

MAD = 466

MSD = 318897

Monthly forecasts generated,

|  |  |
| --- | --- |
| Month | Forecast |
| Jan | 1071.91 |
| Feb | 1073.937 |
| Mar | 1075.969 |
| Apr | 1078.005 |
| May | 1080.044 |
| Jun | 1082.087 |
| Jul | 1084.134 |
| Aug | 1086.185 |
| Sep | 1088.24 |
| Oct | 1090.299 |
| Nov | 1092.361 |
| Dec | 1094.428 |
|  |  |

c) Low

Tt3 = -285 + 62.79t – 0.4958t2

MAPE = 58

MAD = 443

MSD = 289168

Monthly forecasts generated,

|  |  |
| --- | --- |
| Month | Forecast |
| Jan | 947.2864 |
| Feb | 948.4603 |
| Mar | 949.6357 |
| Apr | 950.8125 |
| May | 951.9908 |
| Jun | 953.1706 |
| Jul | 954.3518 |
| Aug | 955.5344 |
| Sep | 956.7186 |
| Oct | 957.9042 |
| Nov | 959.0913 |
| Dec | 960.2798 |

d) Close

Tt4 = -276 + 64.13t – 0.5075t2

MAPE = 57

MAD = 454

MSD = 302533

Monthly forecasts generated,

|  |  |
| --- | --- |
| Month | Forecast |
| Jan | 965.6082 |
| Feb | 966.5837 |
| Mar | 967.5603 |
| Apr | 968.5378 |
| May | 969.5163 |
| Jun | 970.4958 |
| Jul | 971.4763 |
| Aug | 972.4578 |
| Sep | 973.4403 |
| Oct | 974.4237 |
| Nov | 975.4082 |
| Dec | 976.3936 |

e) Volume

Tt5 = 932217965 – 18901592t + 170574t2

MAPE = 38.7

MAD = 1.94 x 108

MSD = 6.49 x 1016

Monthly forecasts generated,

|  |  |
| --- | --- |
| Month | Forecast |
| Jan | 5.96E+08 |
| Feb | 5.97E+08 |
| Mar | 5.98E+08 |
| Apr | 5.99E+08 |
| May | 6E+08 |
| Jun | 6.01E+08 |
| Jul | 6.02E+08 |
| Aug | 6.03E+08 |
| Sep | 6.04E+08 |
| Oct | 6.05E+08 |
| Nov | 6.06E+08 |
| Dec | 6.07E+08 |

Where,

MAPE is the mean absolute percentage error

MAD is the mean absolute deviation

MSD is the mean square deviation

Based on the above data we will fit an MA(2) and AR(2) process to the residual series of each variable.

Ⅱ) The ACF and PACF values

a) Open

Correlogram for variable Open which has been made stationary



|  |  |
| --- | --- |
| ACF | PACF |
| 0.895201 | 0.895201 |
| 0.864665 | 0.318605 |
| 0.834616 | 0.123101 |
| 0.778697 | -0.10488 |
| 0.739381 | -0.02299 |
| 0.690164 | -0.06103 |
| 0.622987 | -0.14533 |
| 0.594234 | 0.087383 |
| 0.566166 | 0.114789 |
| 0.501359 | -0.14436 |
| 0.495087 | 0.148198 |
| 0.489642 | 0.178148 |
| 0.437732 | -0.18885 |
| 0.445284 | 0.092919 |
| 0.393337 | -0.19483 |
| 0.362868 | -0.03842 |
| 0.354301 | 0.030696 |
| 0.309269 | -0.06337 |
| 0.257018 | -0.0948 |
| 0.227091 | -0.06167 |
| 0.163352 | -0.11546 |
| 0.0937 | -0.14724 |
| 0.053836 | -0.05578 |
| -0.03218 | -0.1368 |
| -0.07324 | 0.05661 |
| -0.11538 | -0.06529 |
| -0.19032 | -0.04169 |
| -0.22258 | -0.01529 |
| -0.25335 | -0.00749 |
| -0.30095 | -0.06478 |

b) High

Correlogram for variable High which has been made stationary



|  |  |
| --- | --- |
| ACF | PACF |
| 0.893698 | 0.893698 |
| 0.860783 | 0.308425 |
| 0.832484 | 0.136318 |
| 0.776256 | -0.10102 |
| 0.738784 | -0.00939 |
| 0.689903 | -0.06446 |
| 0.62161 | -0.1523 |
| 0.59344 | 0.084909 |
| 0.566605 | 0.114339 |
| 0.499417 | -0.1549 |
| 0.493107 | 0.150937 |
| 0.486735 | 0.168957 |
| 0.434302 | -0.17965 |
| 0.441524 | 0.096642 |
| 0.388629 | -0.19622 |
| 0.360488 | -0.00554 |
| 0.350429 | -0.00131 |
| 0.303897 | -0.07599 |
| 0.252255 | -0.08047 |
| 0.221736 | -0.06971 |
| 0.160159 | -0.09966 |
| 0.090323 | -0.14401 |
| 0.053785 | -0.03534 |
| -0.03473 | -0.15972 |
| -0.07525 | 0.047032 |
| -0.11471 | -0.04624 |
| -0.1913 | -0.05602 |
| -0.22022 | -0.00614 |
| -0.25534 | -0.02573 |
| -0.30096 | -0.04994 |

c) Low

Correlogram for variable Low which has been made stationary



|  |  |
| --- | --- |
| ACF | PACF |
| 0.893033 | 0.893033 |
| 0.863185 | 0.324342 |
| 0.832704 | 0.122651 |
| 0.775193 | -0.11061 |
| 0.737405 | -0.01379 |
| 0.689436 | -0.05029 |
| 0.620722 | -0.15328 |
| 0.590366 | 0.07028 |
| 0.565507 | 0.135104 |
| 0.498802 | -0.15063 |
| 0.493536 | 0.145459 |
| 0.487356 | 0.174483 |
| 0.436708 | -0.17079 |
| 0.44718 | 0.097455 |
| 0.395949 | -0.19231 |
| 0.367174 | -0.02587 |
| 0.35992 | 0.021973 |
| 0.318005 | -0.06407 |
| 0.26744 | -0.08489 |
| 0.235942 | -0.08722 |
| 0.170269 | -0.13381 |
| 0.09812 | -0.15175 |
| 0.059578 | -0.05072 |
| -0.03136 | -0.14598 |
| -0.07301 | 0.061138 |
| -0.11303 | -0.03136 |
| -0.18877 | -0.05249 |
| -0.21941 | -0.00928 |
| -0.25537 | -0.04301 |
| -0.30301 | -0.07345 |

d) Close

Correlogram for variable Close which has been made stationary



|  |  |
| --- | --- |
| ACF | PACF |
| 0.890719 | 0.890719 |
| 0.859462 | 0.319822 |
| 0.832007 | 0.142505 |
| 0.771824 | -0.12068 |
| 0.737083 | 0.00322 |
| 0.691097 | -0.04431 |
| 0.620315 | -0.16316 |
| 0.592099 | 0.070428 |
| 0.56658 | 0.122732 |
| 0.497669 | -0.15631 |
| 0.493719 | 0.149122 |
| 0.487129 | 0.172321 |
| 0.434304 | -0.16602 |
| 0.445135 | 0.094027 |
| 0.393168 | -0.18755 |
| 0.363578 | -0.01176 |
| 0.357597 | 0.005433 |
| 0.314514 | -0.05731 |
| 0.263975 | -0.08754 |
| 0.235078 | -0.08215 |
| 0.169187 | -0.12578 |
| 0.098103 | -0.15035 |
| 0.059656 | -0.06091 |
| -0.0317 | -0.14986 |
| -0.07201 | 0.063577 |
| -0.11209 | -0.04241 |
| -0.18968 | -0.04994 |
| -0.22081 | -0.0185 |
| -0.25725 | -0.0359 |
| -0.30637 | -0.06975 |

e) Volume

Correlogram for variable Volume which has been made stationary



|  |  |
| --- | --- |
| ACF | PACF |
| 0.707148 | 0.707148 |
| 0.642485 | 0.284886 |
| 0.632632 | 0.231355 |
| 0.536945 | -0.02931 |
| 0.545533 | 0.13865 |
| 0.453041 | -0.10842 |
| 0.465 | 0.135161 |
| 0.440824 | -0.00673 |
| 0.366569 | -0.04241 |
| 0.278063 | -0.2169 |
| 0.319151 | 0.202308 |
| 0.280316 | -0.06159 |
| 0.208578 | -0.0337 |
| 0.186637 | -0.10125 |
| 0.129851 | -0.00306 |
| 0.129329 | -0.03273 |
| 0.01416 | -0.14774 |
| 0.009828 | 0.044762 |
| 0.01924 | 0.009323 |
| -0.01948 | 0.00939 |
| 0.001844 | 0.091294 |
| -0.07251 | -0.10133 |
| -0.06822 | -0.03164 |
| -0.08228 | -0.00313 |
| -0.12816 | -0.00672 |
| -0.11495 | 0.009423 |
| -0.09975 | 0.028057 |
| -0.10653 | 0.018157 |
| -0.10992 | 0.025542 |
| -0.12164 | -0.05173 |

Where ACF is the autocorrelation function and PACF is the partial autocorrelation function.

Ⅲ) Fitting an AR(2) process

This process is also known as Yule process and is defined as,

Xt = α1Xt-1 + α2Xt-2 + εt

Where {εt} is a purely random process and α1, α2 are constants.

Cov(Xt,Xt-k) = α1E[Xt-k­Xt-1] + α2E[Xt-k­Xt-2] + E[Xt-k­ εt]

γ(k) = α1γ(k-1) + α2γ(k-2)

Dividing both sides by γ(0),

ρ(k) = α1ρ(k-1) + α2ρ(k-2) ,k>0

The general solution of the equation is,

ρ(k) = A1U1k + A2U2k

Where, A1,A2 are found from the initial conditions and U1,U2 are the roots of the characteristic equation of the process, i.e. U2- α1U+ α2=0.

Case I : (α12+ 4α2) >= 0

In this case the roots are real.

Therefore,

ρ(k) = [U1(1-U22)U1k – U2(1-U12)U2k] / [(U1-U2)(1+U1U2)]

This consists of a mixture of damped exponentials.

Case II : (α12+ 4α2) < 0

In this case the roots are imaginary.

Let U1 = p(cosθ+isinθ), U2 = p(cosθ-isinθ), where 0<p<1

Therefore,

ρ(k) = pk{sin(θk+Ψ) / sinΨ} ,where (1-p2)cotθ/ (1+p2) = cotΨ

This consists of damped sine wave.

Here, ρ(k) is the ACF of AR(2) process.

Ⅳ) Fitting an MA(2) process

Xt = μ + Zt + β1Zt-1 ,where μ, β1 are constants and Zt denotes a purely random process.

ACF of MA(2) process,

ρ(k) = 0 , k=0

= (β0.β1 + β1.β2) / (β02 + β12 + β22) , k=1

= β0.β2 / (β02 + β12 + β22) , k=2

= 0 , k>2

Also, ρ(k) = ρ(-k).

Ⅴ) Forecasts using single exponential smoothing for variable

a) Open = 558.034

b) High = 654.391

c) Low = 514.463

d) Close = 512.281

e) Volume = 768206689

**Conclusion**

Investing in stock market is largely practiced by people to grow their wealth, outrun inflation, to make diversified investments and for many more reasons. So predicting the market plays a very important role in stock exchange. The more accurately a future stock price is predicted, the lesser is the risk subject to investing in it. Time series analysis is a very scientific and reliable method of predicting the market. Studying and applying time series analysis correctly can be a key to make better investments in the market. The results obtained from analyzing the time series data can help someone make an informed decision and it also presents a clearer picture of whether or not it is wise to buy or sell the share at that particular time.

**Acknowledgement**

I would hereby like to thank my supervisor Prof. Madhura Dasgupta for her expert guidance and constant supervision as well as for providing necessary information regarding the project and also for her dedicated support in completing it. I would like to express my gratitude towards my parents and professors of the Department of Statistics of Saint Xavier’s College, Kolkata for their kind co-operation and encouragement in pursuing the topic. I am highly indebted and would like to extend my sincere gratitude to every teaching and non-teaching staff of Saint Xavier’s College, Kolkata.

**Bibliography**

Books,

The Analysis of Time Series, Christopher Chatfield.

Time Series Analysis and its Applications, Robert H. Shumway.

Article Links,

<https://towardsdatascience.com/stock-market-forecasting-using-time-series-c3d21f2dd37f>

<https://www.researchgate.net/publication/286890497_Time_Series_Data_Analysis_for_Stock_Market_Prediction_using_Data_Mining_Techniques_with_R>