CHAPTER 1

INTRODUCTION

1.1 Time Series

A time series is a set of observations measured sequentially through time. This measurement may be made continuously through time or be taken at a discrete set of time points. A time series depicts the relationship between two variables one of the being times. Mathematically, a time series is defined by the functional relationship $y_t = f(t)$, where y_t is the value of the variable under consideration at time t. Here time may be yearly, monthly, weekly, daily or even hourly usually at equal intervals of time. The adjacent observations of a time series data are dependant and the analysis is concerned with the techniques for analysis of this dependence. Rainfall on successive days, monthly birth total in Kerala etc. are examples of time series.

Analysis of time series is of great importance not only to economist or businessman, but also the people working in various disciplines in natural, social and physical sciences. The following points indicate the utility of time series analysis;

- It enables us to study the past behaviour of the phenomenon under consideration i.e., to determine the type and nature of variations in the data.
- It helps in forecasting
- It helps in evaluation of current achievement.
- It helps in making comparative study.

Time series are used in statistics, signal processing pattern recognition, econometrics, mathematical finance, weather forecasting, earth quake prediction, electroencephalography, control engineering, astronomy, communication engineering, largely in any domain of applied science and engineering which involves temporal measurements.

1.2 Objectives of Time Series Analysis

There are several possible objectives in analysing a time series. This objective may be classified as description, modelling, forecasting and control.

a) Description: To describe the data using summary statistics and/or graphical method.
b) Modelling: To find suitable mathematical model to describe the data generating process.
c) Forecasting: To estimate the future values of the series. Given an observed time series one may want to predict the future values of the series.
d) Control: Good forecast enable the analyst to take actions so us to control a given process, whether it is an industrial process or an economy or whatever.

The chief objective of analysing a time series is to evaluate the past performance or occurrence of particular variables and to forecast or predict the magnitude of a variable in future so us to arrive at a desired conclusion for one's future course of action.

1.3 Components of Time Series

The various forces at work, affecting the values of a phenomenon in a time series can be broadly classified into the following categories, commonly known as components of time series.

- Secular trend
- Seasonal variation
- Cyclic variation
- Random or irregular movements

1. Secular trend: By trend we mean general tendency of the data to increase or decrease during the long period of time. A time series data may show upward trend or downward trend for a period of years and this may be due to factors like increase in population, change in technological progress, large scale shifting consumer's demands etc. it is not necessary that the decrease or decline should be in the same direction. It may possible that different tendencies of increase, decrease or stability are observed in different sections of time. However, over all tendency may be upward, downward or stable.
2. Seasonal variation: Seasonal variations are short term fluctuations in time series which occur periodically in a year. This continuous to repeat year after year. Thus, seasonal variation in time series will be there if the data are recorded quarterly, monthly, weekly, daily, hourly and so on. The major factors that are responsible for the repetitive pattern of seasonal variations are weather conditions and customs of

people. For example, more woollen clothes are sold in winter than in the season of summer, the sales in departmental stores are more during festive season than in normal days etc.

3. Cyclic variations: Cyclical variations are recurrent upward or downward movements in a time series but the period of cycle is greater than a year. Also, these variations are not regular as seasonal variations. It moves like pendulum of a clock and is never ending process.
4. Irregular variations: Irregular variations are fluctuation in time series that are short in duration, erratic in nature and follow no regularity in occurrence pattern. These variations are also referred to as residual variations since by definitions they represent what is left out in a time series after trend, cyclic and seasonal variations have been removed from the data. Irregular fluctuations result due to occurrence of unforeseen events like floods, earth quakes, wars etc.

1.4Mathematical Models for Time Series

In time series analysis it is assumed that there exist two models commonly for the decomposition of a time series into its components.

a) Additive Model: According to additive model the decomposition of time series is done on the assumption that the effects of various components are additive in nature or in other words,

$$Yt = Tt + St + Ct + It$$

Where Yt is the time series value and Tt, St, Ct and It stands for trend, seasonal variations, cyclic variations and irregular variations respectively. In this model Tt, St, Ct and It are absolute quantities and can have positive or negative values. The model assumes that the four components of the time series are independent of each other and none has any effect on the remaining three components. In actual practise this hypothesis is does not hold good as these factors affect each other.

b) Multiplicative Mode: According to the multiplicative model the decomposition of the time series is done on the assumption that the effects of four components are independent of each other. According to the multiplicative model,

$$Yt = Tt \times St \times Ct \times I$$

In this model Tt, St, Ct and It are not absolute amounts as in the case of additive model. They are relative variations and are expressed as rate or indices fluctuating above or below unity. The multiplicative model can be expressed in terms of the logarithms. Thus, if we take log of the multiplicative model, we get log Yt = log Tt + log St + log Ct + log It. It means that the multiplicative model is additive model if we take into account the log of the given time series values.

1.5 Stock price prediction

The stock market is known for being volatile, dynamic, and nonlinear. Accurate stock price prediction is extremely challenging because of multiple (macro and micro) factors, such as politics, global economic conditions, unexpected events, a company's financial performance, and so on. The entire idea of predicting stock prices is to gain significant profits. The art of forecasting stock prices has been a difficult task, in fact investors are highly interested in the research area of stock price prediction. For a good and successful investment, many investors are keen on knowing the future situation of the stock market. Good effective prediction systems for the stock market help traders, investors, and analyst by providing supportive information like the future direction of the stock market. Despite the volatility, stock prices aren't just randomly generated numbers. So, they can be analyzed as a sequence of discrete time data; in other words, time series observations taken at successive points in time (usually on a daily basis). Time series forecasting (predicting future values based on historical values) applies well to stock forecasting.

1.6 Present study

Finance professionals use forecasts to make financial plans. Investors invest their hard-earned capital in stocks with the expectation of gaining from their investment through a positive payoff. Since having an excellent knowledge about share price movement in the future serves the significance of fiscal professionals and investors. This familiarity about the future boosts their confidence by way of consulting and investing. But these

movements predict the share prices without proper forecasting methods, only for the interest of the financial professional and investors. There are many forecasting methods in projecting price movement of stocks such as the Box Jenkins method, Black-Scholes model, and Binomial model. The concept of forecasting stock market return has become fairly popular may be because of the fact that if the future market value of the stocks is successfully predicted, the investors may better guided. The profitability of investing and trading in stock market to a large extent depends on the predictability of the system which in turn prepares the investors in their encounter with their future insecurities and the risks associated with the market.

In this work, I represent my findings and experiments for stock price of daily openings of Netflix. We consider the modelling of Netflix stock price using Autoregressive Integrated Moving Average (ARIMA) and ARCH-GARCH time series model and forecast using the best among these two models. This data contains data for 5 years i.e., from 5th February 2018 to 5th February 2022. The data is collected from the website Kaggle www.kaggle.com

In chapter 2, we consider the basic tools and methodology for modelling and forecasting a time series data using autoregressive integrated moving average (ARIMA) model and ARCH-GARCH model.

Chapter 3 discusses the analysis of the time series data of daily stock price of Netflix using Box and Jenkins ARIMA modelling and Chapter 4 discusses ARCH-GARCH modelling procedure.

Chapter 5 discusses the forecasting procedures using the best model.The process of model fitting and forecasting for stock price prediction was done by the statistical packages SPSS and R.

CHAPTER 2

Time Series Modelling

2.1 Time Series Analysis

The phrase “Time Series Analysis” is used in several ways. Sometimes it refers to any kind of analysis involving time series data. Other times it is used more narrowly to describe attempts to explain behaviour of time series data using only past observations on the variable under consideration. The primary objective of time series analysis is to develop mathematical models that provide plausible description of sample data. In order to provide a statistical setting for describing the character of data that seemingly fluctuate in a random fashion over time, we assume a time series can be defined as a collection of random variables indexed according to the order they are obtained in time.

The theory and practice of time series analysis and forecasting has developed rapidly over the last several years. One of the better-known short-term forecasting methods is often referred to as univariate Box-Jenkins analysis or ARIMA analysis. The acronym ARIMA stands for ‘autoregressive integrated moving. The Autoregressive Conditional heteroskedastic (ARCH-GARCH) modelling is also used for analysing the data and the following sections will introduce the basic ideas of this methodology.

2.2 Normality Test

In statistics, normality tests are used to determine if a data set is well-modelled by a normal distribution and to compute how likely it is for a random variable underlying the data set to be normally distributed.

More precisely, the tests are a form of model selection, and can be interpreted several ways, depending on one's interpretations of probability:

- In descriptive statistics terms, one measures a goodness of fit of a normal model to the data – if the fit is poor then the data are not well modelled in that respect by a normal distribution, without making a judgment on any underlying variable.
- In frequentist statistics statistical hypothesis testing, data are tested against the null hypothesis that it is normally distributed.
- In Bayesian statistics, one does not "test normality" per se, but rather computes the likelihood that the data come from a normal distribution with given parameters μ, σ (for all μ, σ), and compares with the likelihood that the data come from other distributions under consideration, most simply using a Bayes factor (giving the relative likelihood of seeing the data given different models), or more finely taking a prior distribution on possible models and parameters and computing a posterior distribution given the computed likelihoods.

A normality test is used to determine whether sample data has been drawn from a normally distributed population (within some tolerance). A number of statistical tests, such as the student's t-test and the one-way and two-way ANOVA, require a normally distributed sample population.

2.2.1 Histogram

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.

2.2.2 P-P Plot

The Probability-Probability (PP) plot plots the cumulative probability of a variable against the cumulative probability of a particular distribution. After data are ranked and sorted, the corresponding z-scores are calculated for each rank. The scores are then themselves converted to z-scores. The actual z-scores are plotted against the expected z-scores. If the data are normally distributed, the result would be a straight diagonal line that is the observed value should fall on the expected normal distribution Line.

2.2.3 Q-Q Plot

In statistics, Q-Q(quantile-quantile) plots play a very vital role to graphically analyse and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight-line y = x. We plot the theoretical quantiles or basically known as the standard normal variate (a normal distribution with mean=0 and standard deviation=1) on the x-axis and the ordered values for the random variable which we want to find whether it is Gaussian distributed or not, on the y-axis. Which gives a very beautiful and a smooth straight line like structure from each point plotted on the graph. Now, we have to focus on the ends of the straight line. If the points at the ends of the curve formed from the points are not falling on a straight line but indeed are scattered significantly from the positions then we cannot conclude a relationship between the x and y axes which clearly signifies that our ordered values which we wanted to calculate are not Normally distributed.

If all the points plotted on the graph perfectly lies on a straight line, then we can clearly say that this distribution is Normally distributed because it is evenly aligned with the standard normal variate which is the simple concept of Q-Q plot.

2.2.4 Shapiro wilks Normality Test

The Shapiro-Wilk test tests the null hypothesis that a sample x_1,x_2,….x_n came from a normally distributed population. The test statistics is

$$W=\frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Where

- x_i is the i^{th} order statistic, i.e., the ith-smallest number in the sample.
- $\bar{x} = \frac{(x_1+x_2+\ldots.+x_n)}{n}$ is the sample mean.

The coefficients a_i are given by:

$$(a_1, a_2, \ldots a_n) = \frac{m^T V^{-1}}{C}$$

Where C is a vector norm

$C=\|V^{-1}m\| = (m^T V^{-1} V^{-1} m)^{1/2}$

Is made of the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution; finally, V is the covariance matrix of those normal order statistics.

There is no name for the distribution of W. The cut-off values for the statistics are calculated through Monte Carlo simulations.

The three tests used to check normality in time series are Shapiro-Wilks (SW), Kolmogorov-Smirnov (KS), and the Anderson Darling (AD) test.

The Shapiro-Wilks test for normality is one of three general normality tests designed to detect all departures from normality. It is comparable in power to the other two tests. The test rejects the hypothesis of normality when the p-value is less than or equal to 0.05. Failing the normality test allows you to state with 95% confidence the data does not fit the normal distribution. Passing the normality test only allows you to state no significant departure from normality was found.

The Shapiro-Wilks test is not as affected by ties as the Anderson-Darling test, but is still affected.

2.2.4 Kolmogorov Smirnov Test

We can use different test method in order to check normality of a data. One of such tests is a K-S test. The **Kolmogorov–Smirnov test** (**K-S test** or **KS test**) is a non-parametric test of the equality of continuous (or discontinuous), one-dimensional probability distribution that can be used to compare a sample with a reference probability distribution (one-sample K–S test), or to compare two samples (two-sample K–S test). The Kolmogorov–Smirnov test can be modified to serve as a goodness of fit test. In the special case of testing for normality of the distribution, samples are standardized and compared with a standard normal distribution. This is equivalent to setting the mean and variance of the reference distribution equal to the sample estimates, and it is known that using these to define the specific reference distribution changes the null distribution of the test statistic The Kolmogorov-Smirnov test is used to test the null hypothesis that a set of data comes from a normal distribution. The empirical

distribution function F_n for n independent and identically distributed (i.i.d.)ordered observations X_i is defined as

$$F_n = \frac{number \text{ of (elements in the sample} \leq x)}{n}$$

$$= \frac{1}{n} \sum_{i=1}^{n} 1\ [-\infty, x]\ (X_i)$$

Where $1[-\infty, x]\ (X_i)$ is the indicator function, equal to 1 if $X_i \leq x$ and equal to 0 otherwise. The Kolmogorov–Smirnov statistic for a given cumulative distribution function $F(x)$ is

$$D_n = \sup_x |Fn(x)-F(x)|$$

where $\sup_x$ is the supremum of the set of distances. Intuitively, the statistic takes the largest absolute difference between the two distribution functions across all x values. By the Glivenko–Cantelli theorem, if the sample comes from distribution $F(x)$,then D_n converges to 0 almost surely in the limit when n goes to infinity. Kolmogorov strengthened this result, by effectively providing the rate of this converge. In practice, the statistic requires a relatively large number of data points to properly reject the null hypothesis.

2.3 Basic Definitions

Autocorrelation Function (ACF)

The autocorrelation function of a stationary time series $\{X_t\}$, $\rho(k)$ at lag k is defined as the correlation at lag k between X_t and X_{t+k}. Thus, the autocorrelation function at lag k is given by

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)}$$

Where $\gamma(k) = cov(X_t, X_{t+k})$.

Partial Autocorrelation Function (PACF)

The partial autocorrelation function, like autocorrelation function, conveys vital information regarding the dependence structure of a stationary process. In the context of time series, a large portion of correlation between X_t and X_{t-k} can be due to the correlation these variables have with $(X_{t-k}, X_{t-2}, \dots, X_{t-k+1})$. The partial autocorrelation of lag k can be thought of as partial regression coefficient ϕ_{kk} in the representation

$$X_t = \phi_{k1}X_{t-1} + \phi_{k2}X_{t-2} + \cdots + \phi_{kk}X_{t-k} + \varepsilon_t.$$

Thus, the partial correlation at lag k, ϕ_{kk}, measures the correlation between X_t and X_{t-k} after adjusting for the effects of $(X_{t-1}, X_{t-2}, \dots, X_{t-k+1})$.

Autoregressive (AR) process

A time series $\{X_t\}$ is said to be an autoregressive process of order p, abbreviated as AR(p), if it is weigted linear sum of past values plus a random shock so that

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-1} + \cdots + \phi_p X_{t-p} + \varepsilon_t,$$

Where $\{\varepsilon_t\}$ denotes a purely random process with zero mean and a constant variance σ^2. Using backward shift operator B, such that $BX_t = X_{t-1}$, the AR(p) model may be written in the form

$$\phi(B)X_t = \varepsilon_t,$$

Where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ is a polynomial in B of order p.

Moving Average (MA) process

A time series $\{X_t\}$ is said to be a moving average process of order q abbreviated as MA(q), if it can be written in the form

$$X_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q},$$

Where $\{\varepsilon_t\}$ denotes a purely random process with zero mean and a constant variance σ^2. Using backward shift operator B, the MA(q) model may be written in the form

$$X_t = \theta(B)\varepsilon_t$$

Where $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ is a polynomial in B of order q.

ARMA process

A mixed autoregressive moving average model with p autoregressive terms and q moving average terms is abbreviated as ARMA(p, q) model may be written as

$$X_t - \phi_1 X_{t-1} + \phi_2 X_{t-1} + \cdots + \phi_p X_{t-p} = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q}$$

Using backward shift operator B, the ARMA(p, q) modelmay be written in the form

$$\phi(B)X_t = \theta(B)\varepsilon_t,$$

Where $\phi(B)$ and $\theta(B)$ are polynomials in B OF order p and q respectively.

ARIMA process

The AR, MA and ARMA processes apply only to stationary data series. A stationary time series has a mean, variance and autocorrelation function that are essentially constant throughout time. In practice many time series are non-stationary and so we cannot apply stationary AR, MA or ARMA processes directly.

One possible way of handling non stationary series is to apply differencing so as to make them stationary. The first difference namely $X_t - X_{t-1} = (1 - B)X_t$, may themselves be differenced to give second differences, and so on. The d^{th} difference may be written as $(1 - B)^d X_t$. If the original data series is differenced d times before fitting an ARMA(p, q) process, then the model for the original undifferenced series is said to be an ARIMA(p, d, q)

process, where the letter 'I' in the acronym stands for integrated and d denotes the number

$$\phi(B)(1-B)^d X_t = \theta(B)\varepsilon_t,$$

Where $\{\varepsilon_t\}$ denote a purely random process with zero mean and constant variance σ^2.

In addition to the standard autoregressive and moving average parameters, ARIMA models may also include a constant. The interpretation of a (statistically significant) constant depends on the model that is fit. Specifically, (a) if there are no autoregressive parameters in the model, then the expected value of the constant is μ, the mean of the series; (b) if there are autoregressive parameters in the series, then the constant represents the intercept. If the series is differenced, then the constant represents the mean or intercept of the differenced series. For example, if the series is differenced once, and there are no autoregressive parameters in the series, then the constant represents the mean of the differenced series, and therefore the linear trend slope of the undifferenced series.

Stationarity

The ARIMA method is appropriate only for a data series that is stationary. Stationarity implies that the AR coefficients must satisfy certain conditions for an ARIMA model to be stationary. There is a common-sense reason for requiring stationarity: we could not get useful estimates of the parameters of a process otherwise.

If $p = 0$, we have either a pure MA model or a white noise series. All pure MA models and white noise are stationary, so there are no stationarity conditions to check.

For an AR (1) or ARMA (1,q) process, the stationary requirement is that the absolute value of $\phi_1$must be less than one i.e. $|\phi_1| < 1$.

For an AR (2) or ARMA(2,q) process, stationary requirements is a set of three conditions: $|\phi_2| < 1, \phi_1 + \phi_2 < 1$ and $\phi_1 - \phi_2 < 1$

Invertibility

There is another condition that ARIMA models must satisfy called invertibility. This requirement implies that the MA coefficients must satisfy certain conditions. A non-invertible ARIMA model implies that the weights placed on past X observations do not decline as we

move further into the past; but logic says that larger weights should be attached to more recent observations. Invertibility ensures that this result holds.

If $q = 0$, we have either a pure AR process or a white noise series. All pure AR processes and white noise are invertible, so there are no invertibility conditions to check.

For an MA (1) or ARMA(p,1) process, the invertibility requirement is that the absolute value of θ_1 must be less than one i.e., $|\theta_1| < 1$.

For an MA (2) or ARMA(p,2) process, invertibility requirements is a set of three conditions: $|\theta_2| < 1, \theta_1 + \theta_2 < 1$ and $\theta_1 - \theta_2 < 1$.

Characteristics of a Good ARIMA Model

1. A good model is parsimonious.
2. A good AR model is stationary.
3. A good MA model is invertible.
4. A good model has high quality estimated coefficients at the estimation stage.
5. A good model has statistically independent residuals.
6. A good model fits the available data sufficiently well at the estimation stage.
7. A good model has sufficiently small forecast error.

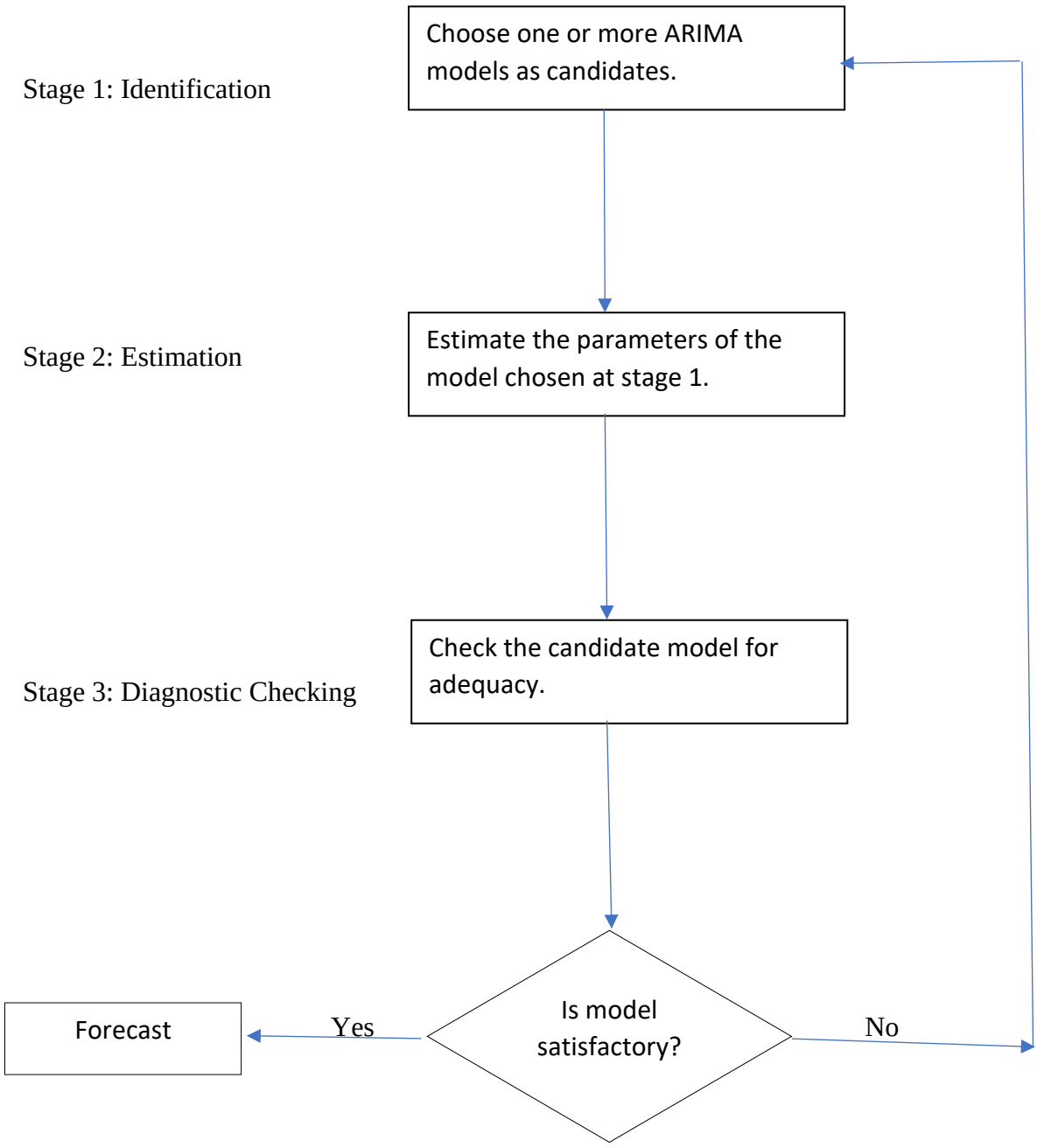
Limitations of ARIMA Model

The ARIMA method is appropriate only for a time series that is stationary i.e., it's mean, variance and autocorrelation should be approximately constant through time. It is recommended that there are at least 50 observations in the input data. It is also assumed that the values of the estimated parameters are constant throughout the series.

2.4 Box-Jenkins Modelling Procedure

The general model introduced by Box and Jenkins (1976) includes autoregressive as well as moving average parameters, and explicitly includes differencing in the formulation of the model. Specifically, the three types of parameters in the model are: the autoregressive parameters(p), the number of differencing required (d) and the moving average parameters(q).

Box and Jenkins propose a practical three-stage procedure for find a good model. The three stage Box and Jenkins procedure is summarized schematically in the following figure.



Stages in Box-Jenkins Iterative approach to model building.

2.4.1 Model Identification

The inputs series for ARIMA needs to be stationary, i.e., it should have a constant mean and variance. Therefore, usually the series, first needs to be differenced until it is stationary. This also of then requires log transforming the data to stabilize the variance. The number of times the series needs to be differenced to achieve stationarity is reflected in the d parameter. In order to determine the necessary level of differencing, one should examine the plot of the data and auto correlogram. Significant changes in level (strong upward or downward changes) usually require first order non seasonal (lag=1) differencing; strong changes of slope usually require second order non seasonal differencing. Seasonal patterns require respective seasonal differencing. If the estimated autocorrelation coefficients decline slowly at longer lags, first order differencing is usually needed. However, one should keep in mind that sometimes series may require little or no differencing and that over differenced series produce less stable coefficient estimates.

At this stage, we also need to decide how many auto-regressive (p) and moving average (q) parameters are necessary to yield an effective but still parsimonious model of the process (parsimonious means that it has the fewest parameters and greatest number of degrees of freedom among all the models that fit the data). In practice, the number of p and q parameters very rarely need to be greater than 2.

Before the estimation can begin, we need to decide on (identify) the specific number and type of ARIMA parameters to be estimated. The major tools used in the identification phase are plots of time series, correlograms of autocorrelations (ACF) and partial autocorrelation (PACF). The decision is not straight forward and in less typical cases requires not only experience but also a good deal of experimentation with alternative models (as well as the technical parameters of ARIMA). However, a majority models that can be identified based on the shape of the auto-correlogram and partial auto-correlogram. Also note that since the number of parameters to be estimated of each kind is almost never greater than 2, it is often practical to try alternative models on the same data. The following table summarizes the important results for selecting p and q.

	ACF	PACF
AR(p)	Tails off as exponential decay or damped sine wave.	Cut off after lag p.
MA(q)	Cut off after lag q.	Tails off exponential decay or damped sine wave.
ARMA(p,q)	Tails off after lag q-p.	Tails off after lag p-q.

2.4.2 Parameter Estimation

At this stage we get precise estimates of the co-efficient of the model chosen at the identification stage. There are several different methods for estimating parameters. All of them should produce very similar estimates, but may be more or less efficient for any given model. In general, during the parameter estimation phase a function minimization algorithm is used to maximize the likelihood of the observed series, given the parameter values. In practice, this requires the calculation of the (conditional) sum of squares (SS) of residuals, given the respective parameters. Different methods have been proposed to compute the SS for residuals: (i) the approximate maximum likelihood method according to McLeod and Sales (1983). (ii) the approximate maximum likelihood method with back casting and (iii) the exact maximum likelihood method according to Mclard (1984).

In general, all methods should yield very similar parameter estimates. Also, all methods are about equally efficient in most real-world time series applications. However, the first method is the fastest and should be used in particular for every long time series (e.g., with more than 30,000 observations). Mclard's exact maximum likelihood method may also become inefficient when used to estimate parameters for seasonal models with long seasonal lags (e.g., with yearly lags of 365 days). On the other hand, we should always use the approximate maximum likelihood method first in order to establish initial parameter

estimates that are very close to the actual final values; thus, usually only a few iterations with the exact maximum likelihood method are necessary to finalize the parameter estimates.

The estimates of the parameter are used in the forecasting stage to calculate new values of the series and confidence intervals for those predicted values. The estimation process is performed on transformed (differenced) data; before the forecasts are generated, the series needs to be integrated (integration is the inverse of differencing) so that the forecasts are expressed in values compatible with the input data.

2.4.3 Diagnostic Checking

Box and Jenkins suggest some diagnostic checks to determine if an estimated model is statistically adequate. A model that fails these diagnostic tests is rejected. Furthermore, the results at this stage may also indicate how a model could be improved. This lead us back to the identification stage. We repeat the cycle of identification, estimation and diagnostic checking until we find a good final model.

In an ARIMA model, the random shocks, ε_t, is assumed to follow the assumptions for a stationary univariate process. The random shocks should be white noise (or independent when their distribution is normal) drawing from a fixed distribution with a constant mean and variance. In practice we cannot observe the random shocks, but we do have estimate of them; we have the residuals, e_t, calculated from the estimated model. If Box-Jenkins model is a good model for the data, the residual should satisfy the assumption of random shocks.

The basic analytical tool at the diagnostic checking stage is the residual ACF. The idea behind the use of residual ACF is this: if the estimated model is properly formulated, then the random shocks (ε_t) should be uncorrelated. If the random shocks are uncorrelated, then the estimates of them (e_t) should also be uncorrelated on average. Therefore, the residual ACF for a properly built ARIMA model will ideally have autocorrelation coefficients that are all statistically zero.

A check of the normality assumption can be made by plotting a histogram of the residuals or using a normal probability plot. The assumption that random shocks have mean zero and constant variance can be checked at the same time. To do this, we use a residual

plot. A residual plot is a scatter plot of residuals against fitted values. If the assumptions are satisfied, we would expect the residuals to vary randomly around zero and spread of the residuals to be about the same throughout the plot. If the points in the plot in a curve around zero, rather than fluctuating randomly, it is an indication that the mean zero assumption is broken. If the residuals seem to increase or decrease in average magnitude with the fitted values, it is an indication that the variance of the residuals is not constant.

2.5 FINANCIAL SERIES

Modelling financial time series is a complex problem. This complexity is not only due to the variety of the series in use (stocks, exchange rates, interest rates, etc.), to the importance of the frequency of d observation (second, minute, hour, day, etc.) or to the availability of very large data set. It is mainly due to the existence of statistical regularities which are common to a large number of financial series and are difficult to reproduce artificially using stochastic models. The properties that we now present are mainly concerned with daily stock prices.

Let p_t denote the price of an asset at time *t* and let $r_t = \ln\left(\frac{p_t}{p_{t-1}}\right)$ be the return. In contrast to the prices, the returns do not depend on monetary units which facilitate comparisons between assets.

2.6 ARCH AND GARCH MODELS

ARCH (*q*) and GARCH (*p, q*) models have become important tools in the analysis of time series data, particularly in financial applications. These models are especially useful when the goal of the study is to analyse and forecast volatility. Data in which the variances of the error terms are not equal, in which the error terms may reasonably be expected to be larger for some points or ranges of the data than for others, are said to be suffer from heteroscedasticity.

Heteroscedasticity is often associated with cross-section data, whereas time series are usually studied in the context of homoscedastic process. In analysis of macroeconomic data, Engle (1982, 1983) and Cragg (1982) found evidence that, for some kind of data, the disturbance of the variance was less stable than usually assumed. Engle's result suggests that in models of inflation, large and small forecast errors appeared to occur in clusters,

suggesting a form of heteroscedasticity in which the variance of the forecast error depends on the size of previous disturbance. He suggests the 'Auto Regressive Conditionally Heteroscedastic' (ARCH) model as an alternative to the usual time series process and their GARCH (generalized ARCH) extension is due to Bollerslev (1986). In these models, the key concept is the conditional variance, that is, the variance conditional on the past. In the classical GARCH models, the conditional variance is expressed as a linear function of the squared past values of the series.

The GARCH model has been proved to be useful in studying the volatility of inflation (Carlson and Robins, 1985), the term structure of interest rates (Engle, Hendry and Trumbull, 1985), the volatility of stock market returns (Engle, Li lien and Robins, 1987) and the behaviour of foreign exchange markets (Domoneitz and Hakkio, 1985) etc.

A process $\{y_t\}$ is called a GARCH (p, q) process if its first two conditional moments exist and satisfy:

$y_t = \sigma_t \varepsilon_t,\ \varepsilon_t \sim N(0,1)$

(i) E ($\frac{y_t}{y_u}$, u < t) =0

(ii) σ_{t^2} = var ($\frac{y_t}{y_u}$, u < t)

(iii) There exists constant ω, $\alpha_i, i = 1,2,\dots,q$ and $\beta_j, j = 1,2,\dots,p$ such that

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i\, y_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \quad , t \in T$$

can be written in a more compact way as,

$$\sigma_t^2 = \omega + \alpha(B) y_t^2 + \beta(B)\sigma_t^2 \quad , t \in T$$

where B is the backward shift operator and α and β are polynomials in degree p and q respectively and

$$\alpha(B) = \textstyle\sum_{i=1}^{q} \alpha_i\, B^i \ ,\ \beta(B) = \sum_{j=1}^{p} \beta_j B^j$$

2.7 Akaike's Information Criterion and Bayesian Information Criterion

Akaike's information criterion, developed by Hirotsugu Akaike (1971) under the name of '*an information criterion (AIC)'* is a measure of the goodness of fit of an estimated statistical model. It is grounded in the concept of entropy, in effect offering a relative measure of the information lost when a given model is used to describe reality and can be said to describe the trade-off between bias and variance in model construction or loosely speaking that of precision and complexity of the model.

The AIC is not a test on the model in the sense of hypothesis testing; rather it is a tool for model selection. Given a data set, several competing models may be ranked according to their AIC, with the one having the lowest being the best. In the general case, the AIC is

$$AIC = -2\ln(L) + 2k,$$

Where k the number of parameters is in the statistical model and L is the maximized value of the likelihood function for the estimated model. Over the remainder of this entry, it wil be assumed that the model errors are normally and independently distributed. Let n be the number of observations and RSS be

$RSS = \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \sum_{i=1}^{n} e_i^{\,2}$, residual sum of squares.

Then AIC becomes, AIC$= n\left[\ln\left(\frac{2\pi RSS}{n}\right) + 1\right] 2k.$

Increasing the number of free parameters to be estimated improves the goodness of fit, regardless of the number of free parameters in the data generating process. Hence AIC not only rewards goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. This penalty discourages over fitting. The preferred model is one with the lowest AIC value. The AIC methodology attempts to find the model that best explains the data with a minimum of free parameters. AIC judges a model by how close its fitted values tend to be the true values, in terms of a certain expected value.

Shibata (1976) has shown that the AIC tends to overestimate the order of auto regression. More recently, Akaike's (1978) has developed a *Bayesian Information criterion (BIC),* which is of the following form

$BIC = \ln(MSE) + k\frac{\ln(n)}{n}$, where MSE$= \frac{1}{n-k}\sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}$

2.8 Test for Stationarity: Unit Roots Test

The usual procedure to check Stationarity is to calculate the sample ACF and see if the coefficients die out. If the coefficients do not die out, the process is not stationary. While the above procedure has traditionally been followed, Unit root tests such as the Dickey-Fuller test can be employed. A unit root tests whether a time series is non-stationary using an autoregressive model.

Dickey –Fuller Test

Let X_1, X_2…, X_n be the observations from AR (1) model

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \varepsilon_t \qquad (1)$$

Where $\{\varepsilon_t\}$ denotes purely random process with mean zero and constant variance σ^2 and $\mu = E(X_t)$

For large n the m.l.e. $\hat{\phi}_1$ of ϕ_1 is approximately $N(\phi_1, \frac{1-{\phi_1}^2}{n})$.

The test hypotheses are $H_0 = \phi_1 = 1$ against $H_1 = \phi_1 < 1$.

To construct a test ofH_0 write the model (1) as

$$\nabla X_t = \phi_0^* + \phi_1^* X_{t-1} + \varepsilon_t,$$

Where $\phi_0^* = \mu(1 - \phi_1)$ and $\phi_1^* = \phi_1 - 1$

Now let $\hat{\phi}_1^*$ be the ordinary least square estimator of ϕ_1^*.

The estimated standard error of $\hat{\phi}_1^*$ is

$$\widehat{S.E}\left(\hat{\phi}_1^*\right) = S\left[\sum_{t=2}^{n}(X_{t-2} - \bar{X})^2\right]^{1/2}$$

Where $S^2 = \frac{1}{n-3}\sum_{t=2}^{n}\left[\nabla X_t - \hat{\phi}_0^* - \hat{\phi}_1^* X_{t-1}\right]^2$ and $\bar{X}$ is sample mean of $X_1, X_2, \ldots, X_{n-1}$

Dickey and Fuller derived the limit distribution as n tends to infinity of the t-ratio,

$$\hat{\tau}_\mu = \frac{\hat{\phi}_1^*}{\widehat{S.E}(\hat{\phi}_1^*)} \qquad (2)$$

Under the unit root assumption $\hat{\phi}_1^* = 0$, from which a test of null hypothesis $H_0: \phi_1 = 1$ can be constructed.

The 0.01, 0.05 and 0.10 quantiles of the limit distribution of $\hat{\tau}_\mu$ are -3.43, -2.86, -2.57 respectively. The augmented Dickey-Fuller test then rejects the null hypothesis of a unit root at 5% level, if $\hat{\tau}_\mu < -2.86$.

The above procedure can be extended to the case where $\{Z_t\}$ follow AR(p) model with mean μ given by,

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \cdots + \phi_p(X_{t-p} - \mu) + \varepsilon_t,$$

Where $\{\varepsilon_t\}$ is purely random process with mean zero and constant variance σ^2.

This model can be rewritten as,

$$\nabla X_t = \phi_0^* + \phi_1^* X_{t-1} + \phi_2^* \nabla X_{t-1} + \cdots + \phi_p^* \nabla X_{t-p+1} + \varepsilon_t,$$

Where $\phi_0^* = \mu(1 - \phi_1 - \cdots - \phi_p)$, $\phi_i^* = \sum_{i=1}^{p}(\phi_i - 1)$, and $\phi_j^* = -\sum_{i=j}^{p} \phi_i$,

$$j = 2,3, \ldots p$$

If the autoregressive polynomial has a unit root at 1, then $\phi(1) = 0 \Rightarrow \phi_1^* = 0$ and the differenced series $\{\nabla X_t\}$ is an AR$(p-1)$ process.

For the large n the t-ratio,

$$\hat{\tau}_\mu = \frac{\hat{\phi}_1^*}{\widehat{S.E}(\hat{\phi}_1^*)} \qquad (3)$$

has the same limit distribution as the test statistics (2). The ADF test in this case is applied in exactly the same manner as for the AR (1) case using the test statistics (3) and the cut off values given above.

2.9 Forecasting and measure of forecast accuracy

Forecasting the future values of an observed time series is an important problem in many areas, including economics, production planning, sales forecasting and stock control. Suppose we have an observed time series y_1, y_2…. y_n. Then the basic problem is to estimate future values such as y_{N+h}. Where the integer h is called the lead time or forecasting horizon. The forecast. Of y_{N+h} made at time N for h steps ahead is typically denoted by {\hat{Y}} _N(h). A wide variety of different forecasting procedures is available and it is important to realize that no single method is universally applicable. Rather, the analyst must choose the procedure that is most appropriate for a given set of conditions.

In statistics, a forecast error is the difference between the actual and the predicted or forecasted value of a time series. Suppose we have a system for forecasting the value of some variable at time t and we have some actual values for the same variable. Let y_t denote the actual value at time t and $\hat{y}_t$ denote the forecast at time t. A forecast is never completely accurate; forecasts will always deviate from the actual values. The objective of forecasting is that to predict future values with no forecast error or with error as slight as possible. There are many measures of forecasting accuracy; the more popular ones are:

(i) Mean Absolute Deviation (MAD)

$$\text{MAD}= \frac{1}{n} \sum_{t=1}^{n} |y_t - \hat{y}_t|$$

(ii) Mean Square Error (MSE)

$$\text{MSE}=\frac{1}{n} \sum_{t=1}^{n} (y_t - \hat{y}_t)^2$$

(iii) Mean Absolute Percentage Error (MAPE)

$$\text{MAPE}= \frac{1}{n} \sum_{t=1}^{n} \left|\frac{y_t-\hat{y}_t}{y_t}\right|$$

CHAPTER 3

3.1 DATA ANALYSIS

3.1.1 Time Series Modelling

In this chapter, we consider the series modelling and forecasting of the daily opening stock price of Netflix and it is obtained from the website Kaggle. The data consist of 1009 observations starting from 5th February 2018 to 4th February 2022. The process of model fitting for daily opening of stock price was done using the statistical packages SPSS and R.

A time series plot of daily opening stock price of Netflix appears in *Figure* 1 Here we can see that the price rises and falls through the time, and is very clear that data exhibit high level of volatility during the period of analysis. Only from the visual inspection of the data we cannot say if these changes in the mean are statistically significant. Accurate forecast is difficult unless the cause of fluctuations can be identified. We take natural logarithm to stabilize variance.

So, we must rely on the appearance of the auto correlation function and partial auto correlation to decide whether the mean is stationary or not. A plot of ample auto correlation and partial auto correlation function of daily closing stock price of Netflix stock price is shown in *Figure 3.1* and *Figure 3.2* respectively.

Time Series Plot Of Daily Opening Stock Price of Netflix

R code:

```
library(readr)

library(tseries)

nflx <- read_csv("E:\\nflx.csv")

View(nflx)

attach(nflx)
```

```
y=ts(Open,start=c(2018,5),end=c(2022,48),frequency=365)

data.frame(y)

plot.ts(y)
```

Output:

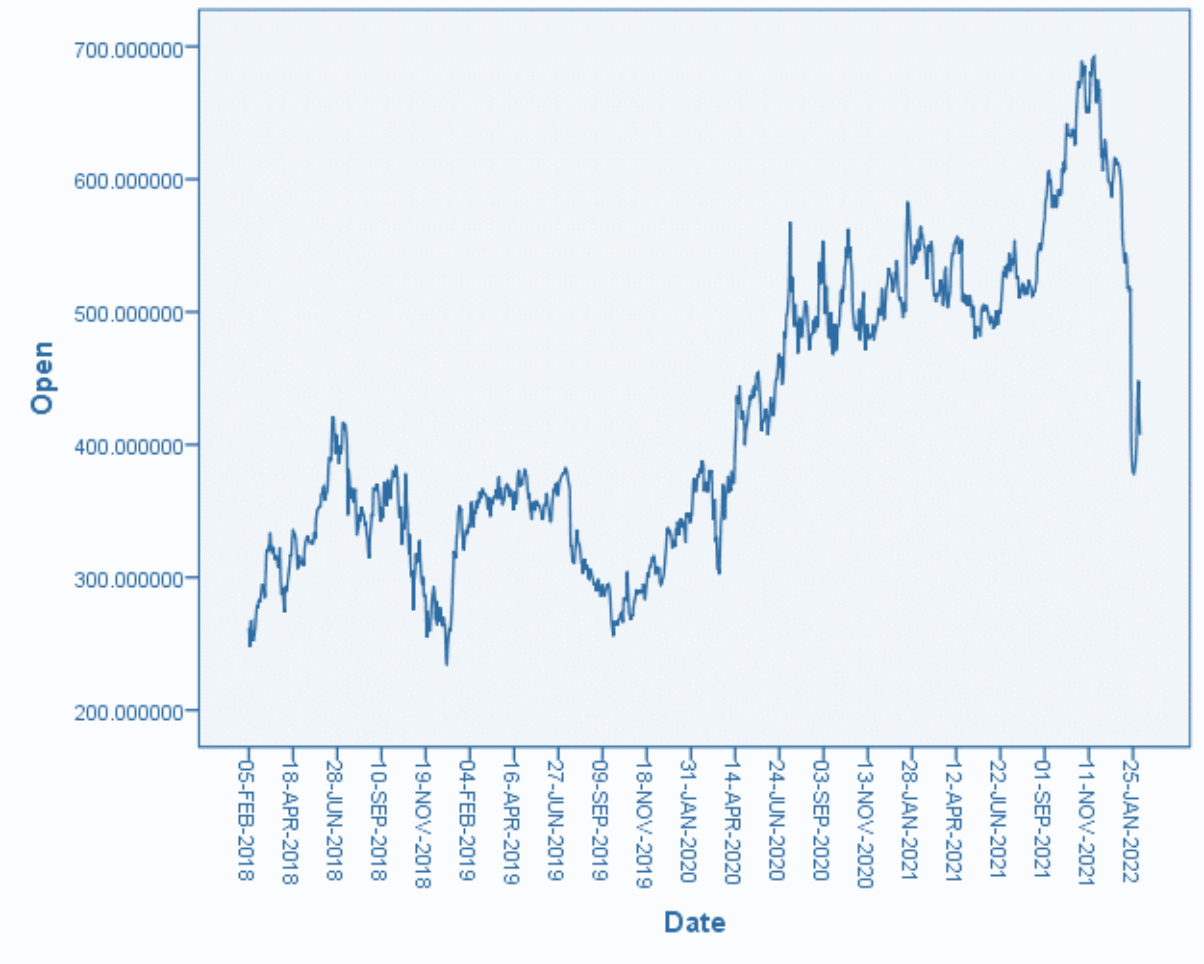


Figure 1: *Time Series plot of daily opening stock price of Netflix from 5th February 2018 to 5th February 2022*

From figure 1 we can see that our data has trend,seasonality and is non stationary in mean.

3.1.2 Normality Test

Normality test identifies the nature of the distribution of data of a given variable. Normal distribution of data refers to the closeness of every observation in a dataset to its mean score. This means that the normality test should ideally be conducted before undertaking a time series analysis. Testing of normality is an important decision as most of the parametric statistical tests that we consider rely on the assumption that variables are normally distributed, unless sample sizes are very large. To check the normality of the corresponding data we will look at this both graphically by constructing Histogram, Q-Q plot and through statistical tests known as Kolmogorov-Smirnov test and Shapiro-Wilk test.

Before Modelling a time series, we should check whether the data is normal or not. We can check Normality using graph and also through performing any corresponding test.

Histogram And Q-Q Plot

If the histogram indicates a symmetric, moderate tailed distribution, then the recommended next step is to do a normal probability plot to confirm approximate normality. If the normal probability plot is linear, then the normal distribution is a good model for the data.

From figure 2.1 it is clear that our data is not normally distributed. However

histogram is the not generally the best type of graph for determining whether our data are normally distributed or not. Normal probability plots are a better choice and they are easy to use. Normal probability plots are also known as quantile-quantile plots, or Q-Q plots. So, now in figure 2.2 we plot a Q-Q plot.

R code

```
hist(y,col="blue")
qqnorm(y,col="green")
qqline(y,col="red")
```

Output

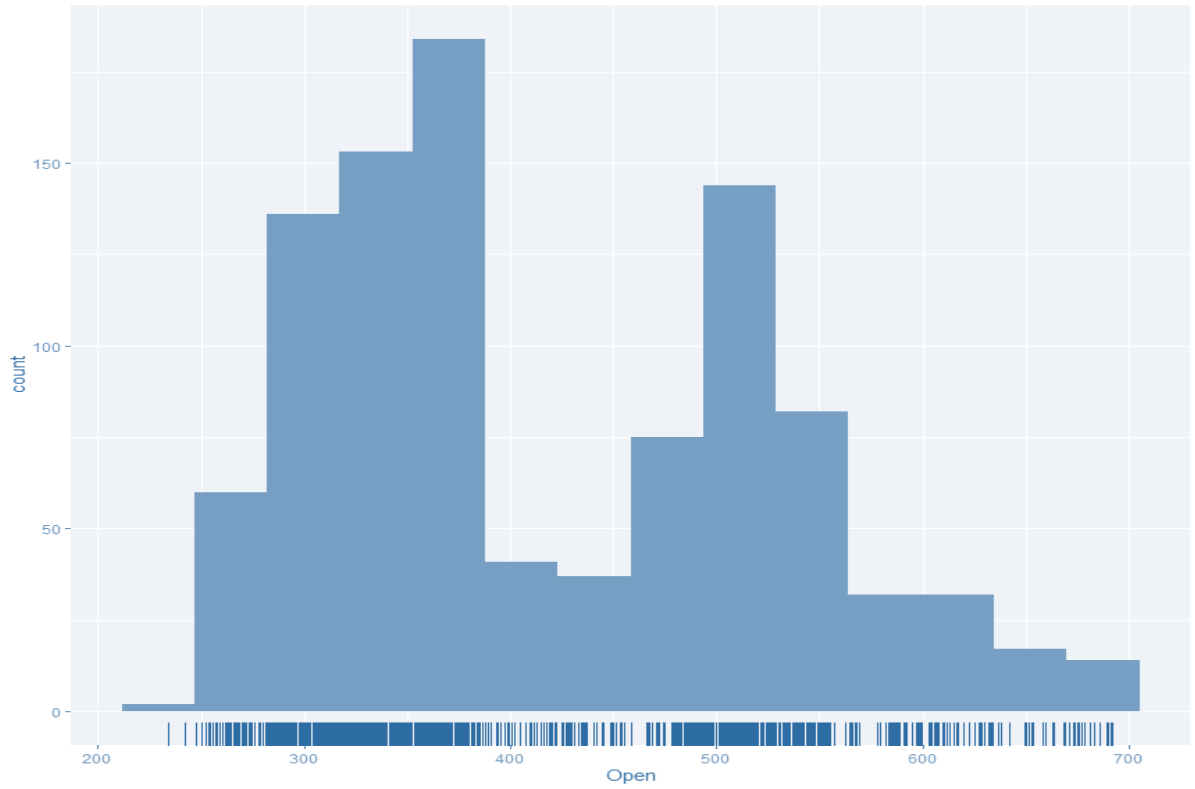


Figure 2.1: *Histogram of daily opening stock price of Netflix*

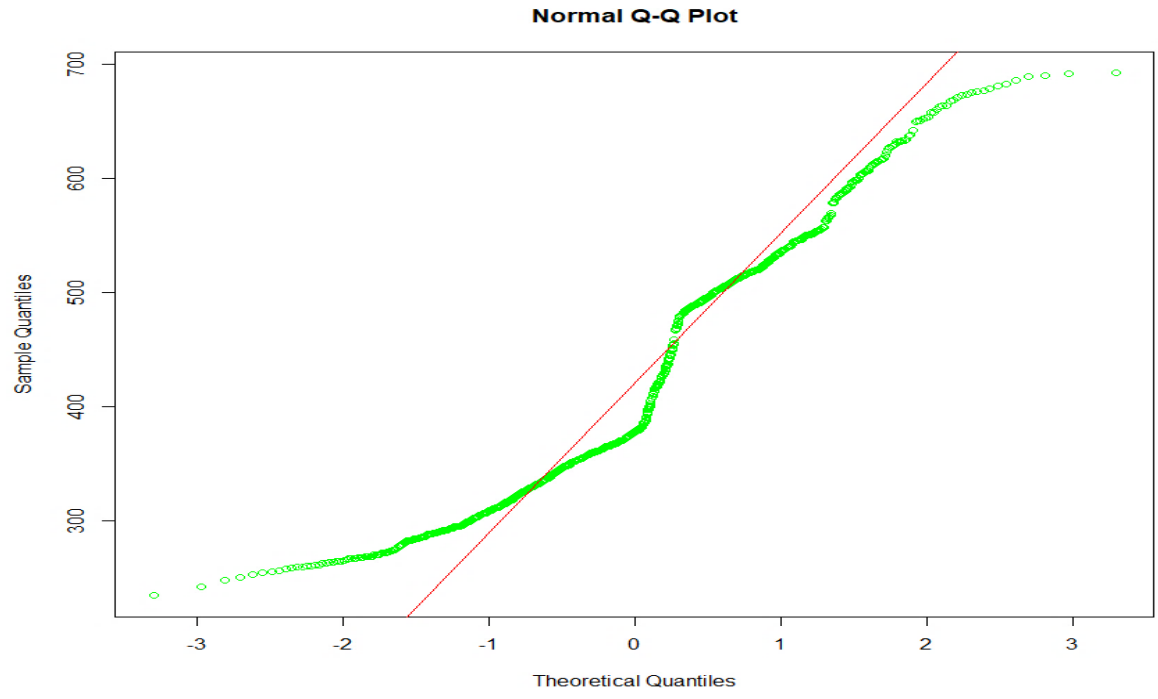


Figure 2.2: *Q-Q plot of daily opening stock price of Netflix*

From figure 2.2 we can see that our data points don't follow the straight line. Hence, our data is not normally distributed.

Shapiro-Wilk Test

R code

```
shapiro.test(y)
```

Output

```
Shapiro-Wilk normality test

data:  y
W = 0.93793, p-value < 2.2e-16
```

Kolmogorov-Smirnov Test

R code

```
ks.test(y,"pnorm")
```

Output

```
One-sample Kolmogorov-Smirnov test

data:  y
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

by using Shapiro-wilk normality test and K-S test in R the p value becomes less than 0.05 and is statistically significant hence rejecting the null hypothesis (data is normal), we can conclude that our data is not normal.

3.1.3 STATIONARITY

From our time series plot we can see that our data is not stationary and has a trend and seasonality. In order to make this assumption more specific we construct the ACF, PACF and also do the Dicky Fuller Test.

In figure 3.1 Autocorrelation function plot of the data is shown.

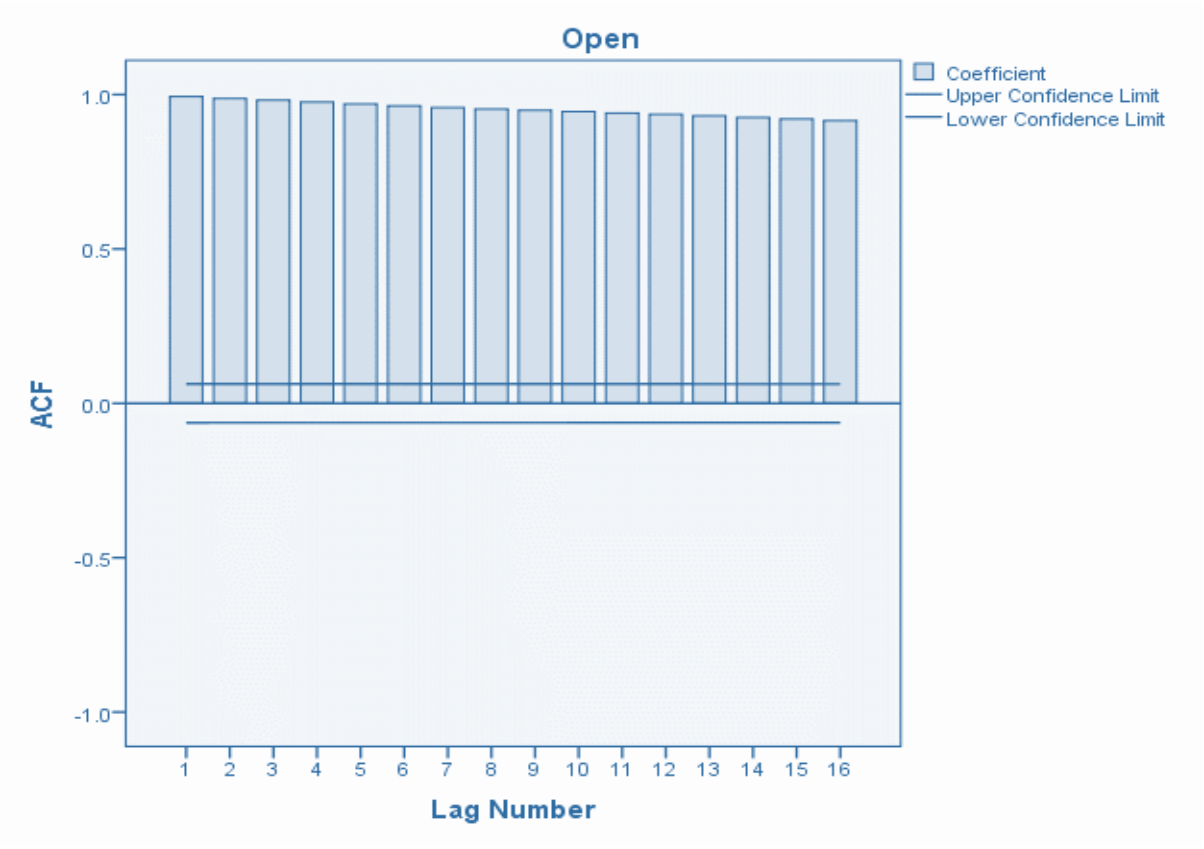


Figure3.1*: Autocorrelation function of daily opening stock price of Netflix.*

The PACF plot is given in figure 3.2

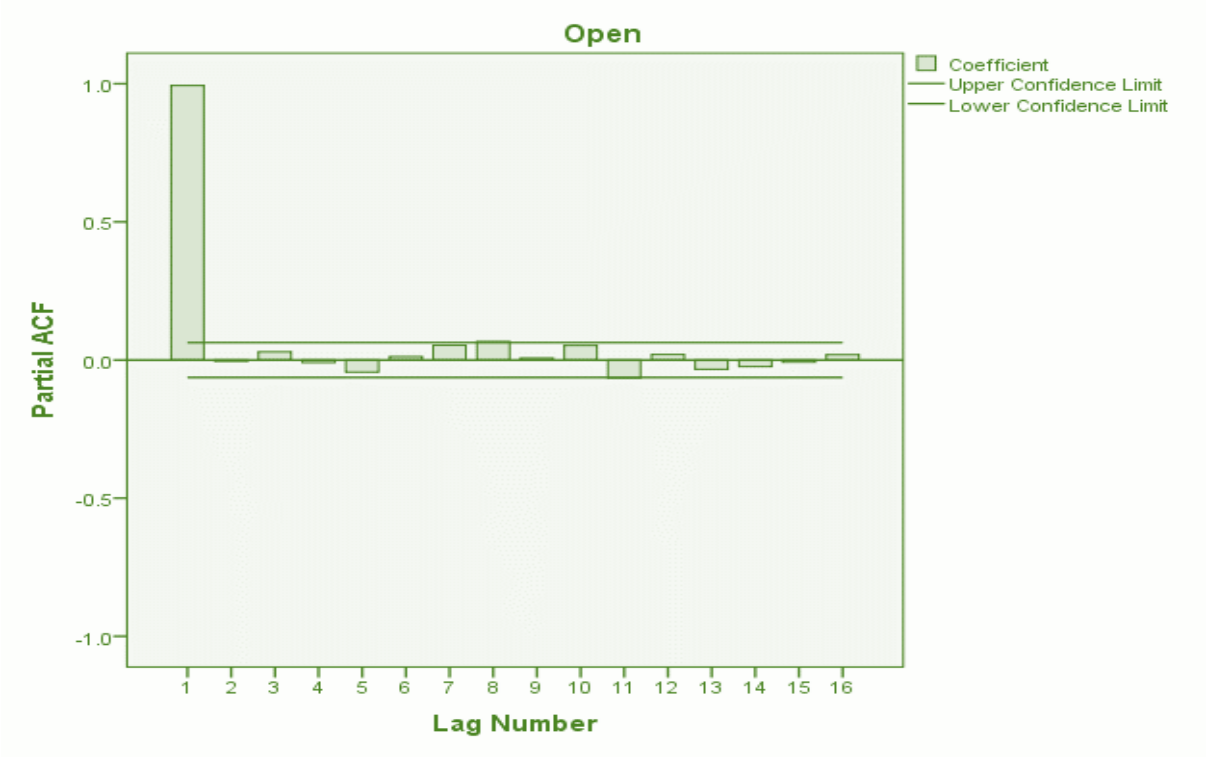


Figure 3.2 *Partial autocorrelation function of daily opening stock price of Netflix.*

Here *Figure 3.1* indicates that the ACF fails to die out rapidly towards zero. Rather than dying down along the first several lags, the graph of ACF displays a slow decrease in the size of ACF values, which is a typical pattern for non-stationary series. From this it is clear that the original data is non-stationary.

Since the data is a financial time series, we convert the series of market prices into return series and find the estimated ACF and PACF for this return series. A time series plot of returns is shown in *Figure* 3.3

In order to test whether the transformed data is stationary, we use a unit root test; there are statistical hypothesis test of stationarity. One of the most popular unit root tests is the Augmented Dickey-Fuller (ADF) test. The R code and output of the ADF test is given below:

R code

```
adf.test(y)
```

output:

Augmented Dickey-Fuller Test

data: y

Dickey-Fuller = -1.8185, Lag order = 10, p-value = 0.6552

alternative hypothesis: stationary

Here the p value is greater than 0.05 hence we accept the null hypothesis therefore the data is non stationary.

In order to make the data normal we apply log to the variable and then check the volatility present in the data

R code:

```
library(rugarch)
library(vrtest)
library(tseries)
ry1=log(y)
Auto.VR(ry1)
ArchTest(ry1)
```

Output:

$stat

[1] 181.6556

$sum

[1] 1006.145

ARCH LM-test; Null hypothesis: no ARCH effects

data: ry1

Chi-squared = 986.18, df = 12, p-value < 2.2e-16

Here the statistical value is very high hence the variance is not constant. Also, when we check Arch test p value is less than 0.05 resulting in rejecting the null hypothesis. Thus, there is ARCH effect. GARCH models are used when the variance of the error term is not constant. i.e., when the error terms are heteroskedastic. Since the variance are not constant and there is ARCH effect GARCH model can be fitted.

Now, in order to convert the data into stationary first we apply log to the variable, If the data is still non stationary, we will need to take difference between a value and lag of a value from the transformed data.

R code:

```
Ry1=diff(log(y))

adf.test(ry)
ry=ts(Ry1=ts(Open,start=c(2018,5),end=c(2022,48),frequency=365)
plot.ts(r)
```

Output:

Augmented Dickey-Fuller Test

data: ry

Dickey-Fuller = -10.277, Lag order = 10, p-value = 0.01

alternative hypothesis: stationary

Here the p value is lesser than 0.05 hence we reject the null hypothesis thus the data becomes stationary.

Time Series Plot of Returns

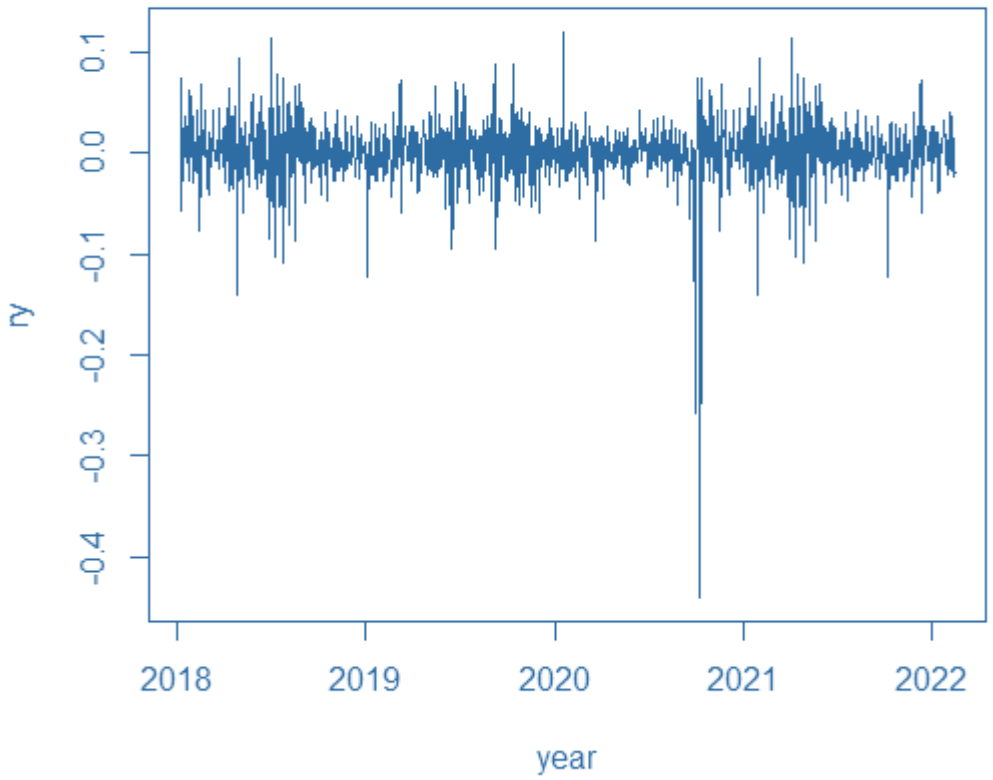


Figure 3.3: *Time series plot of returns*

Time series plot of logged data is shown in Figure 3.3. After applying log and difference to the variable, the trend and seasonality is removed and data becomes clearly stationary in mean. But variance is not constant, noisy data, Thus, data becomes candidate for Arima With arch-garch model.

Now since after applying log and difference our data became stationary let's check the normality of the logged data.

R code:

```
hist(ry)
qqnorm(ry)
qqline(ry)
```

output:

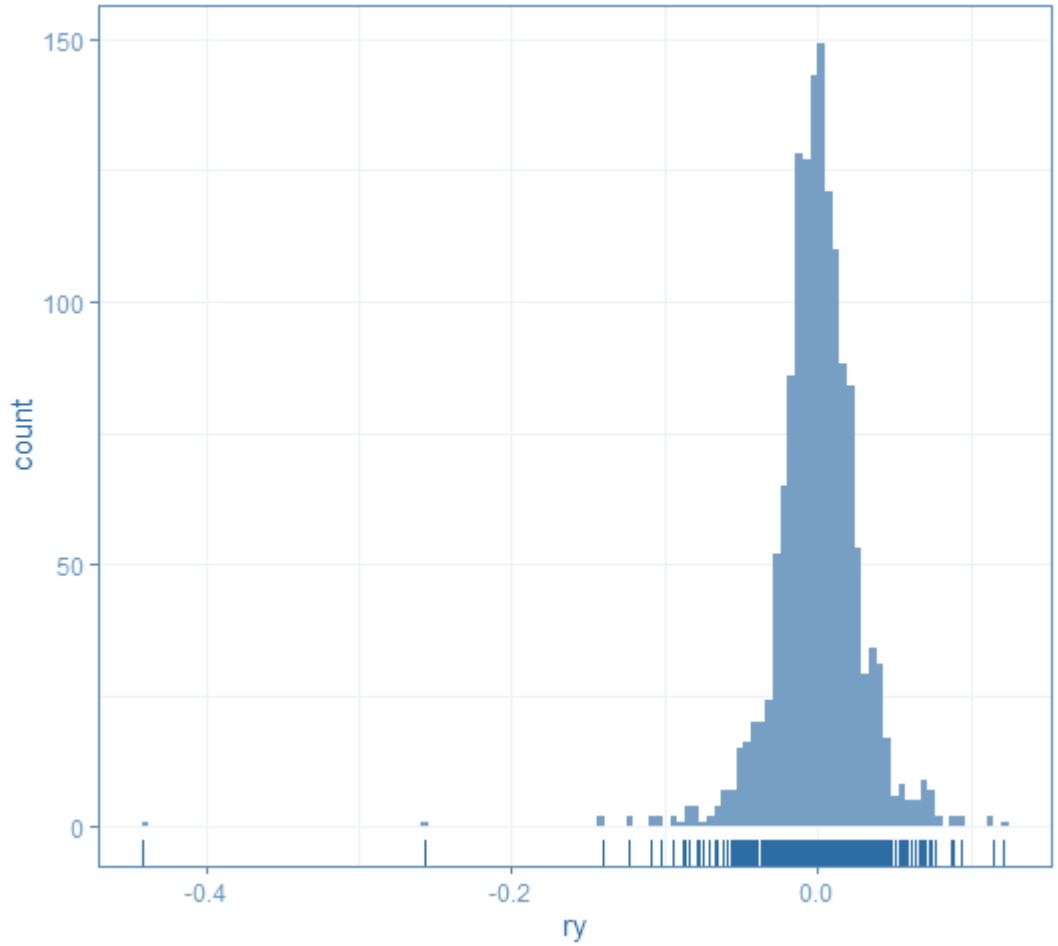


Figure3.4: *Histogram of returns*

From figure 3.4 we can see that our histogram is in a bell shape, thus we can say our data is normally distributed. To make it more clear let's plot the Q-Q plot

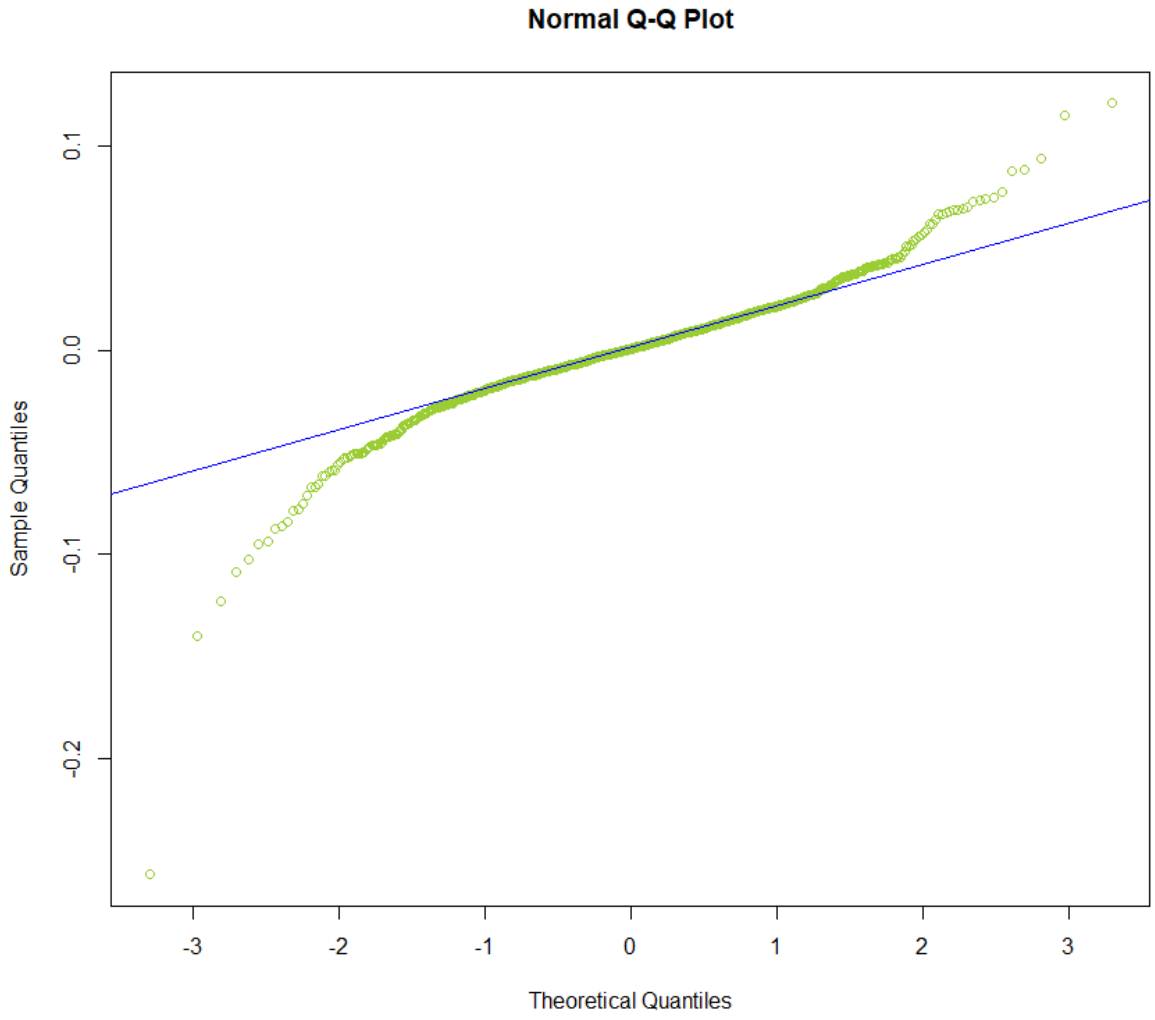


Figure 3.5

From the figure 3.5 clearly our data points follow the straight line. Hence our data becomes Normal.

Now since our data is normally distributed and is stationary, we can fit our model.

3.1.4 Model Fitting

In order to identify the AR order (*p*) and MA order (*q*) and to fit the best fit Arima model we can use the

R code

```
Library(forecast)
auto.arima(ry1)
```

Output:

```
Series: ry1
ARIMA (3,1,3)
ARIMA (3,0,3) with zero mean
Coefficients:
      ar1      ar2      ar3      ma1     ma2     ma3
0.7808  -0.4845  -0.3356  -0.8312  0.5314  0.3479
s.e.  1.3456   1.5760   1.2720   1.3455  1.6236  1.3322

sigma^2 = 0.0007318:  log likelihood = 2211.39
AIC=-4408.78   AICc=-4408.67   BIC=-4374.37
```

As per the R calculation it shows us that the best fitted model is ARIMA (3, 1, 3). That means the current model dependence on the lag, AR lag of 3, difference, d=1 (since we applied one difference) and MA residual lag of 3. i.e., p=3, d=1, q=3.

Now to create a model we use the

R code:

```
modelry=arima(ry, order = c(3, 1, 3)
modelry
tsdiag(modelry)
```

output:

```
Call:
arima(x = ry, order = c(3, 1, 3))

Coefficients:
         ar1      ar2      ar3      ma1      ma2      ma3  intercept
      0.7796  -0.4822  -0.3392  -0.8299   0.5292   0.3509     4e-04
s.e.  1.0413   1.2224   0.9899   1.0424   1.2595   1.0358     9e-04

sigma^2 estimated as 0.0007272:  log likelihood = 2211.52, aic = -4407.04
```

Thus, the equation of the model becomes:

Y=0.0001+ (0.7796)y(t-1)+(-0.4822)y(t-2)+(-0.3392)y(t-3)+(-0.8299)et-1+ (0.5292)et-2+(0.3509)et-3

3.2 Diagnostic Checking

Now in order to check the model we have created is best fit or not, we have to do the Diagnostic check. i.e., we have to check the residuals (the leftover after fitting a model in time series). Residuals are very helpful in checking whether the model has adequately captured the information in the data or not.

The first step is to give a parameter for the residuals. After that we have to check whether the residuals are correlated or not. If there is correlation between the residuals that means there is some information that is left in the residuals which should be used in the forecast. Secondly, we need to check that the residuals have zero mean and finally we have to check whether the residuals are normally distributed

R code:

```
et=residuals(modelry)
 acf(et)

plot.ts(et)

gghistogram(et)

autoplot(modelry)
arimar=arima010$residuals
ggtsdisplay(arimar,lag.max = 20)
```

Output:

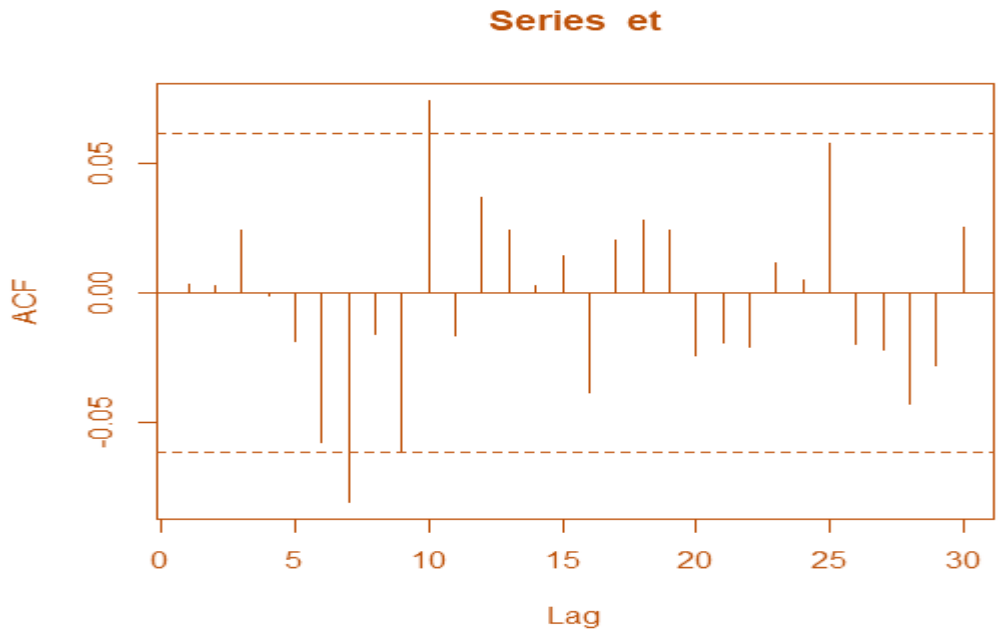


Figure 5: *ACF of residuals*

From the graph it is clear that the first few points do not cross the control limits line. Until then it is independent and hence no autocorrelation is present in between the residuals.

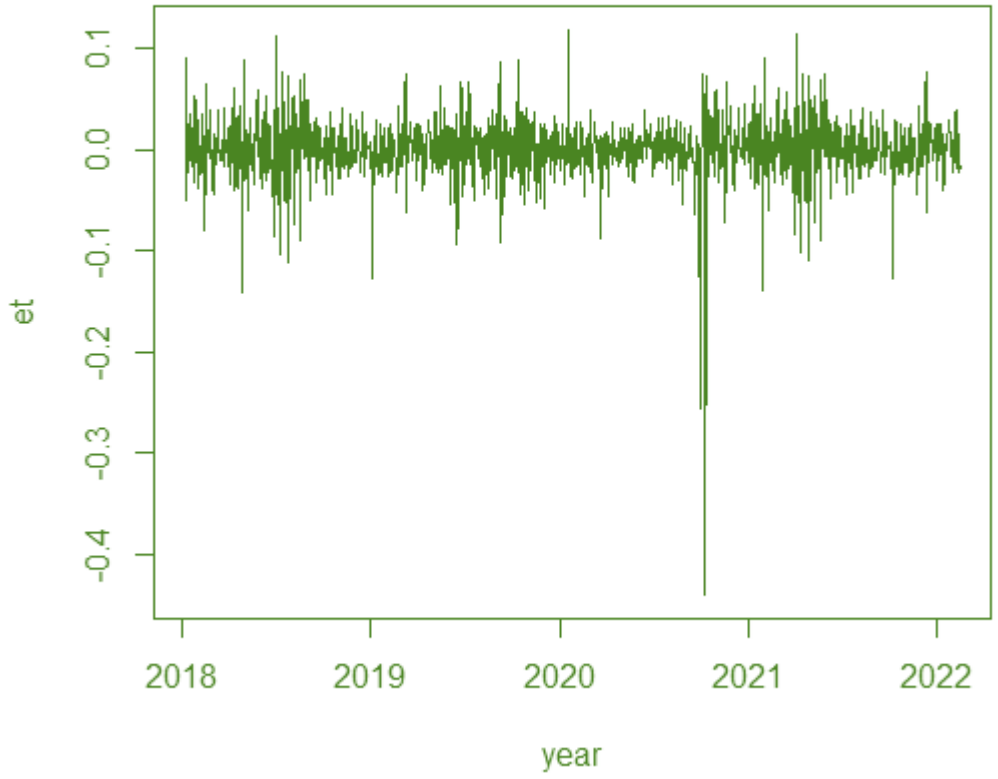


Figure6: *Graph of residuals*

If we put a straight line through zero, we can see that their mean is zero. i.e., there is a constant mean in residuals at zero. But still there is some volatility present which ARIMA model could not cover up.

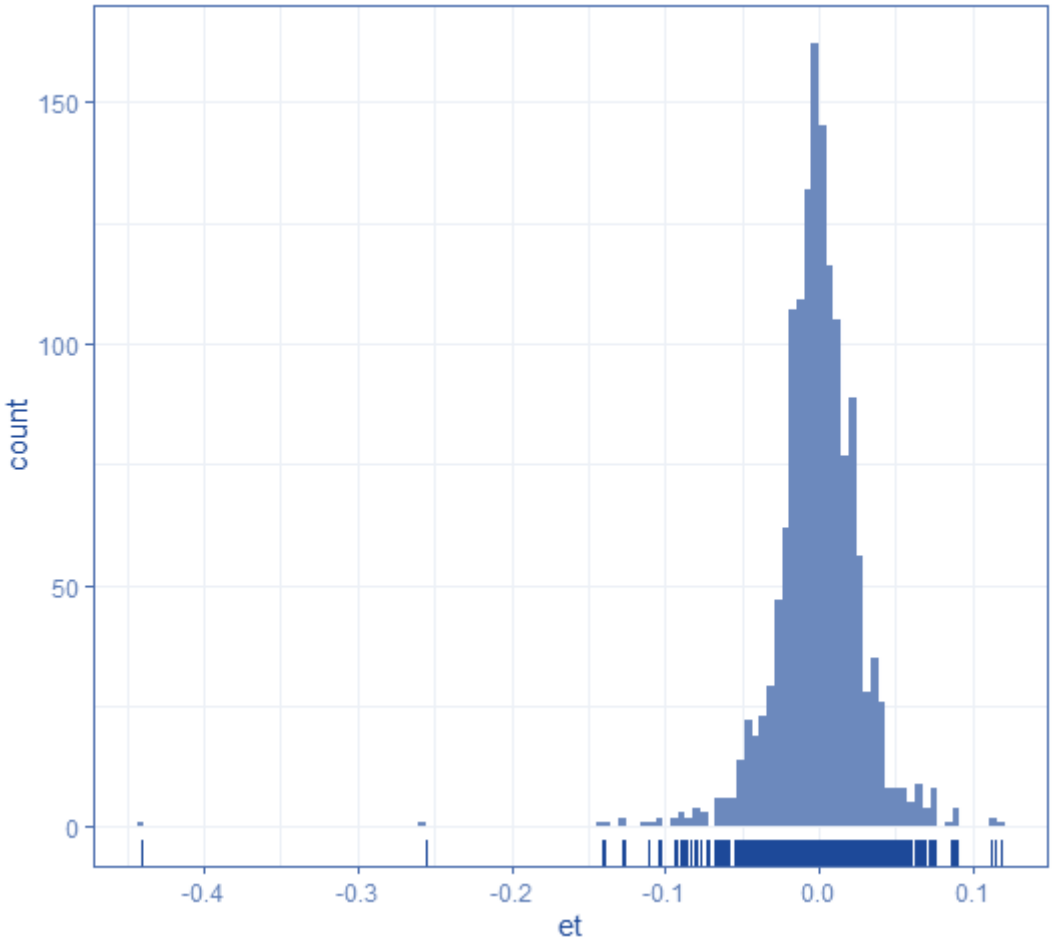


Figure 7: *Histogram*

From the graph it is clear that it is well shaped and it is normally distributed.

From the output we can conclude that the residuals are not correlated, it has a zero mean and it is normally distributed.

There is another method for diagnostic check that is Ljung-Box test.In this test let's take the null hypothesis as residuals follow IID.

R code:

```
Box.test(et,lag = 20, type = c("Box-Pierce","Ljung-Box"),fitdf = 6)
```

Output:

```
Box-Pierce test

data:  et
X-squared = 22.85, df = 14, p-value = 0.0627
```

Here the p value is greater than 0.05 hence we accept the null hypothesis and thus the residuals are independent and is not correlated.

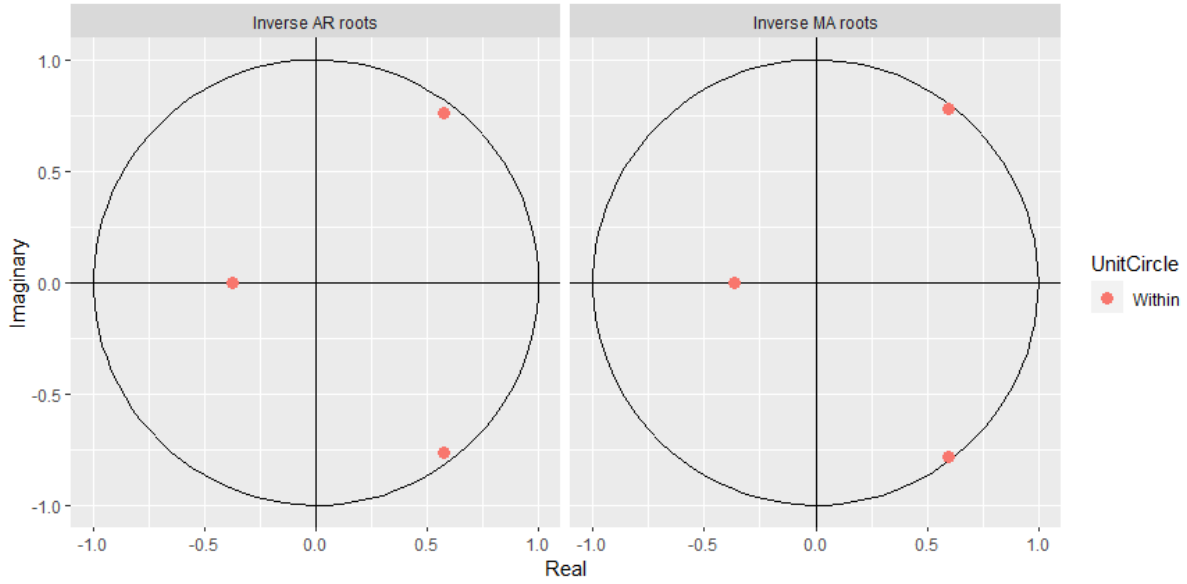


Figure 8: Autoplot

Since there is 3 AR and 3 MA there is 1 root and since the dots are within the circle it shows that the model is stable

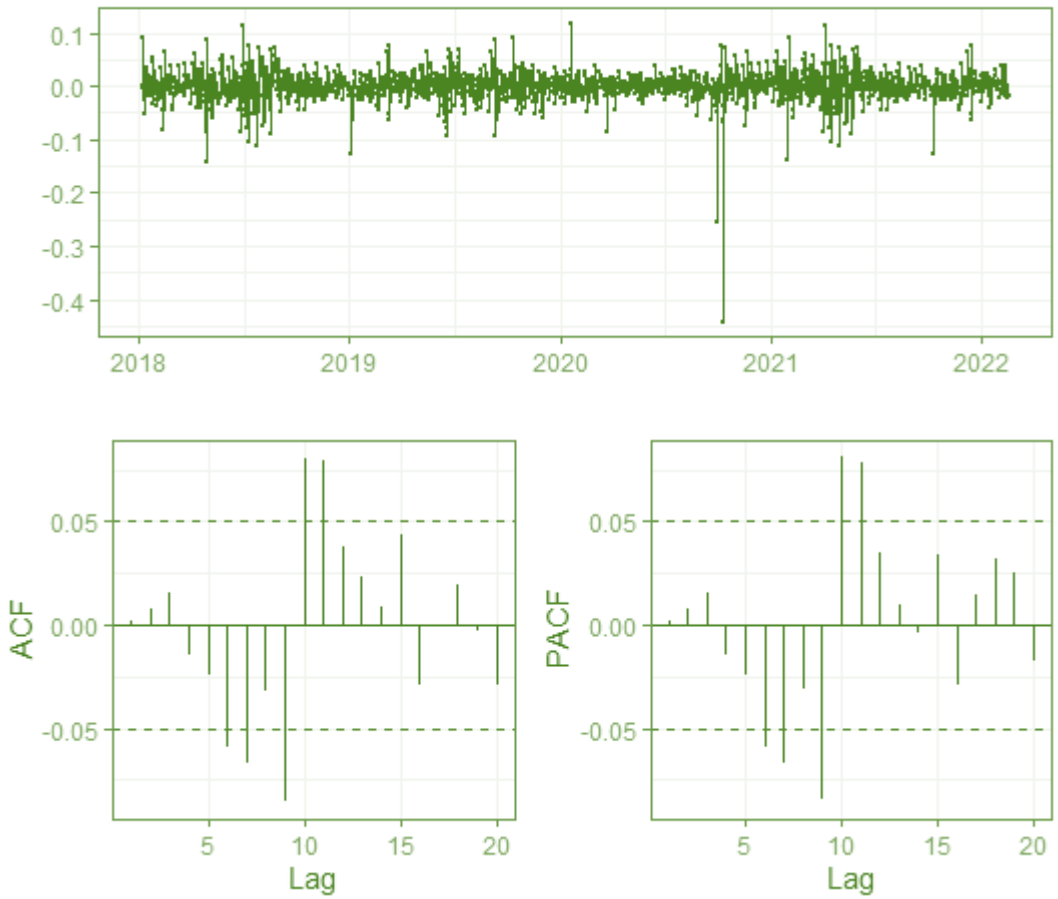


Figure 9

From figure 9 it is clear that there is no more ar and ma effect remaining and arima model is clear but there is still some volatility present.

Since the residuals vary randomly around zero and the spread of the residuals are not the same throughout the plot, we may conclude that the random shocks have mean zero and non-constant variance. Also, from the histogram we can confirm the normality assumption of random shocks. Thus, although the diagnostic checking reveals that the fitted ARIMA (3, 1, 3) model is statistically adequate for modelling the daily opening stock price of Netflix there is still volatility present in the data which this model was not able to cover up. Thus, it is necessary to fit ARCH-GARCH Model.

CHAPTER 4

ARCH AND GARCH MODEL

A change in the variance or volatility over time can cause problems when modelling time series with classical methods like ARIMA. An ARCH is a method that explicitly models the change in variance over time in a time series. The approach of ARCH effect expects the series is stationary. Here our data becomes stationary after applying log. ARCH model is applied to a residual series, et that has mean zero, serially uncorrelated process with non-constant variance. Through diagnostic checking et has mean zero and is uncorrelated. Now to check whether the variance is non constant and if we could apply ARCH model we can use the r-code:

R-code

```
mean=dynlm(et~1)
summary(mean)
```

Output

```
dynlm(formula = et ~ 1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.254637 -0.012675  0.000397  0.014699  0.120562

Coefficients:
             Estimate   Std. Error t value Pr(>|t|)
(Intercept) 2.359e-07  8.498e-04       0        1
```

Residual standard error: 0.02698 on 1007 degrees of freedom

From this output we can analyse that the mean effect is insignificant and hence that our fitted ARIMA model is appropriate. Thus, there is no change in mean value only volatility is left behind in the residuals.

Now to check the ARCH effect,

R code:

```
mean=dynlm(et~1)
summary(mean)
et_sqr=ts(resid(mean)^2)
archeffect=dynlm(et_sqr~L(et_sqr))
summary(archeffect)
```

Output:

```
Time series regression with "ts" data:
Start = 2, End = 1008

Call:
dynlm(formula = et_sqr ~ L(et_sqr))

Residuals:
      Min        1Q    Median        3Q       Max
-0.002884 -0.000653 -0.000519 -0.000132  0.064170

Coefficients:
```

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.688e-04 8.256e-05 8.101 1.58e-15 ***

L(et_sqr) 7.711e-02 3.144e-02 2.453 0.0143 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.002518 on 1005 degrees of freedom

Multiple R-squared: 0.005951, Adjusted R-squared: 0.004962

F-statistic: 6.017 on 1 and 1005 DF, p-value: 0.01434

For estimating the ARCH effect, we find the residuals and then square it up and then create its regression with lag of residual scales, the square of residuals(volatility) is significant hence there is an ARCH effect that is residuals are function of their past. Thus, we go for Arch Model.

R-code

```
garch(ry,grad="numerical",trace=FALSE)
```

Output

Call:

garch(x = ry, grad = "numerical", trace = FALSE)

Coefficient(s):

a0 a1 b1

0.0006464 0.0500001 0.0500000

Here a1 will represent alpha1 value and b1 will represent the beta1 value. Hence, there will be the GARCH order of (1,1). That is alpha1 represents the ARCH (q) order and beta1 represents the GARCH (q) order.

R code:

```
ygarch=ugarchspec(variance.model=list(garchOrder=c(1,1)),mean.model=list(armaOrder=c(3,3)))

    ygarchfit=ugarchfit(ygarch,data=ry)

    ygarchfit

    newsgarch=newsimpact(ygarchfit)


plot(newsgarch$zx,newsgarch$zy,ylab=newsgarch$yexpr,xlab=newsgarch$xexpr,main="News Impact Curve")
```

Output:

```
    *---------------------------------*
    *         GARCH Model Fit         *
    *---------------------------------*


    Conditional Variance Dynamics
    -----------------------------------------
    GARCH Model      : sGARCH(1,1)
    Mean Model       : ARIMA(3,0,3)
    Distribution : norm


    Optimal Parameters
    -------------------------------------------------
```

	Estimate	Std. Error	t value	Pr(>|t|)
mu	0.001308	0.000598	2.1851	0.028878
ar1	1.591356	0.000252	6309.7004	0.000000
ar2	-1.472466	0.000225	-6538.6646	0.000000
ar3	0.448276	0.000097	4634.0590	0.000000
ma1	-1.526878	0.000166	-9184.2356	0.000000
ma2	1.426293	0.000170	8387.2194	0.000000
ma3	-0.353646	0.000051	-6881.5048	0.000000
omega	0.000044	0.000013	3.3321	0.000862
alpha1	0.064768	0.016924	3.8270	0.000130
beta1	0.879271	0.028337	31.0291	0.000000

Robust Standard Errors:

	Estimate	Std. Error	t value	Pr(>|t|)
mu	0.001308	0.000721	1.8142	0.069643
ar1	1.591356	0.072832	21.849584	0.00000
ar2	-1.472466	0.027823	-52.922152	0.00000
ar3	0.448276	0.014258	31.440482	0.00000
ma1	-1.526878	0.033016	-46.246126	0.00000
ma2	1.426293	0.033426	42.670117	0.00000
ma3	-0.353646	0.008387	-42.166834	0.00000
omega	0.000044	0.000020	2.2094	0.027146
alpha1	0.064768	0.024145	2.6824	0.007309
beta1	0.879271	0.028839	30.4888	0.000000

LogLikelihood : 2239.544

Information Criteria

Akaike -4.4816

Bayes -4.4328

Shibata -4.4818

Hannan-Quinn -4.4630

Weighted Ljung-Box Test on Standardized Residuals

---

	statistic	p-value
Lag[1]	7.196	0.007308
Lag[2*(p+q)+(p+q)-1][17]	16.424	0.000000
Lag[4*(p+q)+(p+q)-1][29]	20.881	0.030037

d.o.f=6

H0 : No serial correlation

Weighted Ljung-Box Test on Standardized Squared Residuals

---

	statistic	p-value
Lag[1]	0.01952	0.8889
Lag[2*(p+q)+(p+q)-1][5]	1.18085	0.8176

Lag[4*(p+q)+(p+q)-1][9] 1.80847 0.9266

d.o.f=2

Weighted ARCH LM Tests

	Statistic	Shape	Scale	P-Value
ARCH Lag[3]	0.2448	0.500	2.000	0.6208
ARCH Lag[5]	0.6208	1.440	1.667	0.8475
ARCH Lag[7]	0.8811	2.315	1.543	0.9321

Nyblom stability test

Joint Statistic: 2.5554

Individual Statistics:

mu	0.02252
ar1	0.02374
ar2	0.02892
ar3	0.02622
ma1	0.01673
ma2	0.02982
ma3	0.02498
omega	0.06686
alpha1	0.08107
beta1	0.07689

Asymptotic Critical Values (10%	5%	1%)
Joint Statistic:	2.29	2.54	3.05
Individual Statistic:	0.35	0.47	0.75

Sign Bias Test

---

	t-value	prob sig
Sign Bias	0.7471	0.4552
Negative Sign Bias	0.7757	0.4381
Positive Sign Bias	0.2506	0.8022
Joint Effect	0.8061	0.8480

Adjusted Pearson Goodness-of-Fit Test:

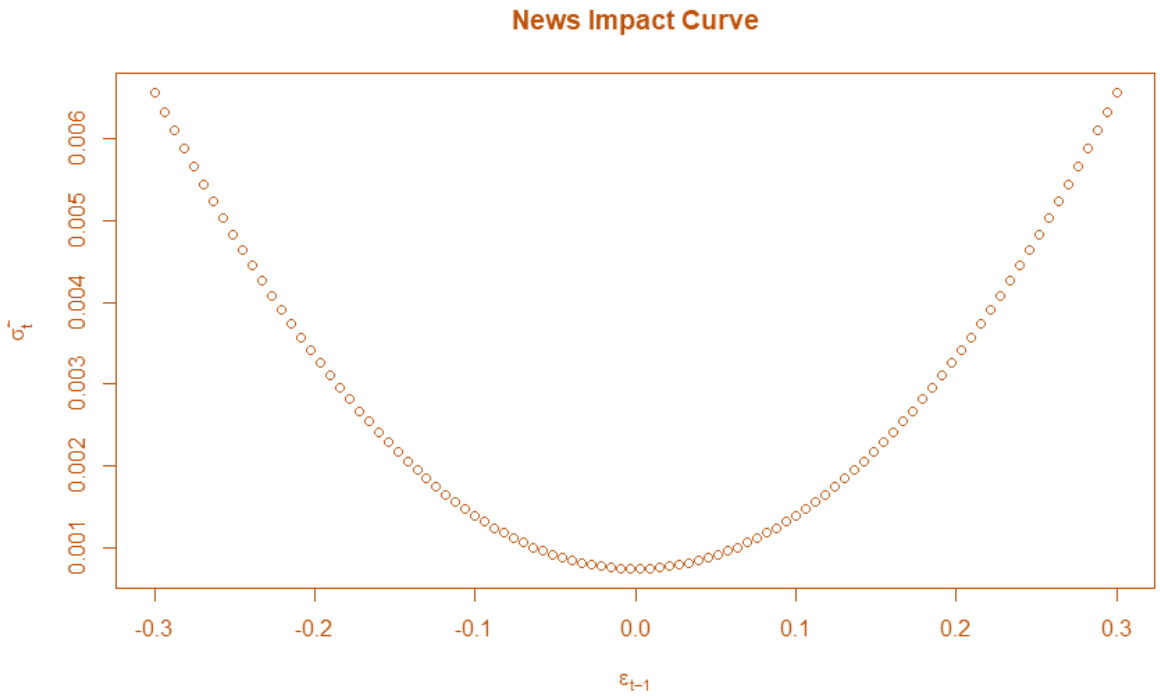
---

	group	statistic	p-value(g-1)
1	20	53.39	4.089e-05
2	30	67.95	5.749e-05
3	40	83.43	4.547e-05
4	50	92.60	1.677e-04

Elapsed time : 2.306941

Here we can see that the fitted model is symmetric GARCH model. That means there is an equal impact of bad news and good news.

Also, here ar1 ma1 omega alpha and beta terms are all significant and also the coefficients of omega, alpha and beta are all positive. Also, the summation of alpha and beta values is less than 1. Hence all the assumptions are proved thus our fitted GARCH model is significant



This graph shows us a symmetric graph.

The estimation results for this model show that the estimated model GARCH(1,1) fit the volatility of ARIMA(3,1,3) .The parameter estimates for the model is

Mu	ar1	ar2	ar3	ma1	ma2	ma3	omega
0.02252	0.02374	0.02892	0.02622	0.01673	0.02982	0.02498	0.06686

Alpha1	beta1
0.08107	0.07689

Thus, the ARIMA-GARCH model for the return series is given by,

$$y_t = 0.02252+0.02374y_{t-1}+0.02892y_{t-2}+0.02622y_{t-3}+0.01673\varepsilon_{t-1}+0.02982\varepsilon_{t-2}+0.02498\varepsilon_{t-3}+\varepsilon_t ,$$

$$\sigma_t{}^2=0.06686+0.08107\varepsilon^2{}_{t-1}+0.07689\sigma^2{}_{t-1}$$

CHAPTER 4

FORECASTING AND CONCLUSION

4.1 Forecasting

One of the most important objectives in the analysis of a time series is to forecast its future values. In forecasting our objective is to produce an optimum forecast that has no error or little error as possible, which leads to the minimum mean square error forecasting.

Since, the best fitted ARIMA (3, 1, 3)-GARCH (1,1) model is statistically adequate. Also, the model satisfies stationarity and invertibility requirements, we may use this model to forecast the daily stock price during the 5th February 2022 to 5th March 2022.

R code:

```
forecastry=forecast(modelry,h=30)

forecastry

plot(forecastry)
```

Output:

Day	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
1009	-3.134688e-03	-0.03769498	0.03142560	-0.05599009	0.04972072
1010	-4.699931e-03	-0.03930393	0.02990407	-0.05762219	0.04822232
1011	-2.107165e-03	-0.03671223	0.03249790	-0.05503105	0.05081672
1012	2.136295e-03	-0.03249927	0.03677186	-0.05083423	0.05510682
1013	4.725076e-03	-0.02994704	0.03939719	-0.04830135	0.05775150
1014	3.817674e-03	-0.03085734	0.03849268	-0.04921318	0.05684853
1015	4.227366e-04	-0.03426424	0.03510971	-0.05262642	0.05347189
1016	-2.664376e-03	-0.03738232	0.03205357	-0.05576089	0.0504321
1017	-3.126241e-03	-0.03785473	0.03160225	-0.05623888	0.0499864
1018	-8.462548e-04	-0.03557650	0.03388399	-0.05396158	0.05226907
1019	2.200937e-03	-0.03254941	0.03695128	-0.05094513	0.05534700

1020	3.633717e-03	-0.03113229	0.03839972	-0.04953630	0.05680373
1021	2.508042e-03	-0.03225819	0.03727427	-0.05066232	0.05567840
1022	-9.390177e-05	-0.03486948	0.03468168	-0.05327856	0.05309076
1023	-2.065484e-03	-0.03685670	0.03272573	-0.05527405	0.05114308
1024	-1.966050e-03	-0.03676039	0.03282829	-0.05517940	0.05124730
1025	-5.534935e-05	-0.03485214	0.03474144	-0.05327244	0.05316174
1026	2.054947e-03	-0.03275349	0.03686338	-0.05117996	0.05528985
1027	2.745035e-03	-0.03206968	0.03755975	-0.05049947	0.05598954
1028	1.617390e-03	-0.03319737	0.03643215	-0.05162719	0.05486197
1029	-3.102034e-04	-0.03513141	0.03451100	-0.05356464	0.05294423
1030	-1.503227e-03	-0.03633184	0.03332539	-0.05476899	0.05176254
1031	-1.121345e-03	-0.03595063	0.03370794	-0.05438814	0.05214545
1032	4.054107e-04	-0.03442622	0.03523705	-0.05286498	0.05367580
1033	1.816133e-03	-0.03302182	0.03665409	-0.05146392	0.05509618
1034	2.050187e-03	-0.03279001	0.03689039	-0.05123330	0.05533368
1035	1.034580e-03	-0.03380594	0.03587510	-0.05224940	0.05431855
1036	-3.484998e-04	-0.03519307	0.03449607	-0.05363868	0.05294168
1037	-1.016383e-03	-0.03586422	0.03383146	-0.05431156	0.05227879
1038	-5.256748e-04	-0.03537358	0.03432223	-0.05382094	0.05276959

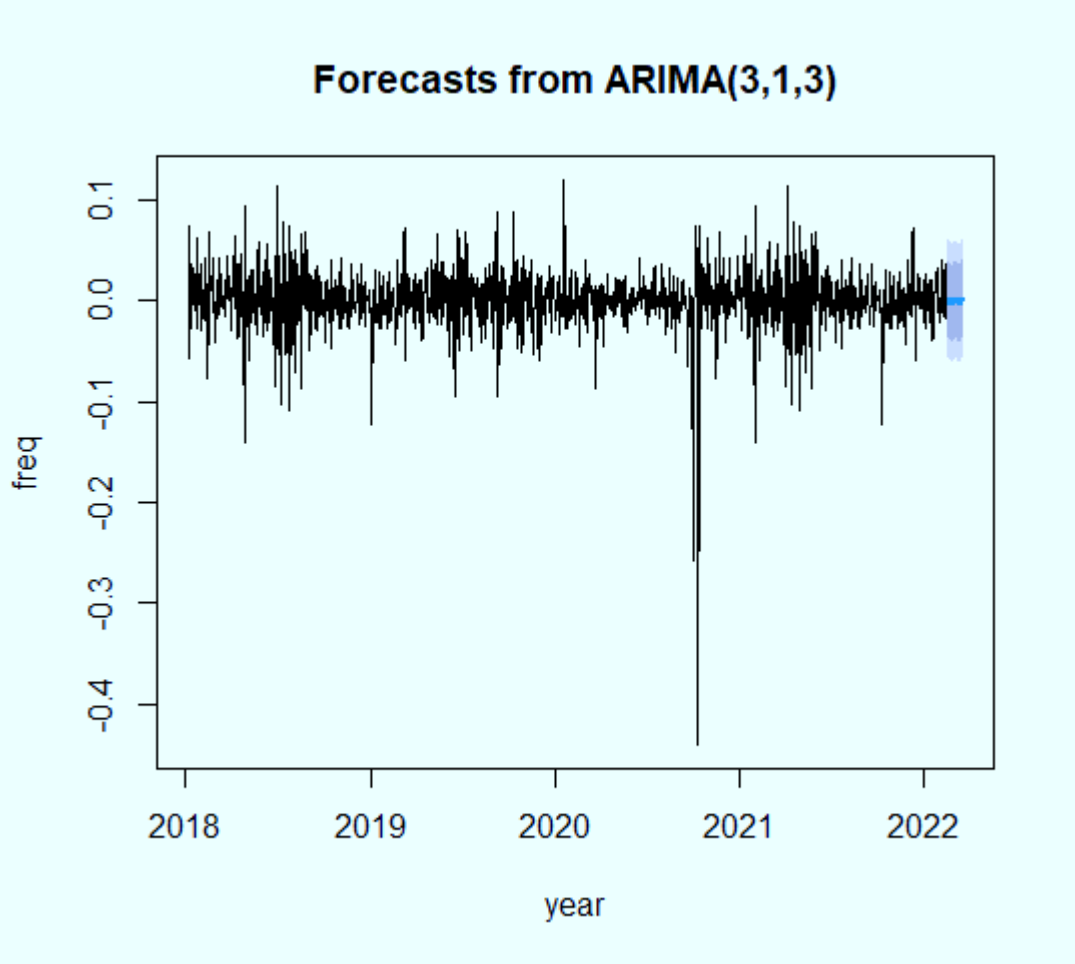


Figure 13: Time Series plot of observed stock price with forecasted values.

4.1.1 ACCURACY

Now we can check the accuracy of our forecasted value using MAPE, ME, MAE, MASE.

R code

Accuracy(modelry)

Output

	ME	RMSE	MAE	MPE	MAPE	MASE
Training set	2.358982e-07	0.02696754	0.01876153	Inf	Inf	0.6906292

	ACF1
	0.000874935

Here mainly we check the accuracy of the forecasted values using MAPE. Our MAPE value tends to infinity i.e., maybe our observed value is zero so when divided it becomes infinite. So, next we check our MAE value which is equal to 0.01876153 which is significant hence our model is accurate.

4.2 CONCLUSION

This project takes up modelling and forecasting of the daily opening stock price of Netflix using ARIMA-GARCH time series model. Our empirical results suggest that ARIMA (3,1,3) models fit the return of stock price series well and GARCH (1,1) model fit the volatility of ARIMA (3,1,3) residuals. Thus, ARIMA (3,1,3)-GARCH (1,1) model is considered as the best model for the daily opening stock price of Netflix and they are capable of predicting the future trend of the price movement.

REFERENCES AND APPENDIX

1)Abraham, B and Ledolter, J. C. (1983): “Statistical Methods for Forecasting”, Wiley, U. S.

2)Box, G. E.P. and Jenkins, G.M. (1976): “Time series analysis, Forecasting and Control”, Holden-Day, San Francisco.

3)Brockwell, P. J. and Davis, R. A. (2002): “Introduction to Time Series and Forecasting”, Springer, U. S.

4)Chatfield, C. (1995): “The analysis of time series an introduction”, Chapman and Hall, New York.

5)Pankratz, A. (1986):” Forecasting with Univariate Box-Jenkins Models”, john Wiley and sons, New York.

6)Wei, W. W. S. (2006): “Time Series Analysis: Univariate and Multivariate Methods”, Pearson Education, New York.

7)www.kaggle.com - https://www.kaggle.com/datasets/jainilcoder/netflix-stock-price-prediction