Implementation of Time Series Approaches to Financial Data

Noshin Nawar Sadat

Supervisor: Dr. Mahbub Majumdar

A thesis submitted in partial fulfillment for the
degree of Bachelors of Computer Science & Engineering

in the

Department of Computer Science & Engineering
BRAC University

August 2016

Declaration

This is to certify that this thesis report, titled "**Implementation of Time Series Approach to Financial Data**" is submitted by Noshin Nawar Sadat (ID:12101017) to the Department of Computer Science and Engineering, School of Engineering and Computer Science, BRAC University, in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering. I, hereby, declare that this thesis and the work presented in it are my own and it has not been submitted to any other University or Institute for award of any other degree or diploma. Every work that has been used as reference for this thesis has been cited properly.

Author

Noshin Nawar Sadat

ID: 12101017

Supervisor

Dr. Mahbub Majumdar

Professor

Department of CSE

BRAC University, Dhaka

Acknowledgements

The completion of this thesis would not have been possible without the support, help and encouragement of several people. First of all, I would like to express my sincere gratitude to Dr. Mahbub Alam Majumdar, my supervisor, for his continuous support and guidance. He has been a constant source of encouragement and enthusiasm throughout the time of this project.

I am also very grateful to Dr. Md. Haider Ali, Professor and Chairperson, Department of Computer Science and Engineering, BRAC University. His classes as well as his support for all his students' work is a source of inspiration for me and everyone else in the department.

I would also like to thank all of my teachers, starting from kindergarten to the University. If it weren't for their teachings, I might not have been able to accomplish whatever I have accomplished till now.

My heartfelt gratitude goes to Ipshita Bonhi Upoma, my dear friend, who listened to me patiently and kept me motivated by warning me about the deadline almost everyday.

I would also like to thank Nuzhat Ashraf Mahsa and Faiza Nuzhat Joyee, for their understanding and support. I have to thank all my school friends as well. Our regular chats over phone helped me to get rid of all my stress within minutes.

Finally, my deepest gratitude goes to my parents and my sister, for their unconditional sacrifices, love and support throughout my life.

ABSTRACT

We study a time series approach to financial data, specifically the ARIMA models, and build a web based platform for stock market enthusiasts to analyze time series of stock market returns data and to fit ARIMA models to the series to forecast future returns. This system also acts as an informative tool by providing helpful instructions to the users regarding the analysis and model-fitting procedure. It uses R to perform the statistical computations.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The analysis of financial data for the prediction of future stock returns has always been an important field of research. According to [17], stock prices may not only signal future changes in the economy of a country but also have direct effects on the economic activities of the country. Although not always effective, stock price movements are considered as useful indicators of the business cycle and they also affect the household consumptions and corporate investments in a country to some extent. Hence, being able to make successful predictions of the future stock prices not only allows investors to make significant amount of profits by preventing loss but also helps the government to take appropriate measures to prepare for possible economic downturns in the future. For that reason, numerous theories and methods are being devised to make accurate predictions of the stock market by the analysis of financial data. In our project, we focused on creating a platform for the prediction of stock returns of companies using time series approach. We built a web-based system where users will be able to analyze stock returns data of forty companies listed in NASDAQ and NYSE and then use that analysis information to fit ARIMA models to those data. All the analysis and model-fitting tasks were done by using the R software. This is just the initiative step towards building a more comprehensive and sophisticated system where users will be able to build their own models based on any time series approach and make predictions of future stock returns in order to take appropriate measures.

1.1 Time Series Approach

In the Time Series Approach of stock price analysis and forecasting, it is considered that the stock price incorporates all important and available information of the stock. By figuring out the underlying structure and function that produced the past observations of the price, time series analysis of stock prices aims to forecast the future prices or trends.

1.2 Motivation

Not all investors are capable of analyzing stock market data on their own for making their trading decisions. They have to depend on the analysis of others instead. Even if they wish to do it by themselves, they face the different problems of collecting the data, acquiring analysis tools, leaning to use those tools etc. Moreover, those who are involved in research in this field, might not be able to access a computer with statistical tools installed all the time. In such cases, having access to an online statistical system to

perform instant analysis on authentic data would be preferable.

We plan to create such a platform for any stock market enthusiast, which will

- Allow them to perform various methods of time series analysis on stock returns data
- Provide authentic data
- Be usable
- Not require any installation
- Be accessible anywhere
- Allow them to save the outputs of their analysis for future references
- Provide helpful information regarding all the tests we allow them to perform and how to interpret those tests

With that final goal in mind, we have initiated this project. Currently, we have created a platform, where users can perform analysis on the time series of daily stock returns data of companies, fit ARIMA models of any order to that data and make forecasts for the next 10 days. This will eventually be evolved and perfected to achieve our final goal.

1.3 Thesis Outline

The outline of this paper is as follows:

- **Chapter 2** gives a basic introduction to the different methods that are used for stock market prediction and it also discusses some papers that have dealt with stock market prediction using time series analysis.
- **Chapter 3** introduces and explains the concepts of ARIMA modeling which is the approach of time series that we have focused on in our work.
- **Chapter 4** describes our implementation work which is a website for stock market prediction using time series approach.
- **Chapter 5** explains the limitations of our project and also provides directions for further research.
- **Chapter 6** summarizes the whole paper and gives conclusive remarks regarding the project.

Chapter 2

Stock Market Prediction Methods

Despite many negative views regarding stock market prediction, people still try to come up with different ways to forecast future price movements in the stock market. Thus, we have now numerous methods of stock market prediction, starting from simple fundamental analysis to more complicated statistical analysis methods. With the advent of computers into the stock market prediction scene, analysis and development of prediction methods have become much easier. This chapter discusses about different such methods of analysis, which can be used singularly or in combination to predict stock price movements.

2.1 Efficient Market Hypothesis

The Efficient Market Hypothesis or EMH, as developed by [11], suggests that it is impossible to beat the stock market as it is an efficient market, where the stock prices reflect all the available and related information. Whenever new information comes up, it gets immediately spread by the news and the stock price gets adjusted to it in no time. Therefore, stocks are always traded at their fair value and it is not possible to make predictions of trends in the market or to identify undervalued stocks. So, the only possible way of making profit in the share business is to buy risky investments. However, this theory has one big flaw. It assumes that the stock market is efficient. However, for the stock market to be efficient, the following criteria have to be met [13]:

- All investors should be able to access high-speed and advanced systems of stock price analysis.
- There should be a stock price analysis method which is universally accepted as correct.
- All the investors in the market should be rational decision makers. Their decisions should not be influenced by their emotions.
- The investors would not wish to make any higher profit than the other investors because that would not be possible.

None of the above mentioned conditions can be met by the present stock market. Hence, we cannot call it efficient. The stock market often shows trends in the price movements and so it is possible to make predictions of future trends based on the past prices, if modeled correctly.

2.2 Random Walk Hypothesis

[14] stated that stock prices cannot be predicted accurately by using price history. He termed stock price movements as a statistical process, called the random walk. According to his theory, the deviation from the mean of each observation is purely random and therefore, unpredictable.

However, several economists and professors of finance have conducted a number of tests and studies which reportedly claim that some sort of trend does exist in the stock market and hence stock market can be predicted to some degree.

2.3 Typical Stock Prediction Methods

The prediction methods of stock market can be divided into two main categories depending on how the share prices are evaluated. These complementary methods are often used by themselves or in combination. They are as follows:

1. **Fundamental Analysis**
   Fundamental analysis is based on the assumption that stock prices in the stock market do not represent the actual real value of the stock. Fundamental analysis asserts that the correct value of the stock can be found out by analyzing the fundamentals of the company. The fundamentals include anything that is related to the economy of the company. It is claimed by the fundamentalists that by buying the undervalued stocks and holding on to them until the market realizes its mistake and changes the prices of the stocks to their actual prices, an investor can make profit.

   The different fundamental factors can be divided into two groups and the analysis methods of these two groups are termed as:

   i) **Quantitative Analysis**
   The quantitative factors are all the factors that can be expressed in numerical terms. This type of analysis involves delving into the financial statements to learn about the companys revenues, assets, liabilities, expenses and all other financial aspects.

   ii) **Qualitative Analysis**
   Qualitative factors are the intangible, non-measurable aspects of the company, such as, the companys business model, competitive advantage, management, competition, customer, market share, government regulations etc. In other words, it involves the analysis of the company itself, the industry in which the company specializes in as well as the economic condition of the country in which the company operates.

   The intrinsic value of the stock is determined by performing these two analysis methods. If the measured intrinsic value turns out to be greater than the actual market value, then the stock is bought. If it is the same as the market price, then it is held. And if it is lower than the market price, then it is sold.

2. **Technical Analysis** Technical analysis is based on the following three principles:

- **Market Discounts Everything**
  The price of a stock reflects all the relevant and available information in the market. Technical analysts believe that all the factors that could affect the company, such as, the company's fundamental factors, the economic factors as well as the market psychology, are all accounted for in the price of that company's stock and hence, there is no need to analyze them separately.
- **Price Moves in Trends**
  The price movements follow a trend and when such trend has been established, the likelihood of the future stock prices to be in that same direction increases.
- **History Tends to Repeat Itself**
  The pattern of price movements in the past tends to repeat itself in the present as market participants have a tendency to react in a consistent manner to similar stimuli all the time.

Technical analysis (or charting) only focuses on the price movements in the market. It involves identifying patterns of price and volume movement in the market by using charts and other tools and using those patterns to predict future activities in the market. It does not care about the undervalued stocks and it does not try to find the intrinsic values of any stock.

There are two types of tools for technical analysis:

a) **Charts**
   Charts are just graphical representations of stock prices over a set time frame. They could vary in time scale or price scale. Depending on the information to be retrieved, the charts that are mostly used are line chart, bar chart, candlestick chart and points and figures chart.

b) **Indicators & Oscillators**
   Indicators use price and volume information to measure flow of money, momentum and trends in the stock market. Indicators are used to either form buy or sell signals, or to confirm price movement. They are also of two types leading and lagging. Leading indicators help predict future price by preceding price movements. Lagging indicators follow price movements and work as a tool of confirmation. Some indicators are constructed in such a way that they fall within a bound range. These are called oscillators. Crossovers and divergence in the indicators are used to form buy or sell decisions. Some popular indicators are Accumulation/Distribution Line, Average Directional Index (ADX), Aroon, Aroon Oscillator, Moving Average Convergence Divergence (MACD), Relative Strength Index (RSI), On Balance Volume (OBV) and Stochastic Oscillator.

Time Series Analysis

Time Series Analysis may be considered as a mathematical version of technical analysis which uses statistical tools to extract meaningful information from a given stock market price series and makes predictions on the basis of those information. However, getting absolutely accurate forecasts of stock prices using time series analysis only is not an easy route.

Machine Learning

Various machine learning algorithms are being applied for stock market prediction. Notable among them are Support Vector Machines, Linear regression, Online Learning, Expert Weighting and Prediction using Decision Stumps. These machine learning algorithms are applied basing on the same assumption as that of technical analysis. It is assumed that all the prices of the stocks have all the relevant information embedded in them [18]. Using machine learning techniques alone, however, will not be able to give accurate results and so a hybrid of several algorithms or a hybrid of algorithms and some other analysis technique can be used. Moreover, the same technique or algorithm might not work for every company's stock price prediction.

2.4 Literature Review

In order to complete this project, a good understanding of Time Series Analysis was required. To understand the basic concepts of time series analysis, the books of [9] and [8] were very helpful. The different topics of time series analysis were explained in great detail with easy to understand examples in these books. Moreover, the lecture note by [20] and the online learning materials provided by PennState Eberly College of Science on [2] were also very useful in our endeavor to understand time series. In order to learn R programming, the online resources of [4] and [5] were of great help.

We went through several papers that dealt with stock market prediction using time series models so that we could understand the method better. One of the papers was written by [7]. They used the closing price data of Nokia Stock Index and Zenith Bank Stock Index to build separate ARIMA models for the two companies. The ARIMA models with smaller BIC and S.E. of regression, higher adjusted R^2 were chosen as the best models. Another criterion for choice was that the residuals of the models would be white noise. They found that their built models were satisfactorily providing short-term predictions.

[12] used Box-Jenkins method to fit ARIMA models to the stock closing prices of AAPL, MSFT, COKE, KR, WINN, ASML, AATI and PEP. They chose to use the closing prices of the past ten years of the companies or since the year the companies went public. The data was collected from Yahoo Finance. After analysis, almost all their models were AR(1) models either in differenced or undifferenced form. They had used the stock price data of eight companies specializing in four different industries to find if there were any similarities between the industries. It was found that the stocks of the same industry did not behave in a similar manner.

A study on the effectiveness of ARIMA models to predict future prices of fifty-six stocks from seven sectors of India was conducted by [15]. For their work, they used the past twenty-three months data. To choose the best model, they used the value of AICc. All their built models were able to predict stock prices with above 85% accuracy.

Another paper attempted to combine traditional time series analysis techniques with information from the Google trend website and the Yahoo finance website to predict weekly changes in stock prices [21]. They collected important news related to a particular stock over a five year span and they used the Google trend index values of this stock to measure the magnitude of these events. They found significant correlation between the

values of the important news/events and the weekly stock prices. They collected weekly stock prices of AAPL from Yahoo Finance website and extracted important news related to AAPL stock by using the Key Developments feature under the Events tab in the Yahoo Finance website. Each piece of news was then analyzed to give a positive or negative value depending on their influence. The weekly search index of the term AAPL were extracted from Google trend. The starting point of the news data and the Google trend data were set to one month before the stock price data in order to find the relation between the news of one time to the prices of stock at a later time. To analyze the historical stock prices, they performed ARIMA time series analysis after first degree differencing of the square root of the raw data. It was found by plotting the autocorrelation function and partial autocorrelation function that the transformed stock prices essentially followed an ARIMA(0,1,0) process.

[16] proposed a hybrid Support Vector Machine and ARIMA model for stock price forecasting. They used daily closing prices of the fifty days (from October 21, 2002 to December 31, 2002) of ten stocks for their research. They used the closing prices in the month of January 2003 as the validation set. The closing prices of February 2003 were used as testing dataset. They tried making one step ahead forecasts of the hybrid model as well as SVM and ARIMA models separately. It turned out that the hybrid model had outperformed the other two models.

[19] collected 50 randomly selected stocks from Yahoo Finance website and applied time series decomposition (TSD), Holts exponential smoothing (HES), Winters exponential smoothing (WES), Box-Jenkins (B/J) methodology, and neural networks (NN) on the dataset to analyze them and predict future prices. To use in the NN model, they divided the dataset into three groups training set, validation set and testing set. Instead of using the data directly, they used normalized data, which reduced the errors. They used back propagation algorithm for training their system. Their models had fit the data with R^2 almost equal to 0.995.

[10] created a system to forecast movements in the stock market in a given day by using time series analysis on the S&P 500 values. They also performed market sentiment analysis on data collected from Twitter to find out whether the addition of it increased the accuracy of the prediction. They collected the S&P 500 values from Yahoo Finance and the Twitter data from Twitter Census: Stock Tweets dataset from Infochimps. The latter included around 2.3 million stock tweets. This dataset was then modified for the purpose of their work. They had three labels for the S&P index movement up, down, and same. To predict the S&P movements, they used five different attributes. To analyze the sentiments in the tweet dataset, they used Naive Bayes Classifier. The sentiments were labeled as up, down or same. After incorporating the sentiment analysis results with the time series analysis results, they found that the accuracy had improved.

From the above discussion, it is clear that although time series analysis alone is a good candidate for stock price forecasting, to get a more accurate result, other factors and techniques should also be incorporated.

Chapter 3

Summary of Theory

In this chapter, we discuss about the time series analysis approach that we have used in our project. We try to summarize what we have learned from the books, lecture notes and other online materials mentioned in Section 2.4 of this paper. With a view to explaining the topics, we have used the dataset on the annual diameter of the hem of womens skirts from 1866 to 1911 provided by [1]. We used the R software as our tool to perform the different analysis on the dataset. We shall discuss about R in detail in Chapter 4.

3.1 Time Series Analysis

Time series is a collection of observations made sequentially over a time interval. Time series data are being generated every day in different fields of application, such as,

- **In Finance:** daily closing prices of stock, daily exchange rates, etc.
- **In Economics:** monthly total exports, monthly data on unemployment, etc.
- **In Physical Sciences:** daily rainfall, monthly average air temperature, etc.
- **In Marketing:** annual or monthly average sales figures, etc.
- **In Demographic Studies:** monthly or annual population of a city, etc.
- **In Medicine:** Brain wave activity during ECG, etc.
- **In Process Control:** Color property of batches of product, etc.

A time series can either be continuous or discrete. It is said to be continuous if it consists of observations taken continuously in time. If the observations are taken in specific, equal time intervals, then it is said to be discrete. In this paper, we will be working with discrete time series.

In the time series, the distance between any two consecutive time points must be the same and each time point must have at most one observation. That is, if the series is an observation of monthly data, then it must have the observations of every month; and for each month, only one observation should be taken.

The following is an example of a time series plot of the annual diameter of the hem of

womens skirts:

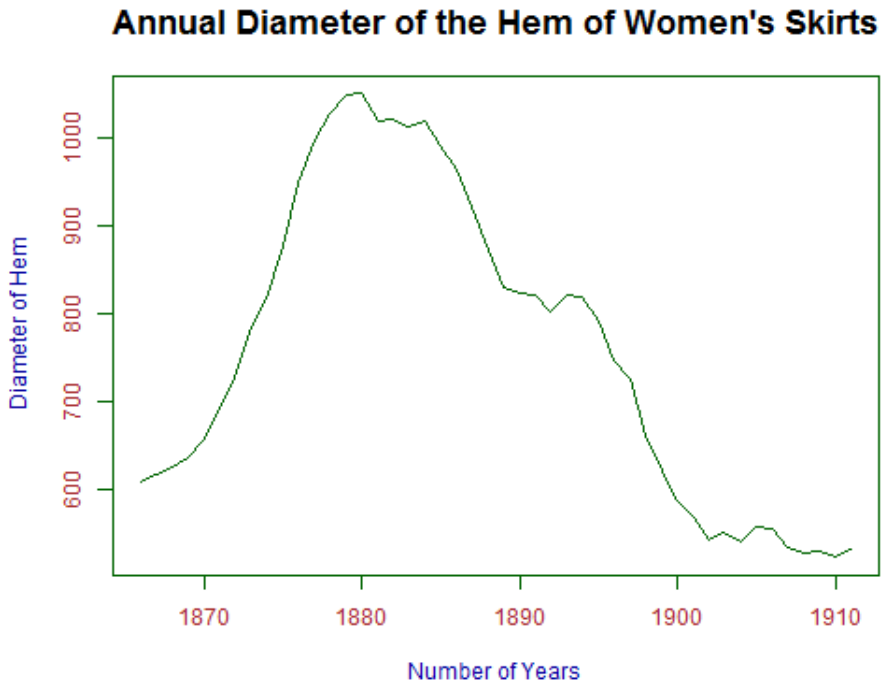

Figure 3.1: Time Series Plot of Annual Diameter of the Hem of Women's Skirts

A **stochastic process** is a sequence of random variables, represented as follows:

$$\{Y_t : t = 0, \pm 1, \pm 2, \}$$

Stochastic processes are used to model observed time series.

Time Series Analysis consists of various techniques of analyzing the time series data with the aim of extracting significant statistics and other important features of data, usually in order to make forecasts of future values based on the past observations. Therefore, during time series analysis, the order of the observations must be maintained. Otherwise, the very meaning of the data would change.

3.2 Some Basic Concepts

3.2.1 Mean Function

If $\{Y_t : t = 0, \pm 1, \pm 2, ...\}$ is a time series, then its mean function is the expected value of the observation at time t,

$$\mu_t = E(Y_t) \qquad \text{for} \qquad t = 0, \pm 1, \pm 2, \pm 3, ... \tag{3.1}$$

3.2.2 Autocovariance Function

If t and s are two time points in the time series, then the covariance of the observations of those two points is given by,

$$\begin{aligned} Cov(Y_t, Y_s) &= E[(Y_t - \mu_t)(Y_s - \mu_s)] && (3.2) \\ &= E[Y_t Y_s] - \mu_t \mu_s && (3.3) \end{aligned}$$

Then the autocovariance function, $\gamma_{t,s}$ is the sequence,

$$\gamma_{t,s} = Cov(Y_t, Y_s) \qquad \text{for} \qquad t, s = 0, \pm 1, \pm 2, \pm 3, ... \tag{3.4}$$

3.2.3 Autocorrelation Function (ACF)

The correlation between two observations at two time points t and s is given by,

$$\begin{aligned} Corr(Y_t, Y_s) &= \frac{Cov(Y_t, Y_s)}{\sqrt{Var(Y_t) Var(Y_s)}} && (3.5) \\ &= \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t} \gamma_{s,s}}} && (3.6) \end{aligned}$$

Then the autocorrelation function $\rho_{t,s}$ is,

$$\rho_{t,s} = Corr(Y_t, Y_s) \qquad \text{for} \qquad t, s = 0, \pm 1, \pm 2, \pm 3, ... \tag{3.7}$$

The value of $\rho_{t,s}$ is always between -1 and $+1$. When its value is close to ± 1, it means that the two observations have strong linear dependence, whereas a value closer to 0 indicates a weak linear dependence. When $\rho_{t,s} = 0$, the two observations are said to be uncorrelated.

3.2.4 Random Walk

If e_1,e_2,.....e_t is a sequence of independent and identically distributed random variables with mean 0 and variance σ_e^2, then the observation at time $t = 1$ will be,

$$Y_1 = e_1$$

At time, t=2,

$$Y_2 = e_1 + e_2$$

This will continue on, until $t = t$, where,

$$Y_t = e_1 + e_2 + ... + e_t$$

Thus the observation at time $t = t$ can be expressed as follows:

$$Y_t = Y_{t-1} + e_t \tag{3.8}$$

Now, the mean of Y_t is given by,

$$\begin{aligned} \mu_t &= E(Y_t) \\ &= E(e_1) + E(e_2) + ... + E(e_t) \\ &= 0 + 0 + ... + 0 \\ &= 0 && (3.9) \end{aligned}$$

And the variance is given by,

$$\begin{aligned} Var(Y_t) &= Var(e_1) + Var(e_2) + ... + Var(e_t) \\ &= \sigma_e^2 + \sigma_e^2 + ... + \sigma_e^2 \\ &= t\sigma_e^2 \end{aligned} \tag{3.10}$$

The covariance between two time point t and s, where $s - t = k$ is,

$$\begin{aligned} \gamma_{t,s} &= Cov(Y_t, Y_s) \\ &= Cov(e_1 + e_2 + ... + e_t, e_1 + e_2 + ... + e_t + e_{t+1} + ... + e_s) \\ &= Cov(e_1, e_1) + Cov(e_1, e_2) + ... + Cov(e_t, e_1) + ... + Cov(e_t, e_t) + ... + Cov(e_t, e_s) \\ &= Var(e_1) + 0 + ... + 0 + ... + Var(e_t) + ... + 0 \\ &= \sigma_e^2 + 0 + ... + 0 + ... + \sigma_e^2 + ... + 0 \\ &= t\sigma_e^2 \end{aligned} \tag{3.11}$$

Hence, the autocovariance function for the process is,

$$\gamma_{t,s} = t\sigma_e^2 \qquad \text{for} \qquad 1 \leqslant t \leqslant s$$

The autocorrelation function of the process can then be expressed as,

$$\begin{aligned} \rho_{t,s} &= \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}} \\ &= \sqrt{\frac{t}{s}} \qquad \text{for} \qquad 1 \leqslant t \leqslant s \end{aligned} \tag{3.12}$$

From (3.12), it is clear that the more the lag $k = s - t$ increases, the more the value of $\rho_{t,s}$ decreases, which indicates that with the increase in lag, the correlation between two observations reduces.

3.2.5 Simple Moving Average (SMA)

The arithmetic moving average which is calculated by adding the observations of a time series over a certain number of time points and then dividing the added result by that number of time points is called simple moving average or SMA.The total number of time points that are taken into account can be varied if needed.

Say, $Y_{-(n-1)}, ...Y_{-3}, Y_{-2}, Y_{-1}, Y_0$ is a time series. If we wish to calculate its simple moving average for n time points at t=0, then,

$$SMA = \frac{Y_{-(n-1)} + ... + Y_{-1} + Y_0}{n} \tag{3.13}$$

Simple moving average helps to smooth out the series by filtering out the noise and thus helps to indicate trends in the series. Often, a weighted moving average is used instead of simple moving average. The weights are given according to the requirement of the analysis.

If a simple moving average of order 12 is done on the time series of annual diameter of hem of women's skirts, we will get the output as shown in Figure 3.2.

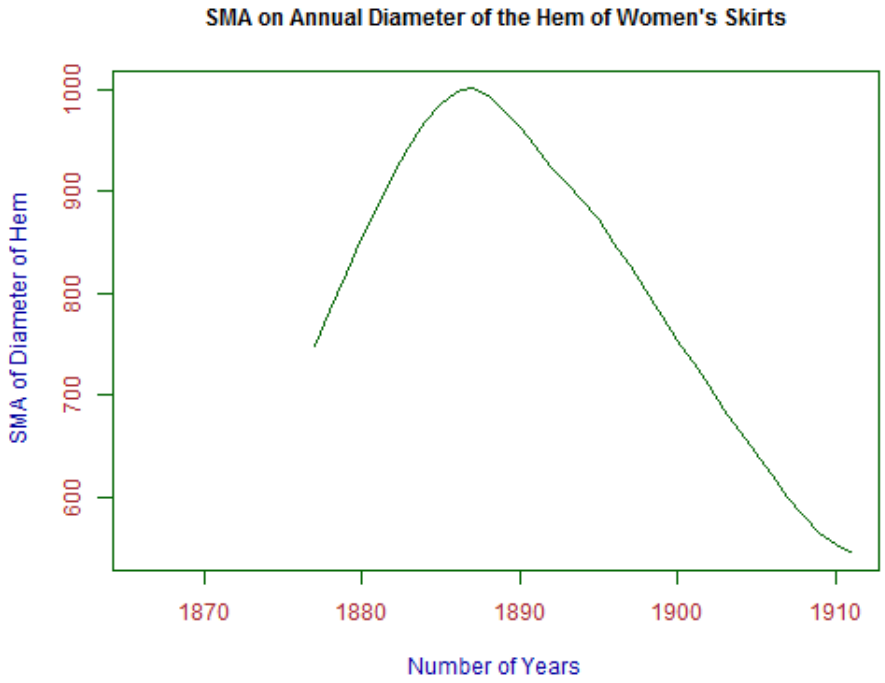

Figure 3.2: Time Series Plot of Simple Moving Average of order 12

It would not be an easy task to see trends visually in most time series data, such as the data of daily stock prices, as there would be a lot more fluctuations in them. Hence, smoothing out the time series data, using the simple moving average method, as we did in Figure 3.2, will help us to see the trends clearly.

3.2.6 Stationarity

Say, $\{Y_t\}$ is a time series. If the joint probability distribution of $Y_{t_1}, Y_{t_2}....., Y_{t_n}$ is same as the joint probability distribution of $Y_{t_1-k}, Y_{t_2-k}....., Y_{t_n-k}$ for all sets of time points $t_1, t_2.....t_n$ and lag k, then $\{Y_t\}$ is said to be a **strictly stationary process**. Shifting the time origin by k does not affect the joint probability distribution. This implies that the joint distribution is dependent on the intervals between $t_1, t_2.....t_n$.

If $n = 1$, the univariate distribution of Y_t will be the same for all t, and both the mean function and the variance will remain constant,

$$\begin{aligned} \mu_t &= \mu \\ Var(Y_t) &= \sigma^2 \end{aligned}$$

If $n = 2$, the bivariate distribution of Y_{t_1} and Y_{t_2} will be the same as that of Y_{t_1-k} and Y_{t_2-k}. Therefore,

$$\begin{aligned} Cov(Y_{t_1}, Y_{t_2}) &= Cov(Y_{t_1-k}, Y_{t_2-k}) \\ \gamma_{t_1,t_2} &= Cov(Y_{t_1-k}, Y_{t_2-k}) \end{aligned}$$

Replacing k by t_1 and k by t_2,

$$\begin{aligned}\gamma_{t_1,t_2} &= Cov(Y_0, Y_{t_2-t_1}) \\ &= Cov(Y_{t_1-t_2}, Y_0) \\ &= Cov(Y_0, Y_{|t_1-t_2|}) \\ &= \gamma_{0,|t_1-t_2|}\end{aligned}$$

Hence, the covariance between Y_{t_1} and Y_{t_2} depends on the lag between them and not on the actual time points t_1 and t_2. Therefore, for stationary processes, we can write the autocovariance functions and autocorrelation functions as follows:

$$\gamma_k = Cov(Y_t, Y_{t-k}) \tag{3.14}$$
$$\rho_k = Corr(Y_t, Y_{t-k}) \tag{3.15}$$

Moreover,

$$\rho_k = \frac{\gamma_k}{\gamma_0} \tag{3.16}$$

$\{Y_t\}$ will be termed as a **weakly stationary process** or **second order stationary process** if its mean function is constant over time and its covariance is independent of actual time t.

In this paper, we will only discuss about univariate, weakly statitionary time series.

3.2.7 White Noise

White noise is a sequence of independent, identically distributed random variables $\{e_t\}$, having mean zero and variance σ_e^2. It is a strictly stationary series, where,

$$\gamma_k = \begin{cases} Var(e_t), & \text{for } k = 0 \\ 0, & \text{for } k \neq 0 \end{cases} \tag{3.17}$$

And,

$$\rho_k = \begin{cases} 1, & \text{for } k = 0 \\ 0, & \text{for } k \neq 0 \end{cases} \tag{3.18}$$

3.3 Trend Estimation

In stationary time series, we assume that the mean function is constant. However, in practicality, that is never the case and so we often need to consider the mean functions to be simple functions of time or trends. These trends can either be **stochastic** or **deterministic**. Stochastic trends are impossible to model because they tend to show completely different characteristics with every simulation, e.g. the random walk model. On the other hand, a deterministic trend can be modeled using deterministic functions to

represent them. For example, a possible model of a time series with deterministic trend could be,

$$Y_t = \mu_t + X_t$$

Here,
μ_t = a deterministic function
X_t = unobserved deviations from μ_t, having zero mean
We might consider μ_t to be periodic. We could also assume it to be a linear or a quadratic function of time. However, it must be kept in mind that whenever we are stating that $E(X_t) = 0$, we are assuming that the trend μ_t will last forever.

3.3.1 Estimating Constant Mean

Say,

$$Y_t = \mu_t + X_t \tag{3.19}$$

The most common estimate of the constant mean for (3.19) is the sample mean, which is calculated from the observed time series $Y_1, Y_2.....Y_n$,

$$\overline{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t$$

Therefore, $E(\overline{Y}) = \mu$
To determine the accuracy of $\overline{Y}$ as an estimate of μ, we will have to make some guesses regarding X_t and test it. Say, $\{Y_t\}$, or equivalently $\{X_t\}$, is a stationary time series with $\rho = \rho_k$. Then,

$$\begin{aligned} Var(\overline{Y}) &= \frac{1}{n^2} Var[\sum_{t=1}^{n} Y_t] \\ &= \frac{1}{n^2} Cov[\sum_{t=1}^{n} Y_t \sum_{s=1}^{n} Y_s] \\ &= \frac{1}{n^2}\sum_{t=1}^{n}\sum_{s=1}^{n} \gamma_{t-s} \end{aligned}$$

Now, putting $t - s = k$ and $t = j$, we get,

$$\begin{aligned} \{1 \le t \le n, 1 \le s \le n\} &\implies \{1 \le j \le n, 1 \le j - k \le n\} \\ &\implies \{k + 1 \le j \le n + k, 1 \le j \le n\} \\ &\implies \{k > 0, k + 1 \le j \le n\} \cup \{k \le 0, 1 \le j \le n + k\} \end{aligned}$$

Therefore,

$$\begin{aligned} Var(\overline{Y}) &= \frac{1}{n^2}[\sum_{k=1}^{n-1}\sum_{j=k+1}^{n}\gamma_k + \sum_{k=-n+1}^{0}\sum_{j=1}^{n+k}\gamma_k] \\ &= \frac{1}{n^2}[\sum_{k=1}^{n-1}(n-k)\gamma_k + \sum_{k=-n+1}^{0}(n+k)\gamma_k] \\ &= \frac{1}{n}\sum_{k=-n+1}^{n-1}(1-\frac{|k|}{n})\gamma_k \\ &= \frac{\gamma_0}{n}\sum_{k=-n+1}^{n-1}(1-\frac{|k|}{n})\rho_k \end{aligned} \tag{3.20}$$

(3.20) is used to evaluate the estimation of μ. By assuming $\{X_t\}$ as different models, the value of ρ_k is placed and the variance of the estimate is approximated. If the variance of the estimate varies with the sample size n, then the estimate is rejected.

3.3.2 Estimating Non-Constant Mean

Regression analysis can be used to estimate the parameters of non-constant mean trend models. For example, if the trend is a linear function of time, then the mean is represented as,

$$\mu_t = \beta_0 + \beta_1 t \tag{3.21}$$

Here,
β_0 = intercept
β_1 = slope
Both the slope and the intercept need to be estimated in this case, which is done by choosing those value of β_0 and β_1 that will minimize the following:

$$Q(\beta_0, \beta_1) = \sum_{t=1}^{n}[Y_t - (\beta_0 + \beta_1 t)]^2$$

One of the solutions of the above equation is shown below:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{t=1}^{n}(Y_t - \overline{Y})(t - \bar{t})}{\sum_{t=1}^{n}(t-\bar{t})^2} \\ \hat{\beta}_0 &= \overline{Y} - \hat{\beta_1}\bar{t} \end{aligned} \tag{3.22}$$

Here, $\bar{t} = \frac{n+1}{2}$
The least square estimate of the slope can be expressed as follows:

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n}(t-\bar{t})Y_t}{\sum_{t=1}^{n}(t-\bar{t})^2} \tag{3.23}$$

Say, $c_1, c_2.....c_m$ are constants and $t_1, t_2.....t_m$ are time points. Then,

$$Var[\sum_{i=1}^{n} c_i Y_{t_j}] = \sum_{i=1}^{n} c_i^2 Var(Y_{t_i}) + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} c_i c_j Cov(Y_{t_i}, Y_{t_j}) \tag{3.24}$$

Using (3.23) and (3.24), we find the variance,

$$\begin{aligned}
Var(\hat{\beta}_1) &= Var(\frac{\sum_{t=1}^{n}(t-\bar{t})Y_t}{\sum_{t=1}^{n}(t-\bar{t})^2}) \\
&= \{\frac{1}{\sum_{t=1}^{n}(t-\bar{t})^2}\}^2 Var(\sum_{t=1}^{n}(t-\bar{t})Y_t) \\
&= \frac{144}{n^2(n^2-1)^2}[\sum_{t=1}^{n}(t-\bar{t})^2 Var(Y_t) + 2\sum_{s=2}^{n}\sum_{t=1}^{s-1}(t-\bar{t})(s-\bar{t})Cov(Y_t,Y_s)] \\
&= \frac{144}{n^2(n^2-1)^2}[\frac{n(n^2-1)}{12}\gamma_0 + 2\sum_{s=2}^{n}\sum_{t=1}^{s-1}(t-\bar{t})(s-\bar{t})\gamma_{s-t}] \\
&= \frac{12\gamma_0}{n(n^2-1)} + \frac{288\gamma_0}{n^2(n^2-1)^2}\sum_{s=2}^{n}\sum_{t=1}^{s-1}(t-\bar{t})(s-\bar{t})\rho_{s-t}] \\
&= \frac{12\gamma_0}{n(n^2-1)}[1 + \frac{24}{n(n^2-1)}\sum_{s=2}^{n}\sum_{t=1}^{s-1}(t-\bar{t})(s-\bar{t})\rho_{s-t}] \qquad (3.25)
\end{aligned}$$

Here, we replaced $\sum_{t=1}^{n}(t-\bar{t})^2$ by $\frac{n(n^2-1)}{12}$.
Using (3.25), the precision of the estimates of the linear trend model can be evaluated in the same way as that of the constant mean model.

If the data is monthly seasonal, then it is assumed that there are twelve constants $\beta_1, \beta_2.....\beta_{12}$, each of which are equal to the average monthly observations of the year. So,

$$\mu_t = \begin{cases} \beta_1, & \text{for } t = 1, 13, ... \\ \beta_2, & \text{for } t = 2, 14, ... \\ . & \\ . & \\ . & \\ \beta_{12}, & \text{for } t = 12, 24, ... \end{cases} \qquad (3.26)$$

(3.26) is also known as the **seasonal means model**. The estimate of any parameter of the seasonal model, say $\hat{\beta}_j$ is given by,

$$\frac{1}{N}\sum_{i=0}^{N-1} Y_{j+12i}$$

Here, N = number of years of monthly data
Since $\hat{\beta}_j$ is like $\overline{Y}$ and it only considers every 12th values, we can transform (3.20) as follows:

$$Var(\hat{\beta}_j) = \frac{\gamma_0}{N}[1 + 2\sum_{k=1}^{N-1}(1-\frac{k}{N})\rho_{12k}] \qquad \text{for } j = 1, 2, ...12 \qquad (3.27)$$

Using (3.27), the precision of the estimates of the seasonal model can be evaluated in the same way as that of the constant mean model.

Some seasonal trends can be modeled using cosine curves. This helps to maintain smoothness in the curve during transitions from one time period to another. Such a model is as follows:

$$\mu_t = \beta cos(2\pi ft + \Phi) \tag{3.28}$$

Here,
β = amplitude of the curve
f = frequency of the curve
Φ = phase of the curve
As the value of t changes, the curve fluctuates between the highest value of β and the lowest value of $-\beta$.
A more convenient form of (3.28) is,

$$\beta cos(2\pi ft + \Phi) = \beta_1 cos(2\pi ft) + \beta_2 sin(2\pi ft) \tag{3.29}$$

Here,

$$\beta = \sqrt{\beta_1^2 + \beta_2^2}, \qquad \Phi = atan(-\frac{\beta_2}{\beta_1}) \tag{3.30}$$

Similarly,

$$\beta_1 = \beta cos(\Phi), \qquad \beta_2 = \beta sin(\Phi) \tag{3.31}$$

Thus, the simplest model for the mean would be

$$\mu_t = \beta_0 + \beta_1 cos(2\pi ft) + \beta_2 sin(2\pi ft) \tag{3.32}$$

Say, frequency, $f = \frac{m}{n}$. Here, m is an integer and $1 \leq m < \frac{n}{2}$. Then,

$$\begin{aligned} \hat{\beta_1} &= \frac{2}{n}\sum_{t=1}^{n}[cos(\frac{2\pi mt}{n})Y_t] \\ \hat{\beta_2} &= \frac{2}{n}\sum_{t=1}^{n}[sin(\frac{2\pi mt}{n})Y_t] \end{aligned} \tag{3.33}$$

Using (3.33) and (3.24), we get the variance of $\hat{\beta}_1$,

$$\begin{aligned} Var(\hat{\beta_1}) &= Var(\frac{2}{n}\sum_{t=1}^{n}[cos(\frac{2\pi mt}{n})Y_t]) \\ &= \frac{4}{n^2}[\sum_{t=1}^{n}\{cos(\frac{2\pi mt}{n})\}^2 Var(Y_t) + 2\sum_{s=2}^{n}\sum_{t=1}^{s-1} cos(\frac{2\pi mt}{n})cos(\frac{2\pi ms}{n})Cov(Y_t, Y_s)] \\ &= \frac{4}{n^2}\frac{n}{2}\gamma_0 + \frac{8}{n^2}\sum_{s=2}^{n}\sum_{t=1}^{s-1} cos(\frac{2\pi mt}{n})cos(\frac{2\pi ms}{n})\gamma_{s-t} \\ &= \frac{2\gamma_0}{n} + \frac{8\gamma_0}{n^2}\sum_{s=2}^{n}\sum_{t=1}^{s-1} cos(\frac{2\pi mt}{n})cos(\frac{2\pi ms}{n})\rho_{s-t} \\ &= \frac{2\gamma_0}{n}[1 + \frac{4}{n}\sum_{s=2}^{n}\sum_{t=1}^{s-1} cos(\frac{2\pi mt}{n})cos(\frac{2\pi ms}{n})\rho_{s-t}] \end{aligned} \tag{3.34}$$

Here, we replaced $\sum_{t=1}^{n}\{cos(\frac{2\pi mt}{n})\}^2$ by $\frac{n}{2}$.
Similarly, the variance of $\hat{\beta}_2$ can be calculated by replacing the cosines with sines in the above derivation.

3.3.3 Analyzing Estimated Outputs

After estimation of μ_t, we can estimate the stochastic component X_t for each t, using the following equation:

$$X_t = Y_t - \mu_t$$

The standard deviation of $\{X_t\}$ can be calculated, if its variance is constant, by the **residual standard deviation**, which is given by,

$$s = \sqrt{\frac{1}{n-p}\sum_{t=1}^{n}(Y_t - \hat{\mu_t})^2} \tag{3.35}$$

Here,
p = number of parameters estimated for μ_t
n-p = degrees of freedom for s
s is an absolute measure of the estimated trend's goodness of fit. The less is its value, the better is the fit.
Another measure of the estimated trend's goodness of fit is the value of R^2, which is also known as the **co-efficient of determination** or **multiple R-squared**. It is unitless and is defined as the square of the sample correlation co-efficient between the series of observation and the approximated trend.
The **adjusted** R^2 approximates unbiased estimates depending on the number of parameter estimated in the trend. Its value is basically just a small adjustment to the value of R^2. if there are several models with different number of parameters, then the adjusted R^2 helps to compare them.
The **standard deviations**, also known as, Std. Error of the estimated co-efficients should not be taken into consideration unless the stochastic component is found to be white noise. When we divide each estimated regression coefficient by their respective standard errors, we get **t-values** or the **t-ratios**. These values do not come to any use if the stochastic component is not white noise.

3.4 Time Series Models

Integrated Autoregressive Moving Average models also known as the ARIMA models are one of the several approaches to time series forecasting. They aim to utilize the autocorrelations in the time series data to make future predictions. ARIMA models are a broad class of parametric models that have gained a lot of popularity in time series forecasting. The fundamental concepts of ARIMA models are discussed in this section.

3.4.1 General Linear Process

A **general linear process** is a process which is a weighted linear combination of the present and past white noise terms. If $\{Y_t\}$ is an observed time series and $\{e_t\}$ is an unobserved white noise series where $e_1, e_2.....e_t$ are independent, identically distributed random variables, then $\{Y_t\}$ can be expressed as a general linear process in the following manner:

$$Y_t = \Psi_0 e_t + \Psi_1 e_{t-1} + \Psi_2 e_{t-2} + ... \tag{3.36}$$

In order to make the infinite series of the right hand side of (3.36) more meaningful, we assume that,

$$\sum_{i=1}^{\infty} \Psi_i^2 < \infty \tag{3.37}$$

Since $\{e_t\}$ is not observable, we can presume that $\Psi_0 = 1$. Therefore,

$$Y_t = e_t + \Psi_1 e_{t-1} + \Psi_2 e_{t-2} + ... \tag{3.38}$$

The Ψs are often considered to form an exponentially decaying sequence, such that,

$$\Psi_j = \phi^j$$

Here, the value of ϕ is strictly between -1 and +1.
Then, (3.38) becomes

$$Y_t = e_t + \phi e_{t-1} + \phi^2 e_{t-2} + ... \tag{3.39}$$

The mean function of Y_t,

$$\begin{aligned} E(Y_t) &= E(e_t + \phi e_{t-1} + \phi^2 e_{t-2} + ...) \\ &= E(e_t) + \phi E(e_{t-1}) + \phi^2 E(e_{t-2}) + ... \\ &= 0 + 0 + 0 + ... \\ &= 0 \end{aligned}$$

The variance of Y_t,

$$\begin{aligned} Var(Y_t) &= Var(e_t + \phi e_{t-1} + \phi^2 e_{t-2} + ...) \\ &= Var(e_t) + \phi^2 Var(e_{t-1}) + \phi^4 Var(e_{t-2}) + ... \\ &= \sigma_e^2 + \phi^2\sigma_e^2 + \phi^4\sigma_e^2 + ... \\ &= \sigma_e^2(1 + \phi^2 + \phi^4 + ...) \\ &= \frac{\sigma_e^2}{1 - \phi^2} \end{aligned}$$

The covariance between two consecutive observations becomes,

$$\begin{aligned} Cov(Y_t, Y_{t-1}) &= Cov(e_t + \phi e_{t-1} + \phi^2 e_{t-2} + ..., e_{t-1} + \phi e_{t-2} + \phi^2 e_{t-3} + ...) \\ &= Cov(e_t, e_{t-1}) + Cov(e_t, \phi e_{t-2}) + ... + Cov(\phi e_{t-1}, e_{t-1}) + Cov(\phi e_{t-1}, \phi e_{t-2}) + \\ &\quad ..... + Cov(\phi^2 e_{t-2}, e_{t-1}) + Cov(\phi^2 e_{t-2}, \phi e_{t-2}) + ... \\ &= 0 + 0 + ... + \phi\sigma_e^2 + 0 + ... + 0 + \phi^3\sigma_e^2 \\ &= \phi\sigma_e^2(1 + \phi^2 + \phi^4 + ...) \\ &= \frac{\phi\sigma_e^2}{1 - \phi^2} \end{aligned}$$

Therefore, covariance between two observations that are k lags apart,

$$Cov(Y_t, Y_{t-k}) = \frac{\phi^k\sigma_e^2}{1 - \phi^2}$$

It is clear from the previous discussion that the autocovariance function is dependent only on the lag k and not on the actual time. So, we can conclude that Y_t is a stationary

process.
Now, the correlation between two consecutive observations is,

$$\begin{aligned} Corr(Y_t, Y_{t-1}) &= \frac{Cov(Y_t, Y_{t-1})}{\sqrt{Var(Y_t)Var(Y_{t-1})}} \\ &= \frac{\frac{\phi\sigma_e^2}{1-\phi^2}}{\sqrt{\frac{\sigma_e^2}{1-\phi^2}\frac{\sigma_e^2}{1-\phi^2}}} \\ &= \frac{\frac{\phi\sigma_e^2}{1-\phi^2}}{\frac{\sigma_e^2}{1-\phi^2}} \\ &= \phi \end{aligned}$$

Therefore, correlation between two observations that are k lags apart,

$$Corr(Y_t, Y_{t-k}) = \phi^k \tag{3.40}$$

Hence, for a general linear process with $\Psi_0 = 1$, we have,

$$E(Y_t) = 0, \qquad \gamma_k = \sigma_e^2 \sum_{i=0}^{\infty} \Psi_i \Psi_{i+k} \qquad for \qquad k \geqslant 0 \tag{3.41}$$

If the general linear process has non-zero mean, then it can be expressed as follows:

$$Y_t = \mu + e_t + \Psi_1 e_{t-1} + \Psi_2 e_{t-2} + ...$$

3.4.2 Moving Average Process

A **moving average (MA)** process is a general linear process having finite number of Ψ weights. It is represented as follows:

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ..... - \theta_q e_{t-q} \tag{3.42}$$

Such a process is known as a moving average process of order q, MA(q).
It is called a moving average process because the weights $1, -\theta_1, -\theta_2, ..., -\theta_q$ are applied to variables $e_t, e_{t-1}.....e_{t-q}$ to get Y_t. Then the weights are shifted once to the right to get Y_{t+1} by the application on the next set of q variables $e_{t+1}, e_t.....e_{t-q+1}$ and so on.

MA(1) Process

An MA(1) process is represented as follows:

$$Y_t = e_t - \theta e_{t-1}$$

Here,

$$\begin{aligned} E(Y_t) &= E(e_t - \theta e_{t-1}) \\ &= E(e_t) - \theta E(e_{t-1}) \\ &= 0 - 0 \\ &= 0 \end{aligned}$$

$$\begin{aligned} Var(Y_t) &= Var(e_t - \theta e_{t-1}) \\ &= Var(e_t) + \theta^2 Var(e_{t-1}) \\ &= \sigma_e^2 + \theta^2 \sigma_e^2 \\ &= \sigma_e^2(1+\theta^2) \\ &= \gamma_0 \end{aligned}$$

$$\begin{aligned} \gamma_1 &= Cov(Y_t, Y_{t-1}) \\ &= Cov(e_t - \theta e_{t-1}, e_{t-1} - \theta e_{t-2}) \\ &= Cov(e_t, e_{t-1}) + Cov(e_t, -\theta e_{t-2}) + Cov(-\theta e_{t-1}, e_{t-1}) + Cov(-\theta e_{t-1}, -\theta e_{t-2}) \\ &= 0 + 0 - \theta\sigma_e^2 + 0 \\ &= -\theta\sigma_e^2 \end{aligned}$$

$$\begin{aligned} \gamma_2 &= Cov(Y_t, Y_{t-2}) \\ &= Cov(e_t - \theta e_{t-1}, e_{t-2} - \theta e_{t-3}) \\ &= Cov(e_t, e_{t-2}) + Cov(e_t, -\theta e_{t-3}) + Cov(-\theta e_{t-1}, e_{t-2}) + Cov(-\theta e_{t-1}, -\theta e_{t-3}) \\ &= 0 + 0 + 0 + 0 \\ &= 0 \end{aligned}$$

$\therefore \gamma_k = Cov(Y_t, Y_{t-k}) = 0$ whenever $k \geqslant 2$.

$$\begin{aligned} \rho_1 &= \frac{\gamma_1}{\gamma_0} \\ &= \frac{-\theta\sigma_e^2}{\sigma_e^2(1+\theta^2)} \\ &= \frac{-\theta}{1+\theta^2} \end{aligned}$$

And, $\rho_k = 0$ for $k \geqslant 2$.

MA(2) Process

An MA(2) process is represented by,

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}$$

Here,

$$\begin{aligned} E(Y_t) &= E(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}) \\ &= E(e_t) - \theta_1 E(e_{t-1}) - \theta_2 E(e_{t-2}) \\ &= 0 - 0 - 0 \\ &= 0 \end{aligned}$$

$$\begin{aligned} Var(Y_t) &= Var(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}) \\ &= Var(e_t) + \theta_1^2 Var(e_{t-1}) + \theta_2^2 Var(e_{t-2}) \\ &= \sigma_e^2 + \theta_1^2\sigma_e^2 + \theta_2^2\sigma_e^2 \\ &= \sigma_e^2(1+\theta_1^2+\theta_2^2) \\ &= \gamma_0 \end{aligned}$$

$$\begin{aligned}
\gamma_1 &= Cov(Y_t, Y_{t-1}) \\
&= Cov(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}, e_{t-1} - \theta_1 e_{t-2} - \theta_2 e_{t-3}) \\
&= Cov(-\theta_1 e_{t-1}, e_{t-1}) + Cov(-\theta_2 e_{t-2}, -\theta_1 e_{t-2}) \\
&= -\theta_1 \sigma_e^2 + \theta_1 \theta_2 \sigma_e^2 \\
&= (-\theta_1 + \theta_1 \theta_2)\sigma_e^2
\end{aligned}$$

$$\begin{aligned}
\gamma_2 &= Cov(Y_t, Y_{t-2}) \\
&= Cov(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}, e_{t-2} - \theta_1 e_{t-3} - \theta_2 e_{t-4}) \\
&= Cov(-\theta_2 e_{t-2}, e_{t-2}) \\
&= -\theta_2 \sigma_e^2
\end{aligned}$$

$\gamma_k = Cov(Y_t, Y_{t-k}) = 0$ whenever $k \geqslant 3$.

$$\begin{aligned}
\rho_1 &= \frac{\gamma_1}{\gamma_0} \\
&= \frac{(-\theta_1 + \theta_1 \theta_2)\sigma_e^2}{\sigma_e^2(1 + \theta_1^2 + \theta_2^2)} \\
&= \frac{-\theta_1 + \theta_1 \theta_2}{1 + \theta_1^2 + \theta_2^2}
\end{aligned}$$

$$\begin{aligned}
\rho_2 &= \frac{\gamma_2}{\gamma_0} \\
&= \frac{-\theta_2 \sigma_e^2}{\sigma_e^2(1 + \theta_1^2 + \theta_2^2)} \\
&= \frac{-\theta_2}{1 + \theta_1^2 + \theta_2^2}
\end{aligned}$$

And, $\rho_k = 0$ for $k \geqslant 3$.

MA(q) Process

For the MA(q) process in (3.42), the mean function is given by,

$$\begin{aligned}
E(Y_t) &= E(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ..... - \theta_q e_{t-q}) \\
&= E(e_t) - \theta_1 E(e_{t-1}) - \theta_2 E(e_{t-2}) - ..... - \theta_q E(e_{t-q}) \\
&= 0 - 0 - 0 - ..... - 0 \\
&= 0
\end{aligned}$$

The variance of Y_t,

$$\begin{aligned}
Var(Y_t) &= Var(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ..... - \theta_q e_{t-q}) \\
&= Var(e_t) + \theta_1^2 Var(e_{t-1}) + \theta_2^2 Var(e_{t-2}) + ... + \theta_q^2 Var(e_{t-q}) \\
&= \sigma_e^2 + \theta_1^2 \sigma_e^2 + \theta_2^2 \sigma_e^2 + ... + \theta_q^2 \sigma_e^2 \\
&= \sigma_e^2(1 + \theta_1^2 + \theta_2^2 + ... + \theta_q^2) \qquad (3.43)
\end{aligned}$$

$$\gamma_k = \begin{cases} \sigma_e^2(-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + ... + \theta_{q-k}\theta_q), & \text{for } k \leqslant q \\ 0, & \text{for } k > q \end{cases} \tag{3.44}$$

$$\rho_k = \begin{cases} \dfrac{-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + ... + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \theta_2^2 + ... + \theta_q^2}, & \text{for } k \leqslant q \\ 0, & \text{for } k > q \end{cases} \tag{3.45}$$

Invertibility
In case of an MA(1) process, both θ and $\frac{1}{\theta}$ give the same ACF. However, this is not acceptable as it would lead to wrong estimations of the parameters during model specification. So, it has to be made sure that no two values of the same parameter leads to the same ACF for an MA function. Invertible MA series are such processes with unique ACF. Let us consider the following MA(1) process,

$$\begin{aligned} Y_t &= e_t - \theta e_{t-1} \\ e_t &= Y_t + \theta e_{t-1} \end{aligned}$$

Replacing t by t-1, we get,

$$e_{t-1} = Y_{t-1} + \theta e_{t-2}$$

Therefore,

$$\begin{aligned} Y_t &= e_t - \theta e_{t-1} \\ e_t &= Y_t + \theta(Y_{t-1} + \theta e_{t-2}) \\ &= Y_t + \theta Y_{t-1} + \theta^2 e_{t-2} \end{aligned}$$

The substitution may continue infinitely into the past if $|\theta| < 1$. Thus an MA(1) model will be inverted into an infinite ordered AR model. And so, MA(1) is said to be invertible if $|\theta| < 1$.
The MA(q) characteristic equation is,

$$\theta(x) = 1 - \theta_1 x - \theta_2 x^2 - ..... - \theta_q x^q$$

The MA characteristics equation is,

$$1 - \theta_1 x - \theta_2 x^2 - ..... - \theta_q x^q = 0$$

To show that the MA(q) model is invertible, we can show that such a coefficient π_j exists that,

$$Y_t = \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + ... + e_t$$

This is only possible if the roots of the MA characteristics equation exceed 1.

3.4.3 Autoregressive Process

In **autoregressive processes**, the linear combination of the most recent p past values plus an error term e_t at time t gives the current value Y_t. All the things that are not explained by the past values of the series are incorporated into e_t. If Y_t is an autoregressive process of order p, that is, AR(p), then it can be expressed as follows:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + e_t \tag{3.46}$$

Here, e_t is independent of all past values $Y_{t-1}, Y_{t-2,....}$.

AR(1) Process
An AR(1) process can be written as,

$$Y_t = \phi Y_{t-1} + e_t \tag{3.47}$$

Say, the process mean has been deducted. So, the mean function of the series is,

$$E(Y_t) = 0$$

The variance is,

$$\begin{aligned} \gamma_0 &= Var(Y_t) \\ &= Var(\phi Y_{t-1} + e_t) \\ &= \phi^2 Var(Y_{t-1}) + Var(e_t) \\ &= \phi^2 \gamma_0 + \sigma_e^2 \\ \therefore \gamma_0 &= \frac{\sigma_e^2}{1-\phi^2} \end{aligned} \tag{3.48}$$

Here, $\phi^2 < 1$.
If we multiply both sides of (3.47) by Y_{t-k} and take expectation, we get,

$$\begin{aligned} E(Y_t Y_{t-k}) &= \phi E(Y_{t-1} Y_{t-k}) + E(e_t Y_{t-k}) \\ \therefore \gamma_k &= \phi \gamma_{k-1} + E(e_t Y_{t-k}) \end{aligned}$$

Since e_t is independent of Y_{t-k},

$$\begin{aligned} E(e_t Y_{t-k}) &= E(e_t) E(Y_{t-k}) \\ &= 0 \end{aligned}$$

$$\therefore \gamma_k = \phi \gamma_{k-1} \qquad for\ k \geqslant 1 \tag{3.49}$$

When $k = 1$,

$$\begin{aligned} \gamma_1 &= \phi \gamma_0 \\ &= \phi \frac{\sigma_e^2}{1-\phi^2} \end{aligned}$$

When $k = 2$,

$$\begin{aligned} \gamma_2 &= \phi \gamma_1 \\ &= \phi (\phi \frac{\sigma_e^2}{1-\phi^2}) \\ &= \phi^2 \frac{\sigma_e^2}{1-\phi^2} \end{aligned}$$

Therefore,

$$\gamma_k = \phi^k \frac{\sigma_e^2}{1-\phi^2} \tag{3.50}$$

And,

$$\begin{aligned} \rho_k &= \frac{\gamma_k}{\gamma_0} \\ &= \frac{\phi^k \frac{\sigma_e^2}{1-\phi^2}}{\frac{\sigma_e^2}{1-\phi^2}} \\ &= \phi^k \qquad for\ k \geqslant 1 \end{aligned} \tag{3.51}$$

Since $\phi^2 < 1$, as the number of lags k increases, ρ_k decreases exponentially.
Now, suppose,

$$Y_{t-1} = \phi Y_{t-2} + e_{t-1}$$

Then, (3.47) can be written as,

$$\begin{aligned} Y_t &= \phi(\phi Y_{t-2} + e_{t-1}) + e_t \\ &= \phi^2 Y_{t-2} + \phi e_{t-1} + e_t \\ \therefore Y_t &= \phi^k Y_{t-k} + \phi^{k-1} Y_{t-k+1} + ... + \phi^2 Y_{t-2} + \phi Y_{t-1} + e_t \end{aligned} \tag{3.52}$$

If the series on the right hand side of (3.52) is infinite instead, then we can write it as follows:

$$Y_t = e_t + \phi e_{t-1} + \phi^2 e_{t-2} + ... \tag{3.53}$$

This is similar to (3.38), if we replace the Ψ_js there by ϕ^j.
The following is an AR characteristic model for AR(1) process:

$$\phi(x) = 1 - \phi x$$

It is used to explain the stationarity of AR(1) process. The corresponding AR characteristic equation is,

$$1 - \phi x = 0$$

To get the stationary condition of AR(1) model,the root of the characteristic equation is used. When the root, taken in its absolute form, exceeds 1, we get the stationarity condition. Thus, $x = \frac{1}{\phi}$ has to exceed 1 in absolute form, which happens only when $|\phi| < 1$.

AR(2) Process
An AR(2) process is written as,

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t \tag{3.54}$$

The following is an AR characteristic model for AR(2) process:

$$\phi(x) = 1 - \phi_1 x - \phi_2 x^2$$

It is used to explain the stationarity of AR(2) process. The corresponding AR characteristic equation is,

$$1 - \phi_1 x - \phi_2 x^2 = 0$$

To get the stationary condition of AR(2) model,the roots of the characteristic equation are used. When the roots, taken in their absolute form, exceed 1, we get the stationarity condition.Now, the roots of the characteristics equation are given by,

$$x = \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2} \tag{3.55}$$

Here, $|x| > 1$ iff,

$$\phi_1 + \phi_2 < 1, \qquad \phi_2 - \phi_1 < 1, \qquad |\phi_2| < 1 \tag{3.56}$$

Multiplying both sides of (3.54) by Y_{t-k} and take expectation, we get,

$$\begin{aligned}
E(Y_t Y_{t-k}) &= \phi_1 E(Y_{t-1}Y_{t-k}) + \phi_2 E(Y_{t-2}Y_{t-k}) + E(e_t Y_{t-k}) \\
\gamma_k &= \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2} + 0 \\
\therefore \gamma_k &= \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2} \qquad for\ k \geqslant 1 \qquad (3.57) \\
\frac{\gamma_k}{\gamma_0} &= \phi_1\frac{\gamma_{k-1}}{\gamma_0} + \phi_2\frac{\gamma_{k-2}}{\gamma_0} \\
\therefore \rho_k &= \phi_1\rho_{k-1} + \phi_2\rho_{k-2} \qquad for\ k \geqslant 1 \qquad (3.58)
\end{aligned}$$

(3.57) and (3.58) are known as the **Yule Walker Equations**.
Variance of AR(2) model,

$$\begin{aligned}
\gamma_0 &= Var(Y_t) \\
&= Var(\phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t) \\
&= Var(\phi_1 Y_{t-1} + \phi_2 Y_{t-2}) + Var(e_t) \\
&= Var(\phi_1 Y_{t-1}) + Var(\phi_2 Y_{t-2}) + 2Cov(\phi_1 Y_{t-1}, \phi_2 Y_{t-2}) + \sigma_e^2 \\
&= \phi_1^2\gamma_0 + \phi_2^2\gamma_0 + 2\phi_1\phi_2\gamma_1 + \sigma_e^2 \\
&= (\phi_1^2 + \phi_2^2)\gamma_0 + 2\phi_1\phi_2\gamma_1 + \sigma_e^2 \qquad (3.59)
\end{aligned}$$

If $k = 1$, (3.57) gives,

$$\begin{aligned}
\gamma_1 &= \phi_1\gamma_0 + \phi_2\gamma_1 \\
\therefore \gamma_1 &= \frac{\phi_1\gamma_0}{1 - \phi_2}
\end{aligned}$$

Therefore, from (3.59),

$$\begin{aligned}
\gamma_0 &= \phi_1^2\gamma_0 + \phi_2^2\gamma_0 + 2\phi_1\phi_2\frac{\phi_1\gamma_0}{1-\phi_2} + \sigma_e^2 \\
&= \frac{\phi_1^2(1-\phi_2)\gamma_0 + \phi_2^2(1-\phi_2)\gamma_0 + 2\phi_1\phi_2\gamma_0 + (1-\phi_2)\sigma_e^2}{(1-\phi_2)} \\
\therefore (1-\phi_2)\gamma_0 &= \gamma_0[\phi_1^2(1-\phi_2) + \phi_2^2(1-\phi_2) + 2\phi_1\phi_2] + (1-\phi_2)\sigma_e^2 \\
\therefore \gamma_0 &= \frac{(1-\phi_2)\sigma_e^2}{1 - \phi_2 - \phi_1^2(1-\phi_2) - \phi_2^2(1-\phi_2) - 2\phi_1\phi_2} \\
\gamma_0 &= \frac{(1-\phi_2)\sigma_e^2}{(1-\phi_2)(1-\phi_1^2-\phi_2^2) - 2\phi_1\phi_2} \qquad (3.60)
\end{aligned}$$

AR(p) Process
The AR characteristic polynomial for an AR(p) model is,

$$\phi(x) = 1 - \phi_1 x - \phi_2 x^2 - ..... - \phi_p x^p \tag{3.61}$$

The corresponding AR characteristic equation is then represented by,

$$1 - \phi_1 x - \phi_2 x^2 - ..... - \phi_p x^p = 0 \tag{3.62}$$

To get the stationary condition of AR(p) model,the roots of the characteristic equation is used. When the roots, taken in their absolute form, exceed 1, we get the stationarity conditions:

$$\phi_1 + \phi_2 + ... + \phi_p < 1 \qquad and \qquad |\phi_p| < 1 \tag{3.63}$$

Multiplying both sides of (3.46) by Y_{t-k} and take expectation, we get,

$$\begin{aligned} E(Y_t Y_{t-k}) &= \phi_1 E(Y_{t-1}Y_{t-k}) + \phi_2 E(Y_{t-2}Y_{t-k}) + ... + \phi_p E(Y_{t-p}Y_{t-k}) + E(e_t Y_{t-k}) \\ \therefore \gamma_k &= \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2} + ... + \phi_p\gamma_{k-p} + 0 \\ \gamma_k &= \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2} + ... + \phi_p\gamma_{k-p} \qquad (3.64) \\ \therefore \rho_k &= \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + ... + \phi_p\rho_{k-p} \qquad fork \geqslant 1 \qquad (3.65) \end{aligned}$$

We get the following Yule Walker equations if we set $k = 1, 2, ...p$, $\rho_0 = 1$ and $\rho_{-k} = \rho_k$ in (3.65):

$$\left.\begin{aligned} \rho_1 &= \phi_1 + \phi_2\rho_1 + \phi_3\rho_2 + ... + \phi_p\rho_{p-1} \\ \rho_2 &= \phi_1\rho_1 + \phi_2 + \phi_3\rho_1 + ... + \phi_p\rho_{p-2} \\ &. \\ &. \\ &. \\ \rho_p &= \phi_1\rho_{p-1} + \phi_2\rho_{p-2} + \phi_3\rho_{p-3} + ... + \phi_p \end{aligned}\right\} \tag{3.66}$$

Multiplying both sides of (3.46) by Y_t and take expectation, we get,

$$\begin{aligned} E(Y_t Y_t) &= \phi_1 E(Y_{t-1}Y_t) + \phi_2 E(Y_{t-2}Y_t) + ... + \phi_p E(Y_{t-p}Y_t) + E(e_t Y_t) \\ \therefore \gamma_0 &= \phi_1\gamma_1 + \phi_2\gamma_2 + ... + \phi_p\gamma_p + \sigma_e^2 \\ \implies 1 &= \phi_1\rho_1 + \phi_2\rho_2 + ... + \phi_p\rho_p + \frac{\sigma_e^2}{\gamma_0} \\ \implies \frac{\sigma_e^2}{\gamma_0} &= 1 - \phi_1\rho_1 - \phi_2\rho_2 - .... - \phi_p\rho_p \\ \therefore \gamma_0 &= \frac{\sigma_e^2}{1 - \phi_1\rho_1 - \phi_2\rho_2 - .... - \phi_p\rho_p} \qquad (3.67) \end{aligned}$$

3.4.4 ARMA Models

Autoregressive moving average models are models of those time series which are partly autoregressive and partly moving average. If a series Y_t consists of an autoregressive process of order p and a moving average process of order q, then it is known as a ARMA(p,q) process. It is expressed as follows:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - .... - \theta_q e_{t-q} \tag{3.68}$$

ARMA(1,1) Model

An ARMA(1,1) model is shown below,

$$Y_t = \phi Y_{t-1} + e_t - \theta e_{t-1} \tag{3.69}$$

Here,

$$\begin{aligned} E(e_t Y_t) &= E[e_t(\phi Y_{t-1} + e_t - \theta e_{t-1})] \\ &= 0 + E(e_t e_t) + 0 \\ &= Var(e_t) \\ &= \sigma_e^2 \end{aligned}$$

$$\begin{aligned} E(e_{t-1} Y_t) &= E[e_{t-1}(\phi Y_{t-1} + e_t - \theta e_{t-1})] \\ &= \phi E(e_{t-1} e_{t-1}) + 0 - \theta E(e_{t-1} e_{t-1}) \\ &= \phi Var(e_t) - \theta Var(e_t) \\ &= \phi\sigma_e^2 - \theta\sigma_e^2 \\ &= \sigma_e^2(\phi - \theta) \end{aligned}$$

Multiplying both sides of (3.69) by Y_{t-k} and take expectation, we get,

$$\begin{aligned} E(Y_t Y_{t-k}) &= E(\phi Y_{t-1} Y_{t-k} + e_t Y_{t-k} - \theta e_{t-1} Y_{t-k}) \\ \therefore \gamma_k &= \phi E(Y_{t-1} Y_{t-k}) + E(e_t Y_{t-k}) - \theta E(e_{t-1} Y_{t-k}) \\ \therefore \gamma_k &= \phi\gamma_{k-1} + 0 - 0 \\ \therefore \gamma_k &= \phi\gamma_{k-1} \end{aligned}$$

If k=0,

$$\begin{aligned} \gamma_0 &= \phi E(Y_{t-1} Y_t) + E(e_t Y_t) - \theta E(e_{t-1} Y_t) \\ &= \phi\gamma_1 + \sigma_e^2 - \sigma_e^2(\phi - \theta)\theta \\ &= \phi\gamma_1 + \sigma_e^2[1 - (\phi - \theta)\theta] \end{aligned}$$

If k=1,

$$\begin{aligned} \gamma_1 &= \phi E(Y_{t-1} Y_{t-1}) + E(e_t Y_{t-1}) - \theta E(e_{t-1} Y_{t-1}) \\ &= \phi\gamma_0 + 0 - \theta\sigma_e^2 \\ &= \phi\gamma_0 - \theta\sigma_e^2 \end{aligned}$$

Therefore,

$$\left.\begin{aligned} &\gamma_0 = \phi\gamma_1 + \sigma_e^2[1 - (\phi - \theta)\theta] \\ &\gamma_1 = \phi\gamma_0 - \theta\sigma_e^2 \\ &\gamma_k = \phi\gamma_{k-1} \qquad for\ k \geqslant 2 \end{aligned}\right\} \tag{3.70}$$

From the first two equations of (3.70), we get,

$$\begin{aligned} \gamma_0 &= \phi(\phi\gamma_0 - \theta\sigma_e^2) + \sigma_e^2[1 - (\phi - \theta)\theta] \\ &= \phi^2\gamma_0 - \phi\theta\sigma_e^2 + \sigma_e^2 - \phi\theta\sigma_e^2 + \theta^2\sigma_e^2 \\ &= \phi^2\gamma_0 - 2\phi\theta\sigma_e^2 + \sigma_e^2 + \theta^2\sigma_e^2 \\ &= \frac{\sigma_e^2 - 2\phi\theta\sigma_e^2 + \theta^2\sigma_e^2}{1 - \phi^2} \\ &= \frac{1 - 2\phi\theta + \theta^2}{1 - \phi^2}\sigma_e^2 \end{aligned} \tag{3.71}$$

If we solve the recursion, we get,

$$\rho_k = \frac{(1-\theta\phi)(\phi-\theta)}{1-2\theta\phi+\theta^2}\phi^{k-1} \qquad for\ k \geqslant 1 \tag{3.72}$$

To get the stationarity conditions of the ARMA(1,1) process, we have to ensure that the absolute value of the root of the AR characteristic equation $1-\phi x = 0$ exceeds 1. This happens when,

$$|\phi| < 1$$

This is the stationarity condition of ARMA(1,1) model.

3.4.5 ARIMA Models

If a process $\{Y_t\}$ is non-stationary, then it means that it has a non-constant mean over time. If we difference consecutive observations of the time series, then the mean gets stabilized to some extent as the changes in the level get removed. If required, differencing can be done more than once on the time series data to achieve stationarity.
If differencing is done once, it is called the first order differencing,

$$\bigtriangledown Y_t = Y_t - Y_{t-1}$$

If it is done twice, then it is called the second order differencing and so on. A second order differencing looks like the following:

$$\begin{aligned} \bigtriangledown^2 Y_t &= \bigtriangledown(\bigtriangledown Y_t) \\ &= (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \\ &= Y_t - 2Y_{t-1} - Y_{t-2} \end{aligned}$$

If we perform first order difference of the seriesin Figure 3.1, we get the output as shown in Figure 3.3.

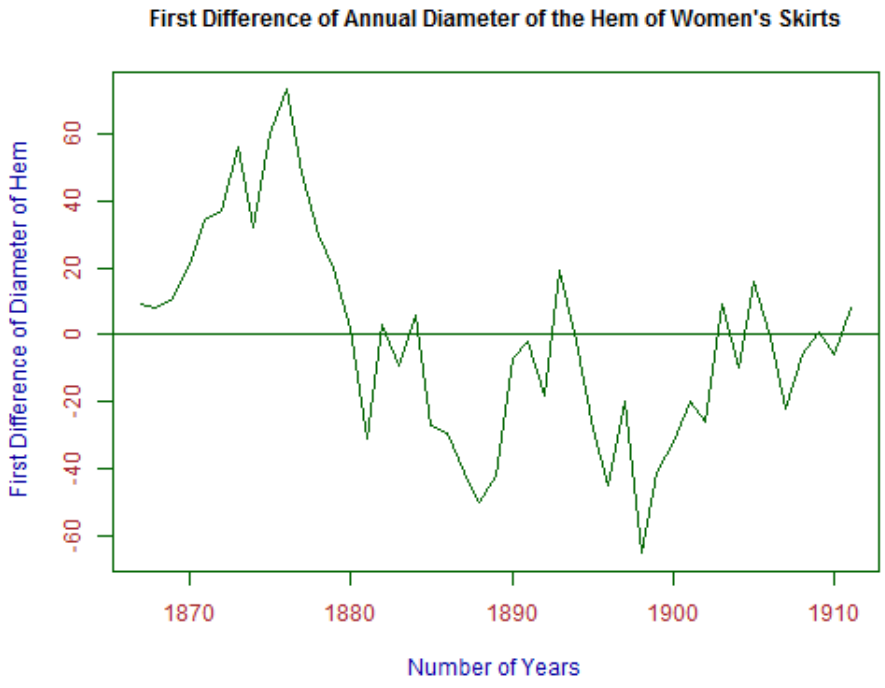

Figure 3.3: Plot of First Order Difference of the Diameter of Hem Series

It can be seen clearly in Figure 3.3 that the trend has been removed substantially from the series. If we perform second order difference on the series, then we will get,

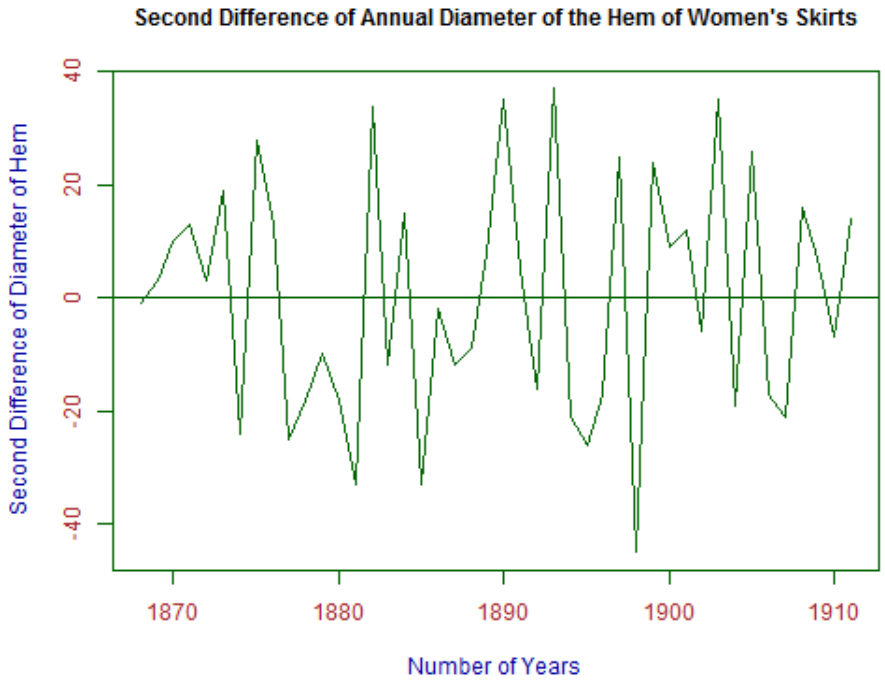

Figure 3.4: Plot of Second Order Difference of the Diameter of Hem Series

$\{Y_t\}$ will be an **integrated autoregressive moving average** process, if its d^{th} difference, denoted by $W_t = \bigtriangledown^d Y_t$, is a stationary ARMA process. Therefore, $\{W_t\}$ follows an ARMA(p,q) model and $\{Y_t\}$ follows an ARIMA(p,d,q) model. Typically, the value of d is at least 0 (meaning no difference) or at most 2.

An ARIMA(p,1,q) series is represented as follows in terms of its observations:

$$\begin{aligned} Y_t - Y_{t-1} &= \phi_1(Y_{t-1} - Y_{t-2}) + \phi_2(Y_{t-2} - Y_{t-3}) + ... + \phi_p(Y_{t-p} - Y_{t-p-1}) \\ &\quad + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ..... - \theta_q e_{t-q} \\ \implies Y_t &= (1+\phi_1)Y_{t-1} + (\phi_2 - \phi_1)Y_{t-2} + (\phi_3 - \phi_2)Y_{t-3} + ... + (\phi_p - \phi_{p-1})Y_{t-p} - \phi_p Y_{t-p-1} \\ &\quad + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ..... - \theta_q e_{t-q} \end{aligned} \tag{3.73}$$

Or, it can be expressed as follows:

$$W_t = \phi_1 W_{t-1} + \phi_2 W_{t-2} + ... + \phi_p W_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ..... - \theta_q e_{t-q} \tag{3.74}$$

(3.73), which looks like an ARMA(p+1,q) process, is also known as the difference equation form of the ARIMA model.

The characteristic polynomial equation of ARIMA (p,d,q) model is as follows:

$$1-(1+\phi_1)x-(\phi_2-\phi_1)x^2-.....-(\phi_p-\phi_{p-1})x^p+\phi_p x^{p+1} = (1-x)(1-\phi_1 x-\phi_2 x^2-....-\phi_p x^p)$$

From the above equation, we can see that one of the roots is $x = 1$, which implies non-stationarity. The other roots are the roots of the characteristic polynomial equation

of the stationary time series $\bigtriangledown Y_t$.

We consider stationary time series to have zero mean. However, if we wish to accommodate a non-zero constant mean μ in the ARMA process $\{W_t\}$, we can suppose that,

$$W_t - \mu = \phi_1(W_{t-1} - \mu) + \phi_2(W_{t-2} - \mu) + ... + \phi_p(W_{t-p} - \mu) + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ..... - \theta_q e_{t-q}$$

We can also add a constant term θ_0 to the model instead:

$$W_t = \theta_0 + \phi_1 W_{t-1} + \phi_2 W_{t-2} + ... + \phi_p W_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ..... - \theta_q e_{t-q}$$

If we take expectations on both sides of the above equation, we get,

$$\begin{aligned}
E(W_t) &= \theta_0 + \phi_1 E(W_{t-1}) + \phi_2 E(W_{t-2}) + ... + \phi_p E(W_{t-p}) + E(e_t) \\
&\quad -\theta_1 E(e_{t-1}) - \theta_2 E(e_{t-2}) - ..... - \theta_q E(e_{t-q}) \\
\therefore \mu &= \theta_0 + \phi_1\mu + \phi_2\mu + ... + \phi_p\mu + 0 - 0 - 0 - ..... - 0 \\
\implies \mu &= \theta_0 + \phi_1\mu + \phi_2\mu + ... + \phi_p\mu \\
\implies \mu &= \theta_0 + (\phi_1 + \phi_2 + ... + \phi_p)\mu \\
\therefore \mu &= \frac{\theta_0}{1 - \phi_1 - \phi_2 - ..... - \phi_p} \qquad (3.75) \\
\therefore \theta_0 &= \mu(1 - \phi_1 - \phi_2 - ..... - \phi_p) \qquad (3.76)
\end{aligned}$$

Often, it is observed in time series that the higher its level is, the more variations it shows around that level and vice versa. That is, its variance increases as its level increases. In such cases, transforming the data-set to its log form will result in a series with constant variance over time. If the level of the original time series varies exponentially with time, then the new log transformed time series will show a linear time trend, which can be removed by differencing. Therefore, the new series can be expressed as,

$$\bigtriangledown \left[log(Y_t)\right] = X_t \qquad (3.77)$$

In stock price predictions, often, the returns are considered to perform analysis. These returns are the differences of the logarithms of the stock prices.
Data can also be transformed by using power functions. Such transformation is known as **power transformations**. If λ is given, then the power transformation is defined by,

$$g(x) = \begin{cases} \dfrac{x^\lambda - 1}{\lambda}, & \text{for } \lambda \neq 0 \\ log(x), & \text{for } k = 0 \end{cases} \qquad (3.78)$$

The value of λ is estimated and used to transform the non-stationary time series. Power transformation only works if the values of the data are positive. Otherwise, they are transformed after adding the absolute form of the lowest value to all the data values and making the data values positive.

3.4.6 Backshift Operator

The **backshift operator** of a time series operates on the time index of the observations of the series to produce the previous observations, e.g. $BY_t = Y_{t-1}$.

Applying the backshift operator on the general MA(q) model, we get,

$$\begin{aligned} Y_t &= e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ..... - \theta_q e_{t-q} \\ &= e_t - \theta_1 B e_t - \theta_2 B^2 e_t - ..... - \theta_q B^q e_t \\ &= e_t(1 - \theta_1 B - \theta_2 B^2 - ..... - \theta_q B^q) \\ &= \theta(B) e_t \end{aligned}$$

Applying the backshift operator on the general AR(p) model, we get,

$$\begin{aligned} Y_t &= \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + e_t \\ \implies e_t &= Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - ..... - \phi_p Y_{t-p} \\ &= Y_t - \phi_1 B Y_t - \phi_2 B^2 Y_t - ..... - \phi_p B^p Y_t \\ &= (1 - \phi_1 B - \phi_2 B^2 - ..... - \phi_p B^p) Y_t \\ &= \phi(B) Y_t \end{aligned}$$

Applying the backshift operator on the general ARMA(p,q) model, we get,

$$\phi(B) Y_t = \theta(B) e_t$$

Applying the backshift operator on the differencing equations, we get,

$$\begin{aligned} \bigtriangledown Y_t &= Y_t - Y_{t-1} \\ &= Y_t - B Y_t \\ &= Y_t(1 - B) \end{aligned}$$

$$\begin{aligned} \bigtriangledown^2 Y_t &= Y_t - 2Y_{t-1} + Y_{t-2} \\ &= Y_t - 2B Y_t + B^2 Y_t \\ &= Y_t(1 - B)^2 \end{aligned}$$

Applying the backshift operator on the general ARIMA(p,d,q) model, we get,

$$\phi(B)(1 - B)^d Y_t = \theta(B) e_t$$

3.5 Box-Jenkins Procedure

The ARIMA models discussed in **Section 3.4** are fitted to time series data for further analysis and forecasting. George Box and Gwilym Jenkins had established a method for finding the best fit of ARIMA models to past values of a time series. This method is known as the Box-Jenkins procedure, which is widely used in time series analysis and forecasting.
The Box-Jenkins procedure consists of three steps:

1. Model Specification
2. Parameter Estimation
3. Model Verification

3.5.1 Model Specification

Model specification involves determining reasonable yet tentative values for p, d and q of the ARIMA(p,d,q) model to fit to the time series data. The tools that are used for this purpose are discussed here.

Sample Autocorrelation Function

If $Y_1, Y_2, ...Y_n$ is the observed time series, then,

$$\begin{aligned} \text{mean, } \overline{Y} &= \frac{\sum_{t=1}^{n} Y_t}{n} \\ \text{autocovariance, } c_k &= \hat{\gamma}_k \\ &= \frac{\sum_{t=k+1}^{n}(Y_t - \overline{Y})(Y_{t-k} - \overline{Y})}{T} \\ \text{autocorrelation, } r_k &= \hat{\rho}_k \\ &= \frac{\hat{\gamma}_k}{\hat{\gamma}_0} \\ &= \frac{\sum_{t=k+1}^{n}(Y_t - \overline{Y})(Y_{t-k} - \overline{Y})}{\sum_{t=1}^{n}(Y_t - \overline{Y})^2} \end{aligned} \tag{3.79}$$

r_k is the sample autocorrelation function and it is used to identify MA(q) process. The plot of r_k against lag k is called a correlogram. We know from (3.45) that for $k > q$, the autocorrelation function ρ_k becomes zero. So, if the value of r_k drops to zero at a particular lag in the correlogram, then we can say that it is a MA process and we can also determine the value of q from it.

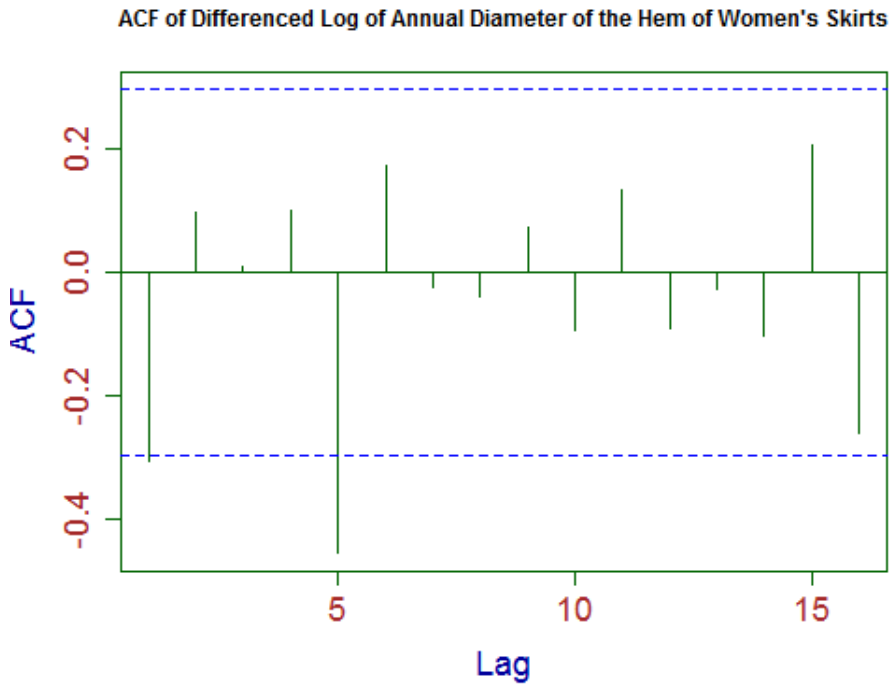

Figure 3.5: ACF Plot of the Diameter of Hem Series

We plot the ACF of our example time series on the diameter of hem of skirts in Figure 3.5. From the plot, we can see that the ACF exceeds the significance bound at lag 1. So, we can guess that the series could be an MA(1) process.

Sample Partial Autocorrelation Function

r_k works as a good indicator of the order q of a moving average process. However, in case of autoregressive processes, the autocorrelation function never becomes zero after a certain amount of lags. It simply dies off. Hence, we have to define a function for the correlation between two observations Y_t and Y_{t-k}. We can define the function in such a way that the effect of the observations $Y_{t-1}, Y_{t-2}, ..., Y_{t-k+1}$ are removed. This function is known as the **partial autocorrelation function** or PACF, and is denoted by ϕ_{kk}.
If $\{Y_t\}$ is a time series with normal distribution, then, ϕ_{kk} is defined by,

$$\phi_{kk} = Corr(Y_t Y_{t-k} | Y_{t-1}, Y_{t-2}, ...Y_{t-k+1}) \tag{3.80}$$

In case we wish to define ϕ_{kk} for both normally distributed and non-normally distributed series, then, we can assume that the prediction of Y_t is based on a linear combination of its intervening variables:

$$\beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_{k-1} Y_{t-k+1}$$

Here, βs are selected in such a way that the mean square error of the prediction gets minimized. Since it is a stationary series, the prediction of Y_{t-k} will also be based on a linear combination of its intervening variables:

$$\beta_1 Y_{t-k+1} + \beta_2 Y_{t-k+2} + ... + \beta_{k-1} Y_{t-1}$$

Then the PACF at lag k will be,

$$\begin{aligned} \phi_{kk} &= Corr(Y_t - \beta_1 Y_{t-1} - \beta_2 Y_{t-2} - ..... - \beta_{k-1} Y_{t-k+1}, Y_{t-k} - \beta_1 Y_{t-k+1} - \beta_2 Y_{t-k+2} - \\ &\quad ..... - \beta_{k-1} Y_{t-1}) \end{aligned} \tag{3.81}$$

We always consider ϕ_{11} to be equal to 1. It can be shown that based on Y_{t-1} alone, the best linear prediction of Y_t is $\rho_1 Y_{t-1}$. Therefore,

$$\begin{aligned} Cov(Y_t - \rho_1 Y_{t-1}, Y_{t-2} - \rho_1 Y_{t-1}) &= Cov(Y_t, Y_{t-2}) - \rho_1 Cov(Y_t, Y_{t-1}) - \rho_1 Cov(Y_{t-1}, Y_{t-2}) + \\ &\quad \rho_1^2 Cov(Y_{t-1}, Y_{t-1}) \\ &= \gamma_2 - \rho_1\gamma_1 - \rho_1\gamma_1 + \rho_1^2\gamma_0 \\ &= \rho_2\gamma_0 - \rho_1^2\gamma_0 - \rho_1^2\gamma_0 + \rho_1^2\gamma_0 \\ &= (\rho_2 - \rho_1^2 - \rho_1^2 + \rho_1^2)\gamma_0 \\ &= (\rho_2 - \rho_1^2)\gamma_0 \end{aligned}$$

$$\begin{aligned} Var(Y_t - \rho_1 Y_{t-1}) &= Var(Y_{t-2} - \rho_1 Y_{t-1}) \\ &= \gamma_0 - \rho_1^2\gamma_0 \\ &= \gamma_0(1 - \rho_1^2) \end{aligned}$$

Therefore,

$$\phi_{22} = \frac{(\rho_2 - \rho_1^2)\gamma_0}{\gamma_0(1-\rho_1^2)} = \frac{\rho_2 - \rho_1^2}{1-\rho_1^2} \tag{3.82}$$

For an AR(1) model, where $\rho_k = \phi^k$,

$$\phi_{22} = \frac{\rho_2 - \rho_1^2}{1-\rho_1^2} = \frac{\phi^2 - \phi^2}{1-\phi^2} = 0$$

Therefore, for AR(1) process, $\phi_{kk} = 0$ for all $k > 1$. So, we can say that the PACF for an AR(p) process would cut off when the lag becomes greater than its order. That is,

$$\phi_{kk} = 0 \qquad for\ all\ k > p \tag{3.83}$$

In case of an MA(1) process, from (3.82), we get,

$$\begin{aligned} \phi_{22} &= \frac{0 - (\frac{-\theta}{1+\theta^2})^2}{1 - (\frac{-\theta}{1+\theta^2})^2} \\ &= \frac{-\theta^2}{(1+\theta^2)^2 - \theta^2} \\ &= \frac{-\theta^2}{1 + 2\theta^2 + \theta^4 - \theta^2} \\ &= \frac{-\theta^2}{1 + \theta^2 + \theta^4} \end{aligned} \tag{3.84}$$

So, for an MA(q) model, ϕ_{kk} never becomes equal to zero. It only dies off. So, it can be used as a tool to exclusively identify an AR process.

The value of ϕ_{kk} can be found by using the following Yule Walker equations:

$$\rho_j = \phi_{k1}\rho_{j-1} + \phi_{k2}\rho_{j-2} + ... + \phi_{kk}\rho_{j-k} \qquad for\ j = 1, 2, ...k \tag{3.85}$$

Here, we assume that the values of $\rho_1, \rho_2...., \rho_k$ are given. By estimating the values of ρs using the sample autocorrelation functions, that is, by replacing ρ_ks by r_ks, we can solve (3.85) to get the values of sample autocorrelation functions ($\hat{\phi_{kk}}$). There is a method called the Levinson-Durbin method, by which we can show that (3.85) can be solved to find an equation for ϕ_{kk}:

$$\phi_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1} \phi_{k-1,j}\rho_{k-j}}{1 - \sum_{j=1}^{k-1} \phi_{k-1,j}\rho_j} \tag{3.86}$$

Here, $\phi_{k,j} = \phi_{k-1,j} - \phi_{kk}\phi_{k-1,k-j}$ for $j = 1, 2, ...k-1$. We plot the PACF of our example time series on the diameter of hem of skirts in Figure 3.6. From the plot, we can see that the PACF exceeds the significance bound at lag 1. So, we can guess that the series could be an AR(1) process.

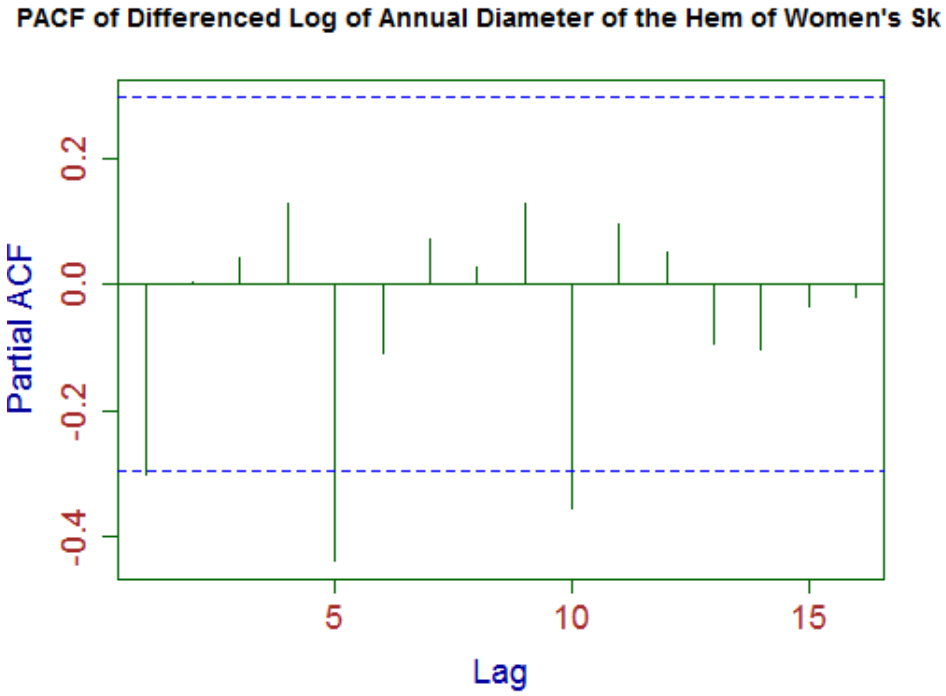

Figure 3.6: PACF Plot of the Diameter of Hem Series

Extended Autocorrelation Function

Both sample ACF and sample PACF are very effective in identifying the MA(q) and AR(p) models respectively. However, in case of mixed ARMA(p,q) models, both the ACF and the PACF tend to tail off instead of becoming zero within a finite number of lags (See Table 3.1).

	AR(p)	MA(q)	ARMA(p,q)
ACF	Dies off	Cuts off after q lags	Dies off
PACF	Cuts off after p lags	Dies off	Dies off

Table 3.1: Behavior of ACF and PACF for Different ARMA Models

Various tools are used in such cases, where the series seems to follow an ARMA model, such as, the extended autocorrelation function (EACF), the corner method, smallest canonical correlation method, etc. In our work, we have only considered about the EACF method.

In the EACF method, it is assumed that if we can determine the autoregressive part of a mixed ARMA model, by "filtering" it out from that model, we can get a pure moving average process. We can then use sample ACF to determine the order of the moving average part.

The coefficients of the autoregressive part are estimated by using a finite sequence of regressions. If $\{Y_t\}$ is a true ARMA(1,1) model, then,

$$Y_t = \phi Y_{t-1} + e_t - \theta e_{t-1}$$

Performing a linear regression of Y_t on Y_{t-1} gives an estimator of ϕ which is inconsistent. By performing another regression of Y_t on Y_{t-1} and on the lag one residuals of the first regression will result in a consistent estimator $\breve{\phi}$. Then, the autoregressive part of the series will be $\breve{\phi}Y_{t-1}$. By deducting it from the series, we will get an approximately pure moving average series:

$$W_t = Y_t - \breve{\phi}Y_{t-1}$$

Similarly, for an ARMA(p,q) process, we can estimate the autoregressive coefficients by a sequence of q regressions. If we consider the AR order to be k and MA order to be j, then the remaining pure MA series will be:

$$W_{t,k,j} = Y_t - \breve{\phi}_t Y_{t-1} - ..... - \breve{\phi}_k Y_{t-k} \tag{3.87}$$

The sample ACF of $W_{t,k,j}$ is known as the extended ACF.
It was suggested by Tsay and Tiao that the sample EACF information should be summarized by a table, where the elements of the (k,j)-th slot will be 'X' if the sample ACF of at lag j+1 of $W_{t,k,j}$ is significantly different from zero, and 'O' otherwise.

Finding 'd'

If the sample ACF of a time series fails to die off quickly as the number of lags increases, it can be assumed that the series is non-stationary. In such cases, by using the different transformation methods, including differencing methods, we can turn it into a stationary series. If any order of differencing is done on the series, then the value of d of the ARIMA(p,d,q) model becomes that order.

Over-differencing
If we difference any stationary series, we get another stationary series. Over-differencing a series can lead to various complications in the modeling process and so, care should be taken while choosing a differencing order.

Dickey Fuller Unit Root Test
The Dickey Fuller unit root test is a method of hypothesis testing to find if a series is stationary or not. In this test, the null hypothesis is that the series is non-stationary. To reject the null hypothesis, the probability value has to be less than 0.1. Let us assume that in the following model, $\{X_t\}$ is a stationary AR(k) process:

$$Y_t = \alpha Y_{t-1} + X_t$$

Here, $\{Y_t\}$ will be stationary if $|\alpha| < 1$, and non-stationary if $\alpha = 1$. Therefore, under the null hypothesis that $\{Y_t\}$ is non-stationary, let $a = \alpha - 1$ and $X_t = Y_t - Y_{t-1}$. Then,we get,

$$\begin{aligned} Y_t - Y_{t-1} &= \alpha Y_{t-1} - Y_{t-1} + X_t \\ &= (\alpha - 1)Y_{t-1} + X_t \\ &= aY_{t-1} + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_k X_{t-k} + e_t \\ &= aY_{t-1} + \phi_1(Y_{t-1} - Y_{t-2}) + \phi_2(Y_{t-2} - Y_{t-3}) + ... + \phi_k(Y_{t-k} - Y + t - k - 1) \\ &\quad + e_t \end{aligned} \tag{3.88}$$

Here, $\{Y_t\}$ will be difference non-stationary if $\alpha = 1$, that is, if $a = 0$. If $-1 < \alpha < 1$, then $\{Y_t\}$ will follow an AR(k+1) model, having AR characteristic equation,

$$\begin{aligned} 1 - \alpha x - \phi_1 x - \phi_2 x^2 - ..... - \phi_k x^k &= 0 \\ (1 - \phi_1 x - \phi_2 x^2 - ..... - \phi_k x^k)(1 - \alpha x) &= 0 \end{aligned}$$

Therefore, if $x = 1$, that is, the equation has unit root, then the process will be considered non-stationary which is the null hypothesis. Otherwise, it will be considered as stationary.Therefore, to find if a series needs differencing or not, all that is required is to find if the characteristic equation has unit root.
In case, there is possibility that the series has a non-zero mean, then an intercept is augmented to (3.88). The test is then known as the Augmented Dickey Fuller Test.

Akaike's Information Criterion (AIC)
In this method, only that model is chosen, which minimizes the AIC. The AIC is given by,

$$AIC = -2log(maximum \qquad likelihood) + 2k \tag{3.89}$$

If the model contains an intercept or constant term, then k=p+q+1. Otherwise k=p+q. 2k here is the penalty function. By adding the number of parameters, it is ensured that models with too many unnecessary parameters do not get chosen. AIC estimates the mean Kullback-Leibler divergence of the estimated model from the actual model. If $Y_1, Y_2....Y_n$ is a model with true probability distribution function $p(y_1, y_2, ..., y_n)$ and if its estimated probability distribution function is $q_\theta(y_1, y_2, ..., y_n)$, having parameter θ, then the Kullback - Leibler divergence of p from q_θ is given by,

$$D(p, q_\theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} ..... \int_{-\infty}^{\infty} p(y_1, y_2, ..., y_n) log[\frac{p(y_1, y_2, ..., y_n)}{q_\theta(y_1, y_2, ..., y_n)}] dy_1 dy_2 .... dy_n$$

AIC is an estimator of $E[D(p, q_{\hat{\theta}})]$, where $\hat{\theta}$ is the maximum likelihood estimator of the vector parameter θ.

Corrected Akaike's Information Criterion (AIC_c)
AIC's estimations are biased. And so, Hurvich and Tsai came up with a new equation for AIC, which is called the corrected AIC or AIC_c. They added a non-stochastic penalty term to the existing equation of AIC (3.89) to eliminate the bias:

$$AIC_c = AIC + \frac{2(k+1)(k+2)}{n-k-2} \tag{3.90}$$

Here,
k = total number of parameters minus the noise variance
n = the effective sample size
If k/n is greater than 10%, AIC_c outperforms both AIC and BIC.

Bayesian Information Criterion (BIC)
In this method, only that model is chosen, which minimizes the BIC. The BIC is given by,

$$AIC = -2log(maximum \qquad likelihood) + klog(n) \tag{3.91}$$

The orders specified by BIC for the model of a true ARMA(p,q)process are always consistent as the sample size increases. However, if the true process is not an ARMA(p,q) process, then the AIC leads to more optimal selection of orders than BIC.

3.5.2 Parameter Estimation

Once a model is specified for a stationary time series (which could possibly have been a non-stationary time series that had been transformed into a staionary series), the parameters of the model are estimated. The different methods of parameter estimation are discussed here.

Method of Moments Estimators
In this method, the sample moments are equated to the theoretical moments and after solving the resulting equations, the estimates of the unknown parameters are acquired.
In case of AR(p) models, by equating ρ_is to r_is where $i = 1, 2, 3....p$, we get,

$$\left.\begin{aligned} r_1 &= \phi_1 + \phi_2 r_1 + \phi_3 r_2 + ... + \phi_p r_{p-1} \\ r_2 &= \phi_1 r_1 + \phi_2 + \phi_3 r_1 + ... + \phi_p r_{p-2} \\ &. \\ &. \\ &. \\ r_p &= \phi_1 r_{p-1} + \phi_2 r_{p-2} + \phi_3 r_{p-3} + ... + \phi_p \end{aligned}\right\} \tag{3.92}$$

These Yule Walker equations are solved to get, $\hat{\phi}_1, \hat{\phi}_2.....\hat{\phi}_p$.
In case of MA(q) models, method of moments is not a good estimator. We know that for MA(1) process,

$$\rho_1 = \frac{-\theta}{1+\theta^2}$$

When we equate ρ_1 to r_1:

- if $r_1 = \pm 0.5$, the solutions are not invertible
- if $|r_1| > 0.5$, no solution exists
- if $|r_1| < 0.5$, only one of the solutions is invertible

In case of ARMA(1,1) models, first we find $\hat{\phi}$ using the following formula:

$$\hat{\phi} = \frac{r_2}{r_1} \tag{3.93}$$

Then, by equating ρs to r's in (3.72), we get,

$$r_1 = \frac{(1-\theta\hat{\phi})(\hat{\phi}-\theta)}{1-2\theta\hat{\phi}+\theta^2} \tag{3.94}$$

We solve (3.94) to get $\hat{\theta}$. Since the model has MA process in it, care has to be taken to make sure that only invertible solutions are taken.

In order to estimate the noise variance, σ_e^2, at first the sample variance of the process is estimated using the following formula:

$$s^2 = \frac{1}{n-1}\sum_{t=1}^{n}(Y_t - \overline{Y})^2 \tag{3.95}$$

Then the relationship among variance, noise variance, θs and ϕs are used to estimate noise variance. For AR(p) models, from (3.67), we get,

$$\hat{\sigma_e}^2 = (1 - \hat{\phi}_1 r_1 - \hat{\phi}_2 r_2 - ..... - \hat{\phi}_p r_p)s^2 \tag{3.96}$$

For MA(q) models, from (3.44), we get,

$$\hat{\sigma_e}^2 = \frac{s^2}{1 + \theta_1^2 + \theta_2^2 + ... + \theta_q^2} \tag{3.97}$$

For ARMA(1,1) models, from (3.71), we get,

$$\hat{\sigma_e}^2 = \frac{1 - \hat{\phi}^2}{1 - 2\hat{\phi}\hat{\theta} + \hat{\theta}^2}s^2 \tag{3.98}$$

Least Squares Estimation (LSE)
In this method, we take a non-zero mean, μ, into consideration and include it in our model. Then we estimate it along with other parameters using least squares.
In case of AR(1) models, after including the non-zero mean, we get,

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + e_t \tag{3.99}$$

In LSE method, estimates are made by minimizing the sum of squares of the differences,

$$(Y_t - \mu) - \phi(Y_{t-1} - \mu)$$

The conditional sum of squares function of an AR(1) model is given by,

$$S_c(\phi, \mu) = \sum_{t=2}^{n}[(Y_t - \mu) - \phi(Y_{t-1} - \mu)]^2 \tag{3.100}$$

ϕs and μs are estimated by the values that minimize $S_c(\phi, \mu)$ when $Y_1, Y_2, ...Y_n$ are given. Say, $\frac{\partial S_c}{\partial \mu} = 0$, then,

$$\sum_{t=2}^{n} 2[(Y_t - \mu) - \phi(Y_{t-1} - \mu)](-1 + \phi) = 0$$

$$\therefore \mu = \frac{1}{(n-1)(1-\phi)}[\sum_{t=2}^{n} -\phi \sum_{t=2}^{n} Y_{t-1}] \tag{3.101}$$

For large n, we get,

$$\frac{\sum_{t=2}^{n} Y_t}{n-1} \approx \frac{\sum_{t=2}^{n} Y_{t-1}}{n-1} \approx \overline{Y}$$

Therefore,from (3.101), we get,

$$\hat{\mu} \approx \frac{(\overline{Y} - \phi\overline{Y})}{1-\phi} = \overline{Y} \tag{3.102}$$

Again,

$$\begin{aligned} \frac{\partial S_c(\phi, \overline{Y})}{\partial \phi} &= 0 \\ \sum_{t=2}^{n} 2[(Y_t - \overline{Y}) - \phi(Y_{t-1} - \overline{Y})](Y_{t-1} - \overline{Y}) &= 0 \\ \hat{\phi} &= \frac{\sum_{t=2}^{n}(Y_t - \overline{Y})(Y_{t-1} - \overline{Y})}{\sum_{t=2}^{n}(Y_{t-1} - \overline{Y})^2} \end{aligned}$$

This is almost similar to r_1 and so for large n, both method of moments and least squares estimators are identical.

For greater orders of AR processes, it can be showed that,

$$\hat{\mu} = \overline{Y} \tag{3.103}$$

In case of AR(2) models, the conditional sum of squares function is given by,

$$S_c(\phi_1, \phi_2, \overline{Y}) = \sum_{t=3}^{n}[(Y_t - \overline{Y}) - \phi_1(Y_{t-1} - \overline{Y}) - \phi_2(Y_{t-2} - \overline{Y})]^2 \tag{3.104}$$

To estimate the values of $\hat{\phi_1}$ and $\hat{\phi_2}$, $\frac{\partial S_c}{\partial \phi_1} = 0$ and $\frac{\partial S_c}{\partial \phi_2} = 0$ are set. Then the resulting equations are divided by $\sum_{t=3}^{n}(Y_t - \overline{Y})^2$. These, when rearranged, turn into Yule Walker equations like (3.92). These equations are then solved for $\hat{\phi_1}$ and $\hat{\phi_2}$. The same principle is followed to find the estimates of parameters of higher order AR processes.

In case of estimation of θs in MA(1) model, the MA(1) model is expressed as its invertible form:

$$Y_t = -\theta Y_{t-1} - \theta^2 Y_{t-2} - ..... + e_t$$

Thus, by using LSE, the value of θ can be estimated by minimizing,

$$S_c(\theta) = \sum (e_t)^2 = \sum [Y_t + \theta Y_{t-1} + \theta^2 Y_{t-2} + ...]^2 \tag{3.105}$$

Here, e_t is the function of the unknown parameter(θ) and the observed series.If we know the value of e_0, which is commonly assumed as zero, to calculate $e_1, e_2, ...e_n$, we use the following equation:

$$e_t = Y_t + \theta e_{t-1} \tag{3.106}$$

which is a rearranged version of the MA(1)model. Thus, we get,

$$\left.\begin{aligned} e_1 &= Y_1 \\ e_2 &= Y_2 + \theta e_1 \\ &. \\ &. \\ &. \\ e_n &= Y_n + \theta e_{n-1} \end{aligned}\right\} \tag{3.107}$$

Here, $Y_1, Y_2, ...Y_n$ are the observed values. Now, we can calculate, $S_c(\theta) = \sum(e_t)^2$ to get the value of θ. For higher order MA(q) models, the same principle applies. $e_t = e_t(\theta_1, \theta_2, ...\theta_q)$ is calculated recursively using the following equation:

$$e_t = Y_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + ... + \theta_q e_{t-q} \tag{3.108}$$

Here, it is assumed that $e_0 = e_{-1} = e_{-2}..... = e_{-q} = 0$. With the help of multivariate numerical method, the sum of squares is minimized.

In case of general ARMA(p,q) models, we use the same technique as the pure MA model. The following equation is used to compute the values of $e_1, e_2, ...e_n$:

$$e_t = Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - ..... - \phi_p Y_{t-p} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + ... + \theta_q e_{t-q} \tag{3.109}$$

Here, it is assumed that $e_p = e_{p-1} = ...e_{p+1-q} = 0$. In order to obtain the least squares estimate of all the parameters, $S_c(\phi_1, \phi_2, \phi_3, ...\phi_p, \theta_1, \theta_2, ...\theta_q)$ is minimized.

Maximum Likelihood Estimator (MLE)
In this method only those values are chosen for the parameters which maximize the likelihood function. The joint probability density of obtaining the actual observed data is called the likelihood function L.

In case of AR(1) model, the probability density function of each white noise term e_t is,

$$\frac{1}{\sqrt{2\pi\sigma_e^2}} \exp(-\frac{e_t^2}{2\sigma_e^2}) \qquad for - \alpha < e_t < \alpha$$

Then, since the white noise terms are independently and identically distributed, the joint probability density function for $e_2, e_3, ...e_n$ is given by,

$$\frac{1}{\sqrt{(2\pi\sigma_e^2)(n-1)}} \exp(-\frac{\sum_{t=2}^{n} e_t^2}{2\sigma_e^2}) \tag{3.110}$$

Say, $Y_1 = y_1$ is given and,

$$\left.\begin{aligned} &Y_2 - \mu = \phi(Y_1 - \mu) + e_t \\ &Y_3 - \mu = \phi(Y_2 - \mu) + e_3 \\ &. \\ &. \\ &. \\ &Y_n - \mu = \phi(Y_{n-1} - \mu) + e_n \end{aligned}\right\} \tag{3.111}$$

Then the joint probability density of $Y_2, Y_3, ...Y_n$ is given by the following equation,

$$f(y_2, y_3, ...y_n|y_1) = \frac{1}{\sqrt{(2\pi\sigma_e^2)^{n-1}}} \exp(-\frac{\sum_{t=2}^{n}[(Y_t - \mu) - \phi(Y_{t-1} - \mu)]^2}{2\sigma_e^2}) \tag{3.112}$$

Since this is an AR(1) process, the marginal probability distribution of Y_1 will be a normal distribution, having mean μ and variance $\frac{\sigma_e^2}{1-\phi^2}$. So, the joint probability density of $Y_1, Y_2, Y_3, ...Y_n$ will be equal to the joint probability density of $Y_2, Y_3, ...Y_n$ multiplied

by the marginal probability density of Y_1.
The likelihood function of AR(1) model is,

$$L(\phi, \mu, \sigma_e^2) = \frac{1}{\sqrt{(2\pi\sigma_e^2)^n}}\sqrt{1-\phi^2}\exp[-\frac{1}{2\sigma_e^2}S(\phi,\mu)] \tag{3.113}$$

Here, $S(\phi, \mu)$ is known as the unconditional sum of squares function and is given by,

$$S(\phi,\mu) = \sum_{t=2}^{n}[Y_t - \mu) - \phi(Y_{t-1} - \mu)]^2 + (1-\phi^2)(Y_1 - \mu) \tag{3.114}$$

Usually, instead of the likelihood function itself, its log is used. The log likelihood function of AR(1) model is given by,

$$l(\phi, \mu, \sigma_e^2) = \frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma_e^2) + \frac{1}{2}log(1-\phi^2) - \frac{1}{2\sigma_e^2}S(\phi,\mu) \tag{3.115}$$

3.5.3 Model Diagnostics

After the models are specified and their parameters are estimated, the next step is to diagnose the fitted models and check if the models fit the data well. There are two approaches for accomplishing this task. They can either be used singularly or together.

Residual Analysis
Residuals are the differences between the actual terms and the predicted terms of a model. For a general ARMA(p,q) model where the existing MA process is inverted to form an infinite autoregressive process, the residual is given by,

$$\hat{e_t} = Y_t - \hat{\pi_1}Y_{t-1} - \hat{\pi_2}Y_{t-2} - ..... \tag{3.116}$$

The residuals will resemble white noise if the models are a good fit, that is, the residuals will have zero mean and a constant standard deviation. Residuals are analyzed using different methods:

1. *Plots of Residuals*

   - The residuals are plotted over time and observed. If the residuals resemble white noise, then there will be no trend visible and the plot would scatter around a horizontal level forming a somewhat rectangular shape. There will be no increase or decrease of variation of the plot around the horizontal line. After fitting an ARIMA(1,2,1) model to our original time series of the diameter of the hem of women's skirts, the residuals we get, is plotted in Figure 3.7.

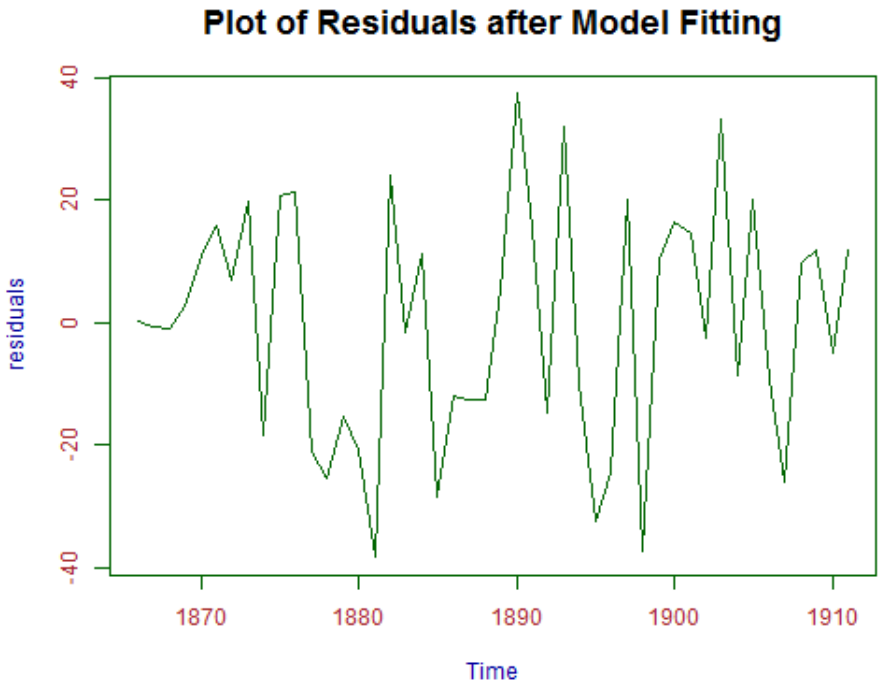

Figure 3.7: Plot of Residuals of the Fitted Model to the Diameter of Hem Series

2. *Normality of the Residuals*

   - Q-Q Plot or Quantile-Quantile plots shows the quantiles of the residuals versus the theoretical quantiles of a normal distribution for the residuals. If the points follow the straight line which passes through the first and third probability quantiles of the series, very closely, then we can say that the residuals are normally distributed. The Q-Q plot of our example series is shown in Figure 3.8. It seems to follow the said straight line pretty closely.
   - Histogram plot of the residuals can also help assess normality. If the histogram is somewhat symmetrical and the tails off at the two ends, then we can say that it resembles normal distribution. The histogram plot of our example series is shown in Figure 3.9. The plot seems to indicate a normal distribution to some extent.

3. *Sample Autocorrelation Function (ACF)*

   - The sample autocorrelation function is plotted to look for correlations among the residuals at different lags. A horizontal line is drawn on either side of zero at 2 approximate standard error of the sample ACF, that is, $\pm\frac{2}{\sqrt{n}}$. This is also known as the significance bound. If the values are within the significance bound, then we can say that the residuals are uncorrelated. Sample partial autocorrelation function is also plotted to ensure that there is no correlation among the residuals. The ACF plot of our example is shown in Figure 3.10. The plot shows significant correlation at lag 5, but not in the smaller lags. We can say that the model was not a perfect fit as there still seems to remain some information left in the residuals.

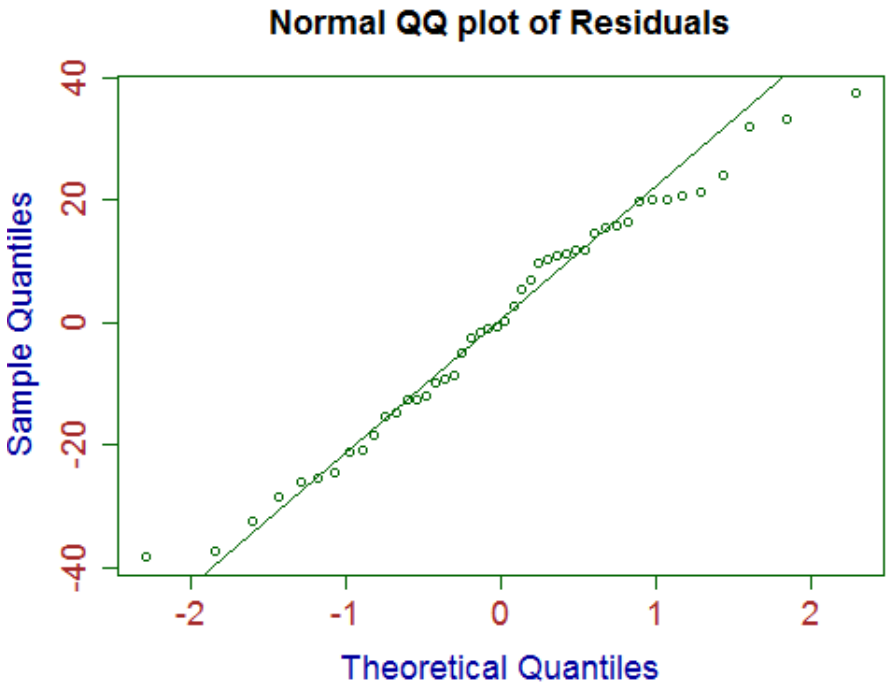

Figure 3.8: Q-Q plot of Residuals of the Fitted Model to the Diameter of Hem Series

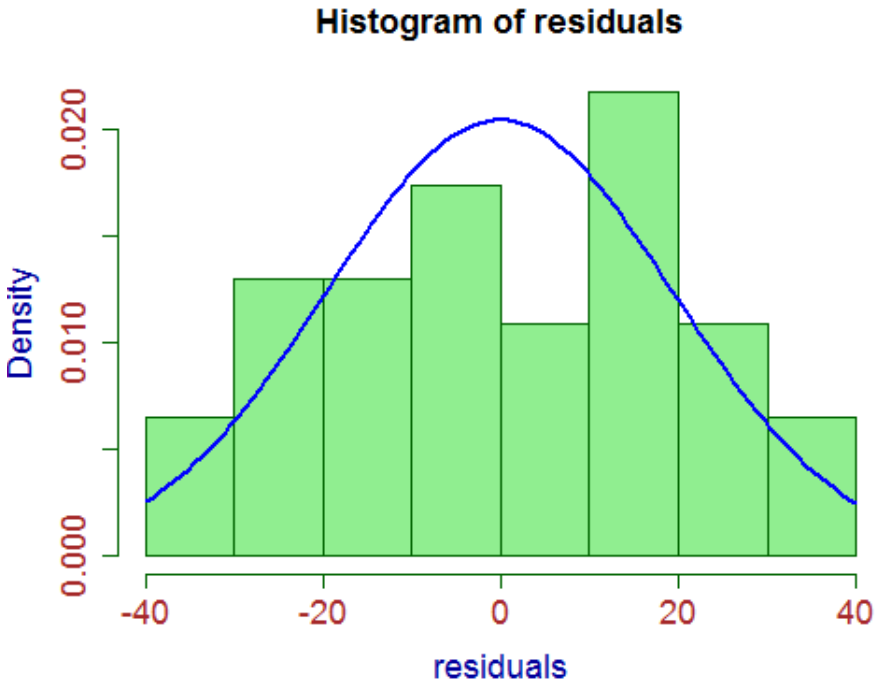

Figure 3.9: Histogram plot of Residuals of the Fitted Model to the Diameter of Hem Series

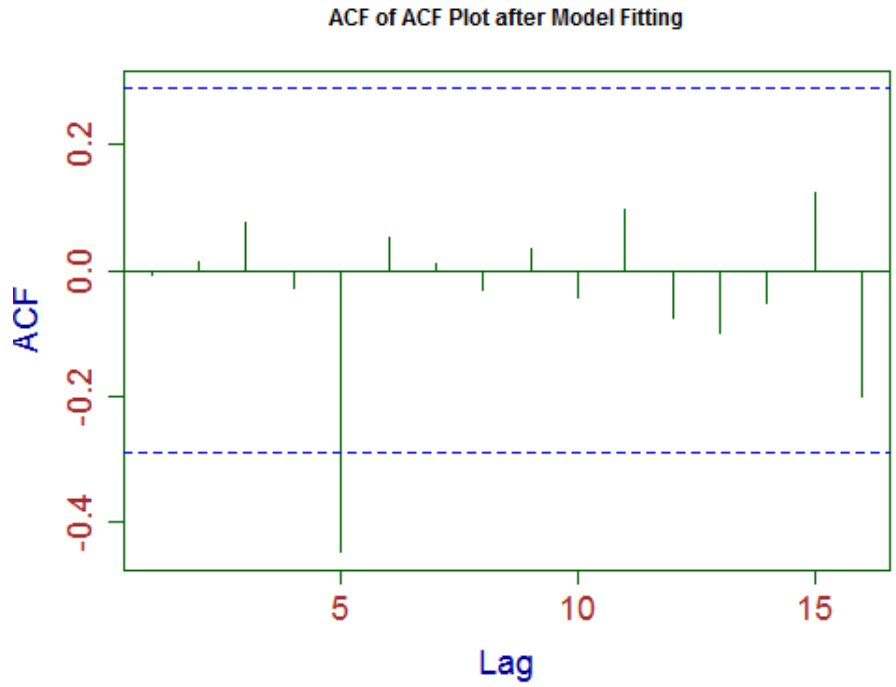

Figure 3.10: ACF plot of Residuals of the Fitted Model to the Diameter of Hem Series

4. *Ljung Box test*

   - The Ljung Box test is based on the following statistic:

$$Q_* = n(n+2)\sum_{k=1}^{m}\frac{\hat{r_k}^2}{n-k} \tag{3.117}$$

   Here, n is the sample size, $\hat{r_k}$ is the k-th sample autocorrelation function of the residuals. If the fitted model is correct, then the Ljung Box test Q_* will have a χ^2 distribution with k-p-q degrees of freedom. If the probability value is greater than 0.05, then the residuals are uncorrelated.

Overfitting and Parameter Redundancy

In this method, once a model is fitted to a series, another model which contains our model as a special case, should be fitted to the series. The orders of the latter model should not be too big compared to those of the former. That is, if the original fitted model was an AR(1) model, then the "bigger" model should be either ARMA(1,1) or AR(2). This will help to analyze more accurately. After fitting the latter model, if the estimated values of the new parameters are significantly different from 0 and/or if the estimates of the parameters which are common to both the models are significantly different, then the former model might have not been a good fit. One way of choosing the bigger model is to see what the residual analysis indicates. For example, if the ACF of the residuals after fitting an MA(1) model show significant correlation at lag 2, then MA(2) model should be chosen instead of ARMA(1,1).

3.6 Forecasting

3.6.1 Minimum Mean Square Error Forecasting

If the time series $Y_1, Y_2, ...Y_t$ is given, then the minimum mean square error forecast of Y_{t+l} which is l time units after t, is,

$$\hat{Y}_t(l) = E(Y_{t+l}|Y_1, Y_2.....Y_t) \tag{3.118}$$

3.6.2 Forecasting ARIMA Models

AR(1) Models

In case of an AR(1) process with non-zero mean, to forecast l time units into the future, we have from (3.99) and (3.118),

$$\begin{aligned} Y_{t+l} - \mu &= \phi(Y_{t+l-1} - \mu) + e_{t+l} \\ E(Y_{t+l}|Y_1, Y_2....Y_t) - \mu &= \phi[E(Y_{t+l-1}|Y_1, Y_2....Y_t) - \mu] + E(e_{t+l}|Y_1, Y_2....Y_t) \\ \hat{Y}_t(l) - \mu &= \phi[\hat{Y}_t(l-1) - \mu] + 0 \\ \hat{Y}_t(l) &= \mu + \phi[\hat{Y}_t(l-1) - \mu] \qquad \texttt{for } l \geqslant 1 \end{aligned} \tag{3.119}$$

From (3.119), we can see that we can make forecasts upto any lead time by recursively forecasting for smaller lead times. This equation is also known as the difference equation form of the forecasts. To get a more explicit expression for $\hat{Y}_t(l)$,

$$\begin{aligned} \hat{Y}_t(l) &= \phi[\hat{Y}_t(l-1) - \mu] + \mu \\ &= \phi\{\phi[\hat{Y}_t(l-2) - \mu]\} + \mu \\ &\quad . \\ &\quad . \\ &\quad . \\ &= \phi^{l-1}[\hat{Y}_t(1) - \mu] + \mu \\ &= \phi^l(Y_t - \mu) + \mu \end{aligned} \tag{3.120}$$

As $|\phi| < 1$,

$$\hat{Y}_t(l) = \mu \qquad \texttt{for large l} \tag{3.121}$$

A one-step ahead forecast error will be given by,

$$\begin{aligned} e_t(1) &= Y_{t+1} - \hat{Y}_t(1) \\ &= [\phi(Y_t - \mu) + \mu + e_{t+1}] - [\phi(Y_t - \mu) + \mu] \\ \therefore e_t(1) &= e_{t+1} \end{aligned} \tag{3.122}$$

$$\therefore Var(e_t(1)) = \sigma_e^2 \tag{3.123}$$

In general linear process form, an AR(1) model can be written as follows:

$$Y_t = e_t + \phi e_{t-1} + \phi^2 e_{t-2} + ... \tag{3.124}$$

Now, the l step ahead forecast error would be:

$$\begin{aligned} e_t(l) &= Y_{t+l} - \hat{Y}_t(l) \\ &= Y_{t+l} - \mu - \phi^l(Y_t - \mu) \\ &= e_{t+l} + \phi e_{t+l-1} + \phi^2 e_{t+l-2} + ... + \phi^{l-1} e_{t+1} + \phi^l e_t + ... - \phi^l(e_t + \phi e_{t-1} + \phi^2 e_{t-2} + ...) \\ \therefore e_t(l) &= e_{t+l} + \phi e_{t+l-1} + ... + \phi^{l-1} e_{t+1} \end{aligned} \tag{3.125}$$

(3.125) can also be written as,

$$e_t(l) = e_{t+l} + \Psi_1 e_{t+l-1} + ... + \Psi_{l-1} e_{t+1} \tag{3.126}$$

Here, $E[e_t(l)] = 0$ and,

$$Var(e_t(l)) = \sigma_e^2(1 + \Psi_1^2 + \Psi_2^2 + ... + \Psi_{l-1}^2) \tag{3.127}$$

Therefore, as the lead increases, the forecast error also increases.

MA(1) Models
In case of an MA(1) model with non-zero mean, to forecast 1 time unit into the future, we have,

$$\begin{aligned} Y_{t+1} &= \mu + e_{t+1} - \theta e_t \\ E[Y_{t+1}|Y_1, Y_2, ...Y_t] &= \mu + 0 - \theta E[e_t|Y_1, Y_2, ...Y_t] \\ \hat{Y}_t(1) &= \mu - \theta e_t \end{aligned} \tag{3.128}$$

Then, the one step ahead forecast error is,

$$\begin{aligned} e_t(1) &= Y_{t+1} - \hat{Y}_t(1) \\ &= (\mu + e_{t+1} - \theta e_t) - (\mu - \theta e_t) \\ &= e_{t+1} \end{aligned}$$

To forecast l time unit into the future, we have,

$$\begin{aligned} Y_{t+l} &= \mu + e_{t+l} - \theta e_{t+l-1} \\ E[Y_{t+l}|Y_1, Y_2, ...Y_t] &= \mu + 0 - 0 \\ \hat{Y}_t(1) &= \mu \qquad \texttt{for } l > 1 \end{aligned} \tag{3.129}$$

ARMA(p,q) Models
The difference equation form for forecasting of general ARMA(p,q) model is given by,

$$\begin{aligned} \hat{Y}_t(l) &= \phi_1 \hat{Y}_t(l-1) + \phi_2 \hat{Y}_t(l-2) + ... + \phi_p \hat{Y}_t(l-p) + \theta_0 - \theta_1 E(e_{t+l-1}|Y_1, Y_2, ...Y_t) \\ &\quad -\theta_2 E(e_{t+l-2}|Y_1, Y_2, ...Y_t) - ... - \theta_q E(e_{t+l-q}|Y_1, Y_2, ...Y_t) \end{aligned} \tag{3.130}$$

Here,

$$E(e_{t+j}|Y_1, Y_2, ...Y_t) = \begin{cases} 0, & \text{for } j > 0 \\ e_{t+j}, & \text{for } j \leqslant 0 \end{cases} \tag{3.131}$$

In case the models are invertible, e_t can be written as a linear combination of the infinite sequence $Y_t, Y_{t-1}, Y_{t-2}, ...$, using π-weights. However, as j increases, the π-weights die out exponentially fast. In fact, for $j > t - q$, π_j is assumed to be negligible.
For leads=1,2,....q, in (3.130), the noise terms $e_{t-(q-1)}, ...e_{t-1}, e_t$ already exist. However, for $l > q$, the autoregressive portion and the mean θ_0 determine the general nature of the forecast.

$$\hat{Y}_t(l) = \phi_1 \hat{Y}_t(l-1) + \phi_2 \hat{Y}_t(l-2) + ... + \phi_p \hat{Y}_t(l-p) + \theta_0 \qquad \texttt{for } l > q \tag{3.132}$$

Moreover, $\theta_0 = \mu(1 - \phi_1 - \phi_2 - ..... - \phi_p)$. Therefore, (3.132) can be written as:

$$\hat{Y}_t(l) - \mu = \phi_1[\hat{Y}_t(l-1) - \mu] + \phi_2[\hat{Y}_t(l-2) - \mu] + ... + \phi_p[\hat{Y}_t(l-p) - \mu] \qquad \texttt{for } l > q \quad (3.133)$$

As l increases, $\hat{Y}_t(l) - \mu$ decays to zero for any stationary ARMA model and the long term forecast then gives the process mean, μ.
An ARIMA model can be written as follows:

$$Y_{t+l} = C_t(l) + I_t(l) \qquad \texttt{for } l > 1 \quad (3.134)$$

Here,
$C_t(l)$ = a certain function of $Y_t, Y_{t-1}, Y_{t-2}, ...$ and,

$$I_t(l) = e_{t+l} + \Psi_1 e_{t+l-1} + \Psi_2 e_{t+l-2} + ... + \Psi_{l-1} e_{t+1} \qquad \texttt{for } l \geqslant 1 \quad (3.135)$$

Then,

$$\begin{aligned} \hat{Y}_t(l) &= E(C_t(l)|Y_1, Y_2, ...Y_t) + E(I_t(l)|Y_1, Y_2, ...Y_t) \\ &= C_t(l) \end{aligned}$$

And,

$$\begin{aligned} e_t(l) &= Y_{t+l} - \hat{Y_t(l)} \\ &= C_t(l) + I_t(l) - C_t(l) \\ &= I_t(l) \\ &= e_{t+l} + \Psi_1 e_{t+l-1} + \Psi_2 e_{t+l-2} + ... + \Psi_{l-1} e_{t+1} \end{aligned}$$

So, for general ARIMA process, we can write,

$$E(e_t(l)) = 0 \qquad \texttt{for } l \geqslant 1 \quad (3.136)$$

And,

$$Var(e_t(l)) = \sigma_e^2 \sum_{j=0}^{l-1} \Psi_j^2 \qquad \texttt{for } l \geqslant 1 \quad (3.137)$$

Non-stationary Models
Using an ARMA(p+1,q) model, we can express an ARIMA(p,1,q) model as follows:

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + ... + \varphi_p Y_{t-p} + \varphi_{p+1} Y_{t-p-1} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ..... - \theta_q e_{t-q} \quad (3.138)$$

Here,

$$\left.\begin{aligned} &\varphi_1 = 1 + \phi_1, \varphi_j = \phi_j - \phi_{j-1} \qquad \texttt{for } j = 1, 2, ...p \\ &\varphi_{p+1} = -\phi_p \end{aligned}\right\} \quad (3.139)$$

If the order differencing of the ARIMA model is d, then there will be $p + d$ such φ coefficients. Using (3.130) and (3.131), and replacing the ps with $(p+d)$s and ϕ_js by φ_js, we can do the forecasting.

3.6.3 Limits of Prediction

In a general ARIMA series, if the white noise terms $\{e_t\}$ are independently and identically distributed, then the forecast error, $e_t(l)$, will also be normally distributed. That is,

$$e_t(l) = Y_{t+l} - \hat{Y}_t(l)$$

will be normally distributed. Therefore, using the standard normal percentile, $Z_{1-\frac{\alpha}{2}}$, it can be stated that,

$$P[-Z_{1-\frac{\alpha}{2}} < \frac{Y_{t+l} - \hat{Y}_t(l)}{\sqrt{Var(e_t(l))}} < Z_{1-\frac{\alpha}{2}}] = 1 - \alpha$$

Here, $(1-\alpha)$ is the given level of confidence. The above equation can also be written as follows:

$$P[\hat{Y}_t(l) - Z_{1-\frac{\alpha}{2}}\sqrt{Var(e_t(l))} < Y_{t+l} < \hat{Y}_t(l) + Z_{1-\frac{\alpha}{2}}\sqrt{Var(e_t(l))}] = 1 - \alpha$$

Hence, it can be said with $(1-\alpha)100\%$ confidence, that the forecast of Y_{t+l} will be within the limits:

$$\hat{Y}_t(l) \pm Z_{1-\frac{\alpha}{2}}\sqrt{Var(e_t(l))} \tag{3.140}$$

3.6.4 Updating ARIMA Forecasts

Say, we have made a forecast, $l+1$ steps into the future and so we have $\hat{Y}_t(l+1)$. Now, once we get the observation of the next time unit, that is, $t = t+1$, we would like to update our forecast with the origin at $t = t+1$, that is, $\hat{Y_{t+1}}(l)$. Therefore, using (3.134) and (3.135), we get,

$$\begin{aligned} Y_{t+l+1} &= C_t(l+1) + e_{t+l+1} + \Psi_1 e_{t+l} + \Psi_2 e_{t+l-1} + ... + \Psi_l e_{t+1} \\ \hat{Y_{t+1}}(l) &= C_t + \Psi_l e_{t+l} \\ \hat{Y_{t+1}}(l) &= \hat{Y}_t(l+1) + \Psi_l[Y_{t+1} - \hat{Y}_t(1)] \end{aligned} \tag{3.141}$$

3.6.5 Forecasting Transformed Series

Transformation by Differencing
In case the transformations were done by differencing, then two approaches could be used to forecast them:

1. Forecasting the original non-stationary series using the difference equation form and replacing the ϕs with φs.
2. Forecasting the stationary differenced series first and then summing the series to undo the differencing.

Log Transformations
If log transformations were done on the original time series Y_t to get $Z_t = log(Y_t)$, then the minimum mean square error forecast on the original series is expressed as,

$$\exp\{\hat{Z}_t(l) + \frac{1}{2}Var[e_t(l)]\} \tag{3.142}$$

This only works properly if the variables are normally distributed. However, if Z_t itself has a normal distibution, then a different method would be preferred.

Chapter 4

Stock Returns Forecasting System

In this chapter, we discuss about our proposed web-based system for stock returns forecasting. We also explain how we implemented the system and the challenges we faced while doing so.

4.1 Proposed System

The development of the stock returns forecasting system is aimed towards aiding any stock market enthusiast who wishes to use their own model for stock price forecasting. At present, the system provides facilities to fit ARIMA models of any order to the time series of stock returns of a company. We intend to include other models into the picture in the future. This system not only helps the users to do their own analysis and model-fitting, but it also provides them with useful descriptions of the different concepts to help them understand better. Therefore, even if someone has no background in time series analysis, they can use this website and get a rough idea of what is going on.
A good advantage of such a system is that it does not require any kind of installation or coding knowledge to perform the analysis. Users can access it anytime and anywhere as long as there is an internet connection. They can also download the outputs of their analysis for future use.

4.1.1 Why use stock returns data instead of stock prices?

In order to fit ARIMA models to any time series, the first and foremost condition is that the time series has to be stationary. Unfortunately, the time series of stock market prices almost always rejects the Augmented Dickey Fuller Unit Root Test (See Section 3.5.1 for details on page 37). That is, they are almost always non-stationary in nature. Taking the first difference of the logarithm of the time series of stock market prices leads to a stationary time series in most cases. Even if it is not stationary, it can be turned into one by differencing it further. Therefore, for our convenience, instead of using time series of stock prices, we use the stock returns time series which is calculated using the following formula:

$$r_t = 100(log(p_t) - log(p_{t-1}))$$

We multiply by 100 because otherwise, it would have resulted in round-off errors as the returns values are too small.

4.1.2 How the Proposed System Works

- Users choose whether they wish to analyze a particular stock returns data or to fit ARIMA models to them.
- If they choose to analyze the data
  - They are asked to select a particular stock symbol from the dropdown list and a past time interval. If the time interval is less than ten trading days, we reject their input and ask them to choose a bigger time interval. We also ask them if they would like to perform differencing on their selected series. We set the limit of differencing to 3 as more differencing of the series is not recommended. On the right of the input form, they can see what outputs they will get after they click the Analyze button along with their description.
  - When they give the required inputs and click the Analyze button, the inputs are sent to the server. The R software in the server extracts the dataset from the database and executes several functions on them to generate the outputs. The outputs are then saved in a directory of the server by R, which are then fetched and shown to the users on the website. Users also get the choice of downloading the generated outputs.
- If they choose to fit models to the data
  - They are asked to select a particular stock symbol from the dropdown list, a past time interval, and the orders of the ARIMA model they wish to fit. If the time interval is less than ten trading days or less than the selected orders of the model, we reject their input and ask them to give proper inputs. To fit the models, we also give them the choice to select a method for parameter estimation and to select whether they wish to fit the model considering the series to have a non-zero mean. By default, these are set to "CSS-ML" and "No mean" respectively. Here, CSS stands for Conditional Sum of Squares and ML stands for Maximum Likelihood. The mean is not included by default because when the series is being used to fit an ARIMA model, we expect that it would be a stationary series with no mean. On the right of the input form, they can see what outputs they will get after they click the Fit Model button along with their description.
  - When they click the Fit Model button, similar steps are executed as the Analysis section. Only the functions executed by R and the generated outputs are different.

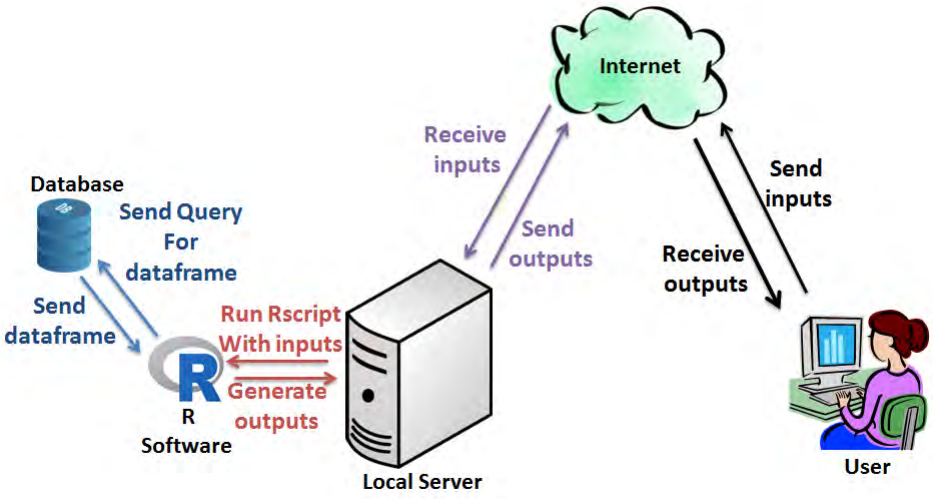

Figure 4.1: An overview of the System

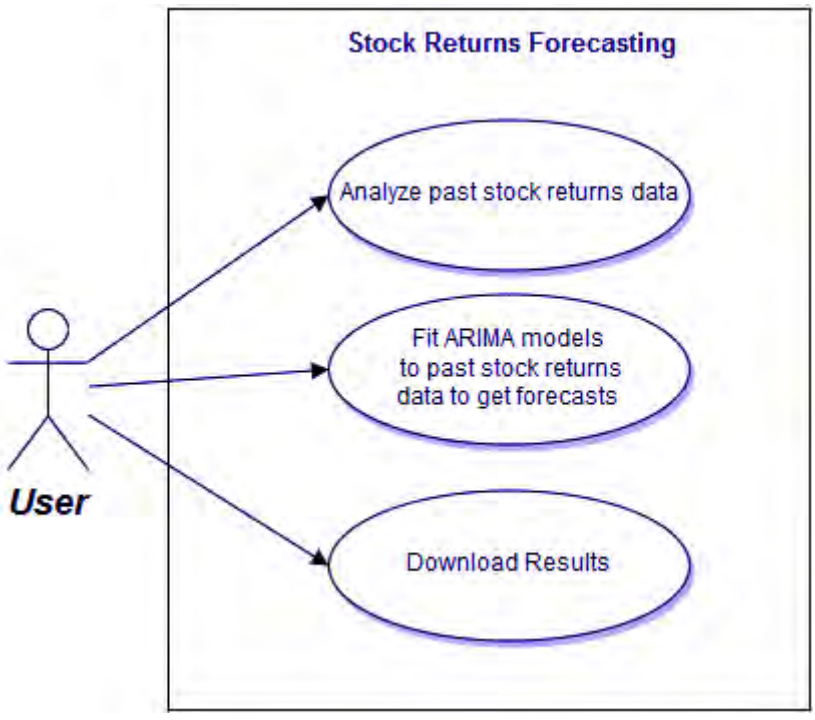

Figure 4.2: Usecase Diagram of the System

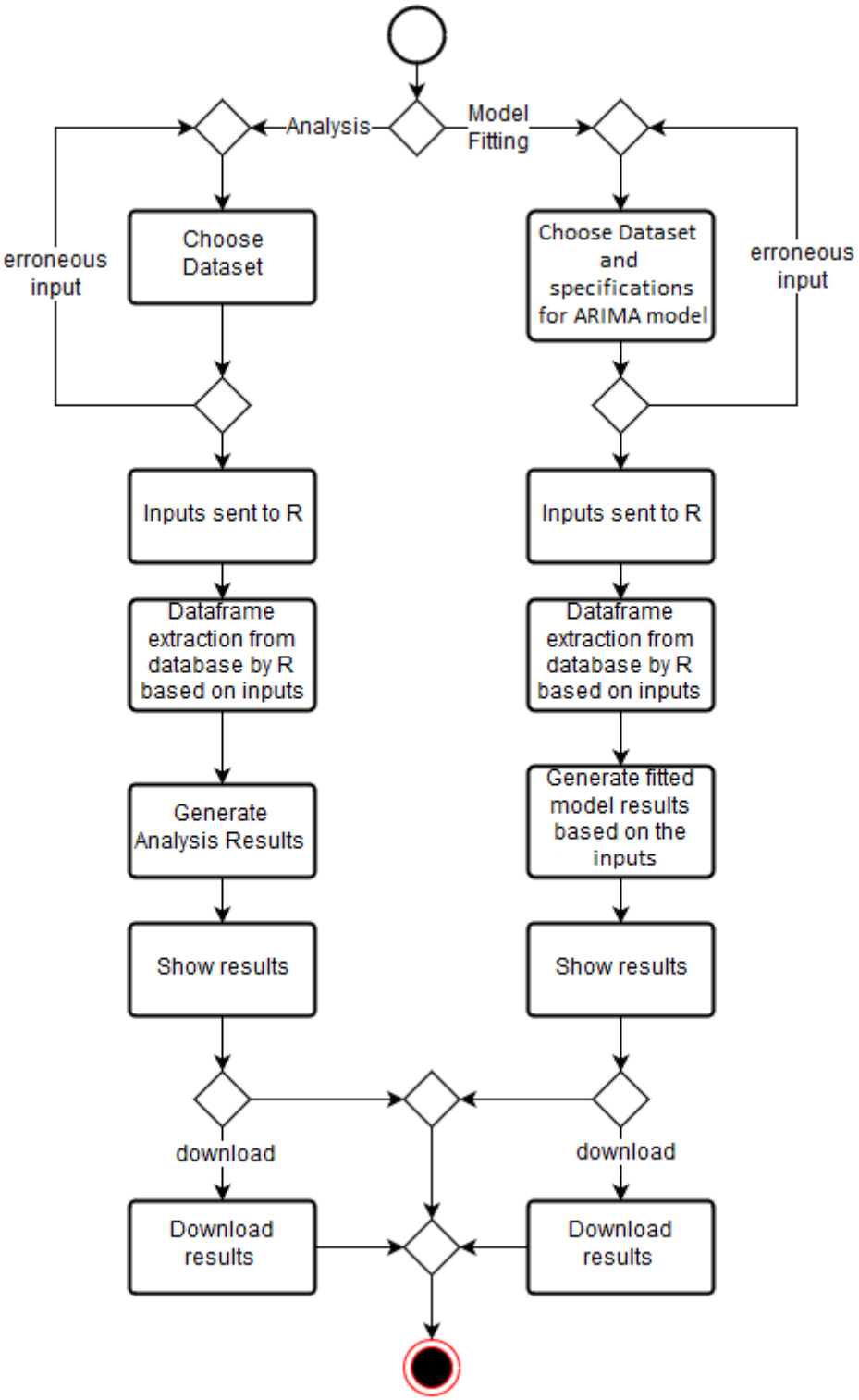

Figure 4.3: An Activity Diagram of the System

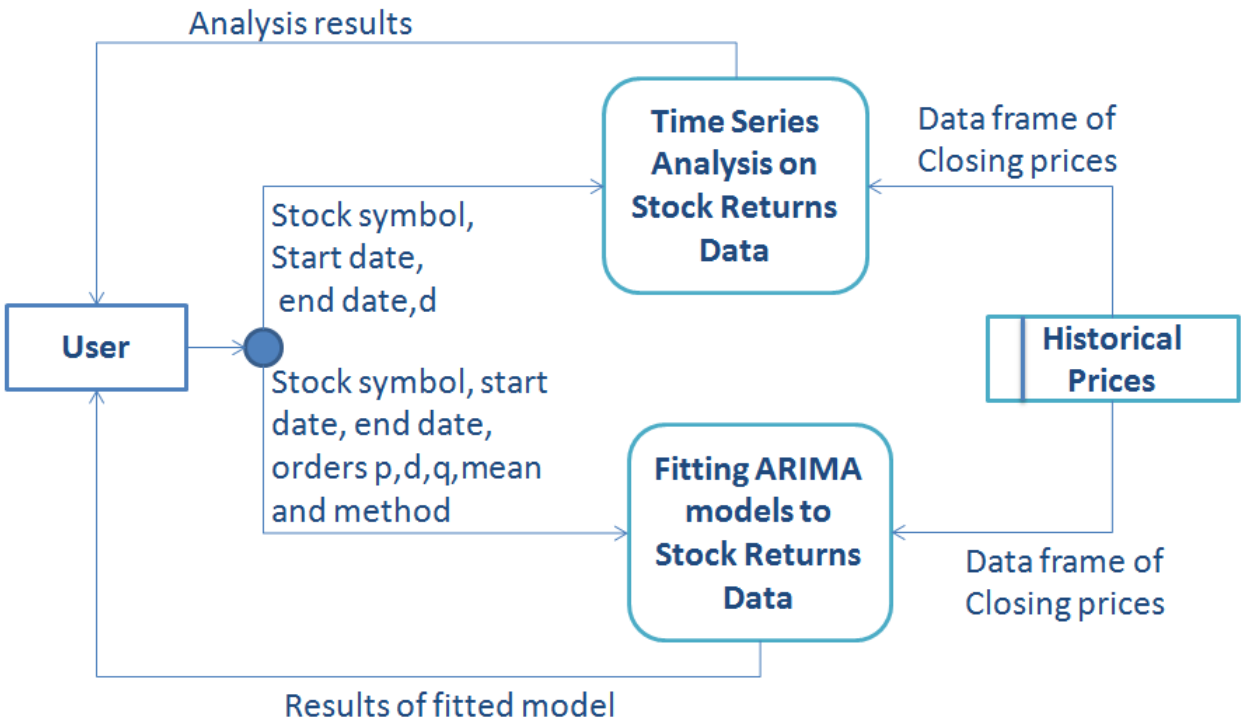

Figure 4.4: A Data Flow Diagram of the System

4.2 System Outputs

In this section, we use the time series of YHOO stock closing prices for the first half of 2016 and explain the different outputs provided by the system. We also mention the different functions of R that we used to get those results.

4.2.1 Outputs of Analysis Section

In this section, the stock returns are calculated using the closing prices of the YHOO stock for the first half of the year 2016 with the help of the formula mentioned in Section 4.1.1 and then they are used as a time series to give the following outputs:

Plot of Stock Returns
The stock returns are plotted (Figure 4.5) as time series using the plot(ts()) function of R. Plotting the time series helps us to have a rough idea about its nature. From the figure, it seems as if the series has constant mean and variance.

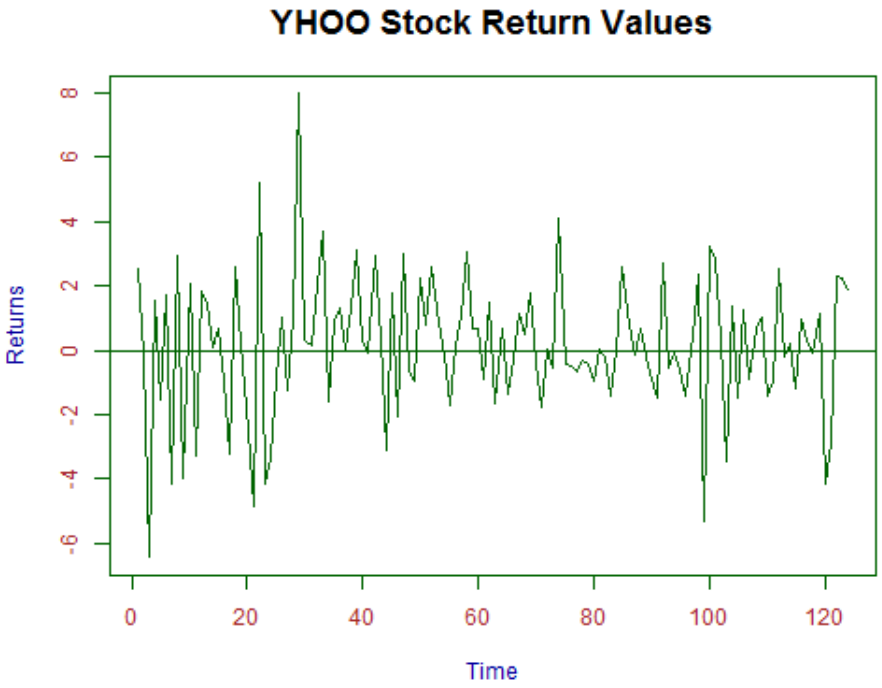

Figure 4.5: Time Series Plot of Stock Returns of YHOO

Summary of Stock Returns

```
Summary of YHOO stock returns
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-6.4210 -0.9612  0.1712  0.1445  1.3840  7.9590
```

Figure 4.6: Summary of Stock Returns Data of YHOO

A summary of the time series data (Figure 4.6) is provided using the summary() function of R, where,

- Min minimum value in the dataset
- 1st Qu. First Quartile 25% of the values are below the given quantity
- Median The median value of the dataset
- 3rd Qu. Third Quartile 75% of the values are below the given quantity
- Max maximum value in the dataset

Plot of ACF

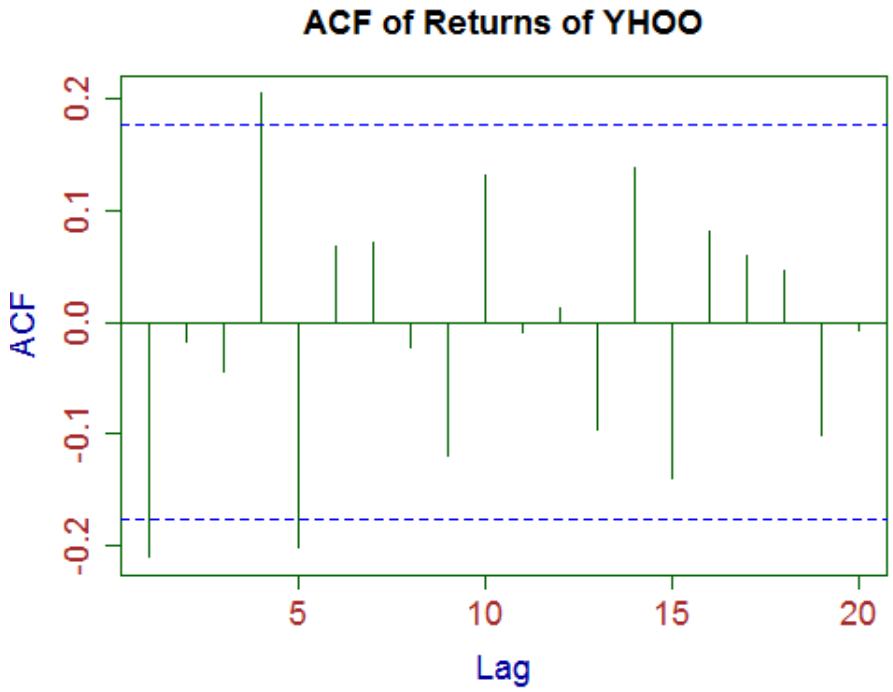

Figure 4.7: Sample ACF Plot of Stock Returns Data of YHOO

The Autocorrelation Function (ACF) of the stock returns time series is plotted using the acf() function of R. The plot of ACF helps us to identify the order of the pure MA(q) model of a time series. Starting from 0, the lag after which the ACF stops crossing the significance bound (blue dashed line), is the order, q, of the MA(q) model. The significance bound is set at $\pm\frac{2}{\sqrt{n}}$, where n is the length of the series. If the ACF does not cross the significance bound in the first lag, but does so in case of later lags, then we assume that q=0. From the plot Figure 4.7, we can see that the ACF at lag 1 has crossed the significance bound. So, we can assume that the series has an MA(1) process in it.

Plot of PACF

The Partial Autocorrelation Function (PACF) of the stock returns time series is plotted using the pacf() function of R. The plot of PACF helps us to identify the order of the pure AR(p) model of a time series. Starting from 0, the lag after which the PACF stops crossing the significance bound (blue dashed line), is the order, p, of the AR(p) model. The significance bound is set at $\pm\frac{2}{\sqrt{n}}$, where n is the length of the series. If the PACF does not cross the significance bound in the first lag, but does so in case of later lags, then we assume that q=0. From the plot in Figure 4.8, we can see that the PACF at lag 1 crosses the significance bound and so, we can assume that it has an AR(1) process in it.

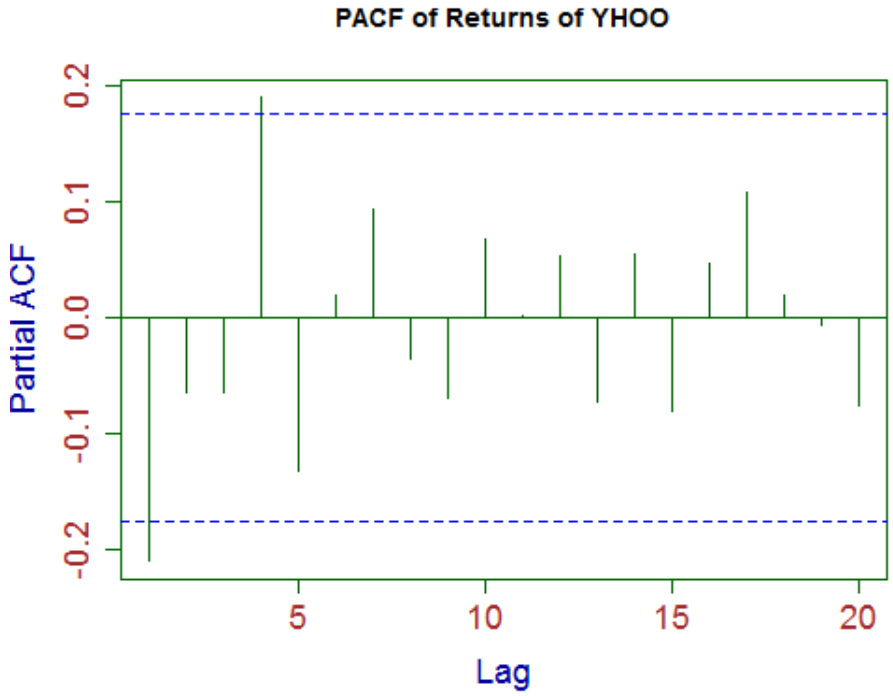

Figure 4.8: Sample PACF Plot of Stock Returns Data of YHOO

EACF

We get the Extended Autocorrelation Function (EACF) by using the eacf() function in R. For a mixed ARMA(p,q) model, Extended Autocorrelation Function (EACF) helps us to identify the possible values of p and q of an ARMA model of a time series. Let, the AR order be k and the MA order be j. Then, in the $symbol table of the output, the element in the k-th row and j-th column is set to x if for AR order k, the lag j+1 ACF is significantly different from zero. Otherwise, it is set to o. The trick to understand what it means is to look for a triangle of zeroes in the $symbol table. The upper left-hand vertex of the triangle will indicate the order of the ARMA(p,q) model. In our case (Figure 4.9), the EACF table does not look too clear. As an exact triangle has not been formed anywhere. So, we can try fitting ARMA(0,5),ARMA(1,1),ARMA(2,1), ARMA(3,3) and ARMA(4,3) on the series and choose the one which best fits the model based on the AIC, AICc or BIC values.

Q-Q Plot

The Quantile-Quantile plot or the Q-Q plot is plotted using the qqnorm() and qqline() functions of R. This plot helps us to find whether a time series is normally distributed or not. If the plot of the values looks like a straight line, then we can say that the series is normally distributed. From the Figure 4.10, we can see that the most of the values seem to align with the straight line in the middle and then they move away at the two ends. We can say that it is somewhat normally distributed.

```
EACF of Returns
$eacf
          [,1]       [,2]        [,3]        [,4]       [,5]
[1,] -0.2099945 -0.01796287 -0.043671257  0.204495568 -0.20152316
[2,] -0.2915203  0.06499715  0.008868365  0.119980593 -0.14549746
[3,] -0.5006120 -0.05738287  0.028147393  0.107567540  0.02084675
[4,]  0.2561344  0.24244278  0.219140937  0.092386899  0.01045093
[5,]  0.4493516  0.23166189 -0.227738418 -0.105587133  0.04623187
[6,]  0.1024771  0.19932371 -0.160625349 -0.256871506  0.19302609
[7,] -0.1483395  0.45338962  0.128461699 -0.238821426  0.15092354
[8,]  0.2559312  0.44643507  0.166636943 -0.007522336  0.01935079
           [,6]        [,7]        [,8]       [,9]        [,10]
[1,]  0.0674850843  0.071319393 -0.021643369 -0.12004831  0.132404652
[2,]  0.1185196226  0.086815793 -0.012701136 -0.07463142  0.146692063
[3,]  0.1453355153 -0.060070818  0.049139544 -0.02297846  0.149893317
[4,]  0.0569191985  0.028228214  0.004946821  0.01827108  0.079396304
[5,] -0.0024412899  0.004046366  0.006544267  0.02215228  0.078772124
[6,] -0.0132572627 -0.001104642  0.006894132  0.03817702  0.040869496
[7,]  0.0008571225 -0.001163262  0.013220497  0.03630442  0.048514337
[8,] -0.0048571297 -0.033952787  0.015681496  0.02824216 -0.004435138
          [,11]       [,12]       [,13]       [,14]
[1,] -0.009057778  0.012624312 -0.09633422  0.138213075
[2,]  0.028275100 -0.012009203 -0.03431644  0.030146791
[3,]  0.017593423 -0.020830879 -0.03731650  0.006533043
[4,]  0.045332986 -0.006686095 -0.05980045  0.002705704
[5,]  0.052356263 -0.007529453 -0.03784949  0.049220673
[6,] -0.029238181  0.023845332  0.01196169  0.051103561
[7,]  0.010752604  0.024362802 -0.01235748  0.078072269
[8,] -0.020001347  0.014940502 -0.05409412 -0.024017071

$ar.max
[1] 8

$ma.ma
[1] 14

 $symbol
   0   1   2   3   4   5   6   7   8   9   10  11  12  13
 0 "x" "o" "o" "x" "x" "o" "o" "o" "o" "o" "o" "o" "o" "o"
 1 "x" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o"
 2 "x" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o"
 3 "x" "x" "x" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o"
 4 "x" "x" "x" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o"
 5 "o" "x" "o" "x" "x" "o" "o" "o" "o" "o" "o" "o" "o" "o"
 6 "o" "x" "o" "x" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o"
 7 "x" "x" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o"
```

Figure 4.9: EACF of Stock Returns Data of YHOO

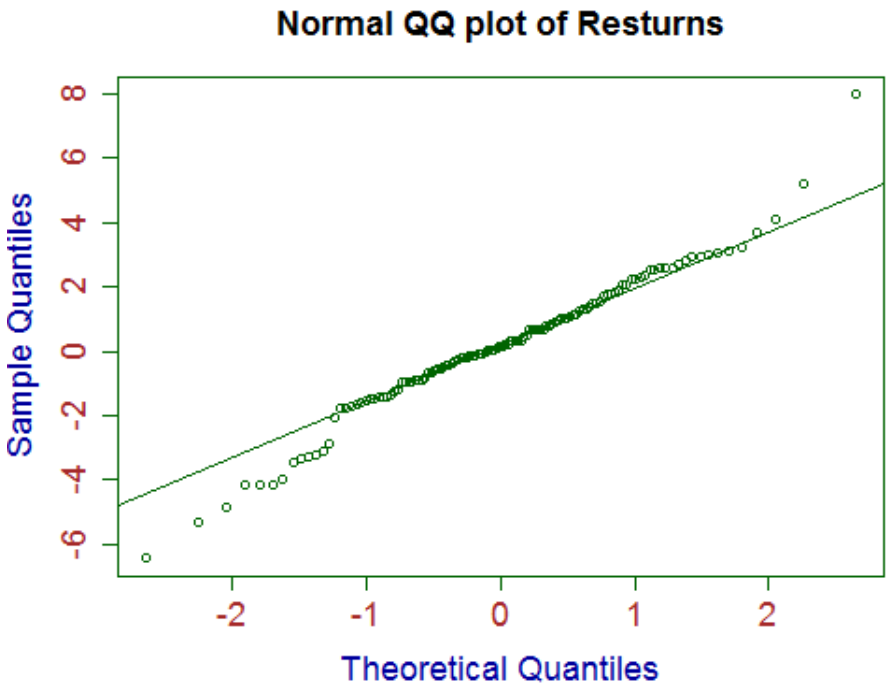

Figure 4.10: Q-Q Plot of Stock Returns Data of YHOO

Histogram Plot

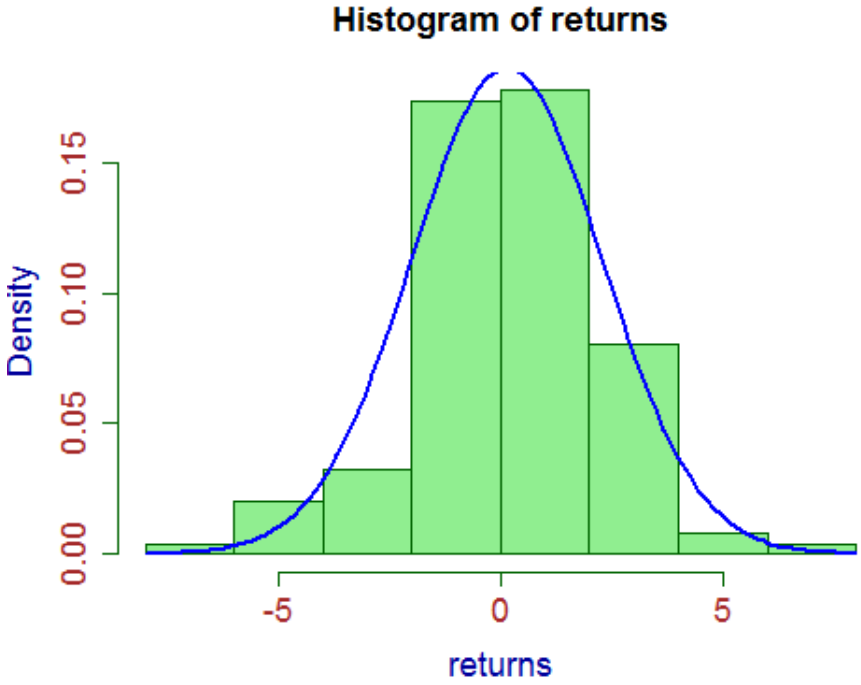

Figure 4.11: Histogram Plot of Stock Returns Data of YHOO

The Histogram of probability densities of the returns series is plotted using the hist() function of R. Along with it, we also drew a curve which represents the theoretical normal distribution of the series. the This plot helps us to find whether a time series is normally distributed or not. The plot will be somewhat symmetric and tail off at both the high and low ends as a normal distribution would. From the Figure 4.11, we can say that the series roughly follows normal distribution.

ADF Test

```
Augmented Dickey Fuller Test on Returns

        Augmented Dickey-Fuller Test

data:  returns
Dickey-Fuller = -5.1127, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

Figure 4.12: ADF Test Results of Stock Returns Data of YHOO

The Augmented Dickey Fuller Test (Figure 4.12) is run on the returns series using the adf.test() function of R. This test helps us to find whether a time series is stationary or

not. If the probability value i.e. p.value becomes greater than 0.1, then we cannot reject the null hypothesis which states that the series is non-stationary. From the output of our example (Figure 4.12), we can say that the series is stationary as the p.value is not greater than 0.1.

4.2.2 Outputs of Model-Fitting

In this section, we take the same inputs as the analysis section and we also take inputs for the orders p,d and q of the ARIMA(p,d,q) model. We try and fit the several ARIMA models to the series that we found we could use during analysis in Section 4.2.1. We chose to consider zero mean while fitting the model and the combination of conditional sum of squares and maximum likelihood methods for parameter estimation.

ARIMA	AIC	AIC_c	BIC
(1,0,1)	540.37	540.57	548.83
(0,0,5)	539.49	540.21	556.42
(2,0,1)	542.32	542.66	553.6
(3,0,3)	537.23	538.19	556.97
(4,0,3)	541.73	542.98	564.29

Table 4.1: Values of AIC, AIC_c and BIC after fitting different models

After we tried fitting the different models, the ARIMA(3,0,3) model showed the least value for AIC_c (See Table 4.1), so we chose to show only the outputs of fitting ARIMA(3,0,3) model to our returns series. The typical outputs that are shown in our system are discussed.

Plot of Stock Returns and Summary of Stock Returns

The plot of time series of stock returns and its summary are shown as output in this section as well for reference.

Summary of Fitted Model

After the ARIMA(3,0,3) model is fitted to the series, as shown in Figure 4.13, the estimates of the coefficients and their respective standard errors are given. Also, the following are shown:

Sigma-square
The variance of the series as assumed by the fitted model.
Log Likelihood
It quantifies the relative abilities of the estimates to explain the observed data.
AIC
This is the Akaikes Information Criterion. Models with the least value of AIC should be chosen.

```
Fitted Model Summary
Series: returns
ARIMA(3,0,3) with zero mean

Coefficients:
          ar1      ar2      ar3     ma1     ma2     ma3
      -1.4069  -1.4195  -0.9071  1.2741  1.2680  0.7400
s.e.   0.1007   0.0895   0.1042  0.1508  0.1707  0.1649

sigma^2 estimated as 4.143:  log likelihood=-261.61
AIC=537.23   AICc=538.19   BIC=556.97

Training set error measures:
                    ME     RMSE      MAE      MPE     MAPE      MASE
Training set 0.1613876 1.985489 1.503944 86.86134 164.8643 0.5986001
                    ACF1
Training set -0.03934442
```

Figure 4.13: Summary of Fitted Model of Stock Returns Data of YHOO

AICc
This is the corrected form of AIC and is said to outperform both AIC and BIC in model selection. Models with the least value of AICc should be chosen.
BIC
Bayesian Information Criterion. Models with the least value of BIC should be chosen.
ME
It is the mean error of the fitted values.
RMSE
It is the root mean square error of the fitted values.
MAE
It is the mean absolute error of the fitted values.
MPE
It is the mean percentage error of the fitted values. Since it gives error in percentage, it can be used to compare models with different datasets.
MAPE
It is the mean absolute percentage error of the fitted values. Since it gives error in percentage, it can be used to compare models with different datasets.
MASE
It is the mean absolute scaled error of the fitted values. It is also used to compare models with different datasets.
ACF1
It is the first order autocorrelation coefficient. It is the correlation coefficient of the first N-1 observations and the next N-1 observations.

Plot of Residuals of Fitted Model

Plotting the residuals (Figure 4.14) helps us to have a rough idea about its nature. From the figure, it seems like the residual series has constant mean and variance.

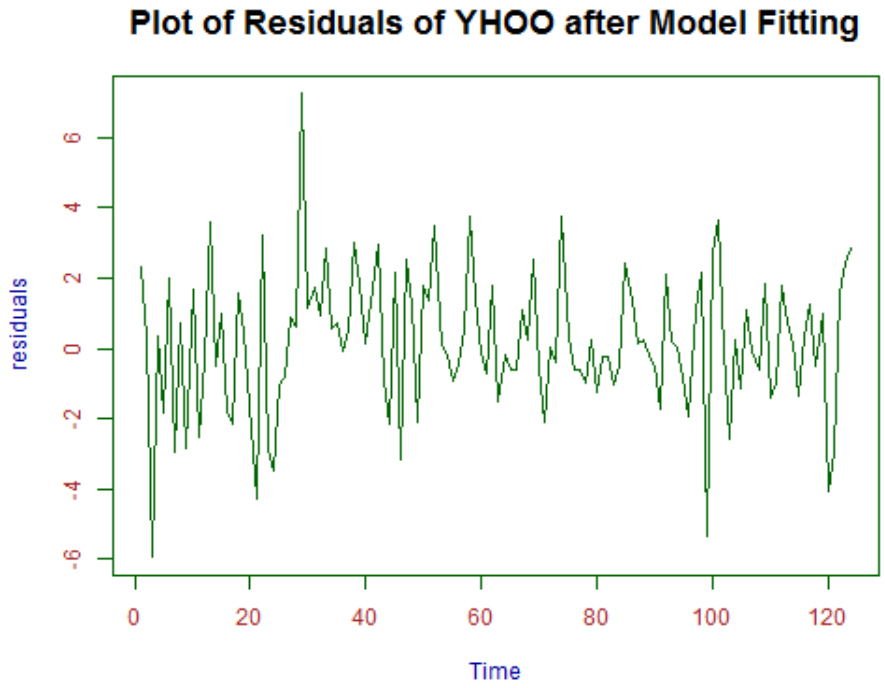

Figure 4.14: Time Series Plot of Residuals of Fitted Model on YHOO Stock Returns

ACF Plot of Residuals

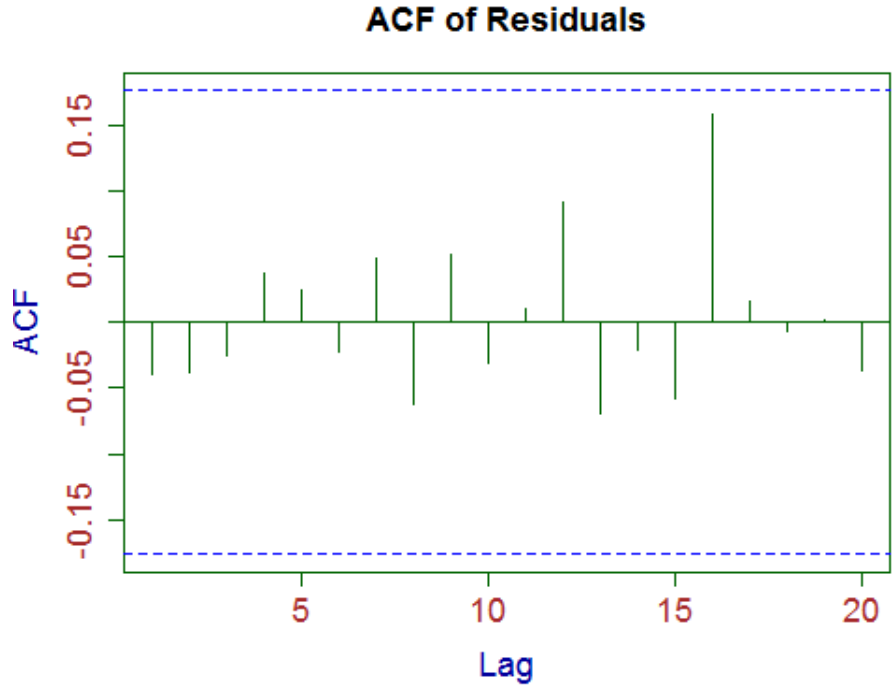

Figure 4.15: ACF Plot of Residuals of Fitted Model on YHOO Stock Returns

The ACF plot (Figure 4.15) in the figure shows that the residual series has no MA(q) process in it as none of the ACFs crossed the significance bound.

PACF Plot of Residuals

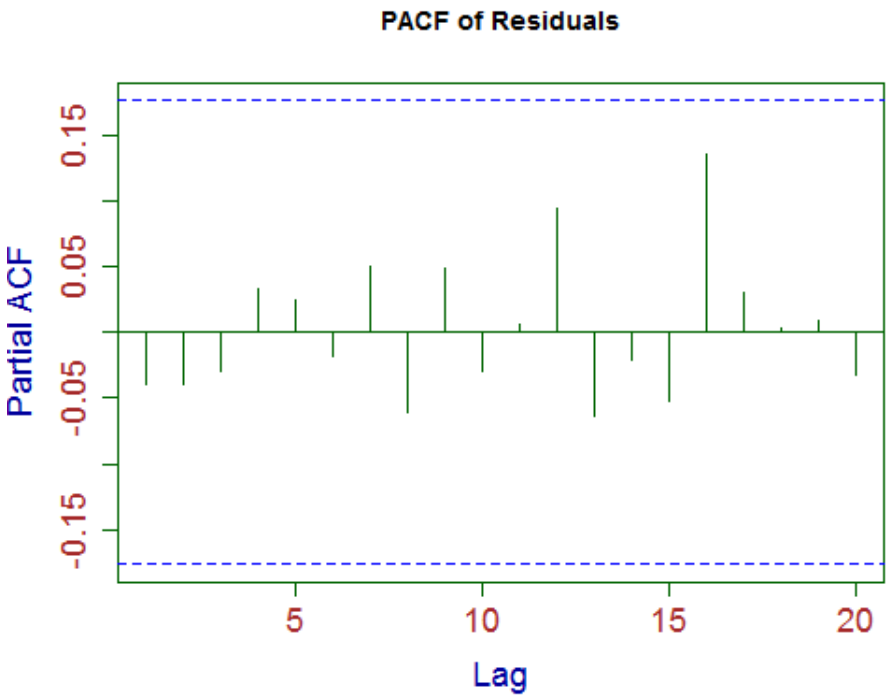

Figure 4.16: PACF Plot of Residuals of Fitted Model on YHOO Stock Returns

The PACF plot (Figure 4.16) in the figure shows that the residual series has no AR(p) process in it as none of the PACFs crossed the significance bound.

Q-Q Plot of Residuals

The Q-Q plot (Figure 4.17) of the residuals shows that the residuals are normally distributed to a great extent.

Ljung-Box Test

Ljung-Box Test was run on the residuals using the Box.test() (Figure 4.18) function of R with the default lag which is 1. This test allows us to find whether error terms are correlated or not. If p.value > 0.05, then we cannot reject the null hypothesis that the error terms are uncorrelated. Since the p.value here is 0.6613, we cannot reject the null hypothesis that the adjacent error terms are uncorrelated.

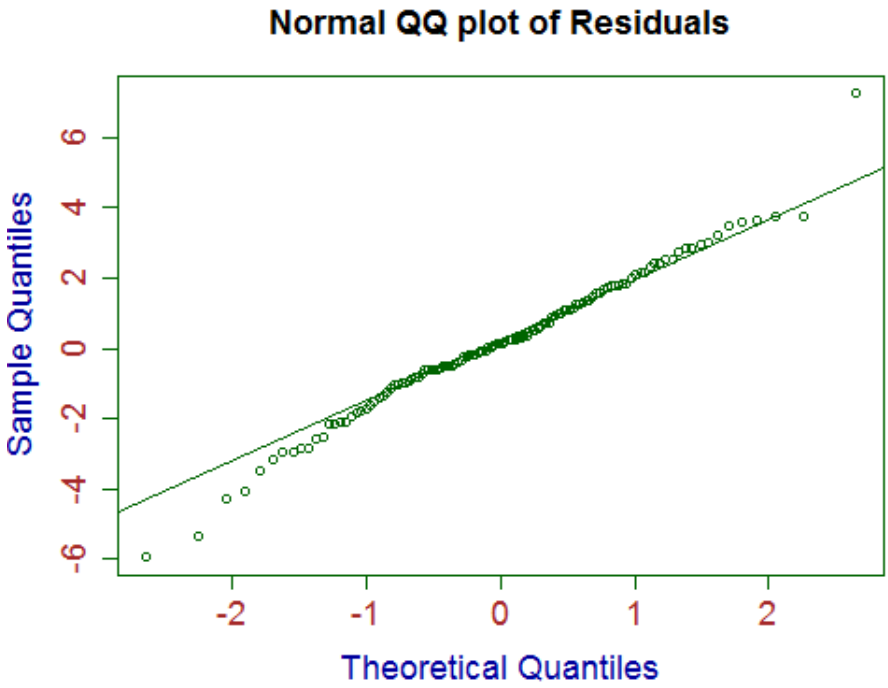

Figure 4.17: Q-Q Plot of Residuals of Fitted Model on YHOO Stock Returns

```
BOX TEST RESULT ON RESIDUALS:

        Box-Pierce test

data:  residuals
X-squared = 0.19195, df = 1, p-value = 0.6613
```

Figure 4.18: Results of Ljung-Box Test on Residuals of Fitted Model on YHOO Stock Returns

Plot of Forecast of Fitted Model

We plot the forecast of next 10 trading days along with a 80% prediction interval for the forecast and a 95% prediction interval for the forecast by using the plot.forecast() function in Figure 4.19. If the values of the next ten days are also available, then those are also plotted to show the outputs.

Forecast Values

Using the forecast.Arima() function, we get the values of forecast of next 10 trading days along with a 80% prediction interval for the forecast and a 95% prediction interval for the forecast (Figure 4.20).

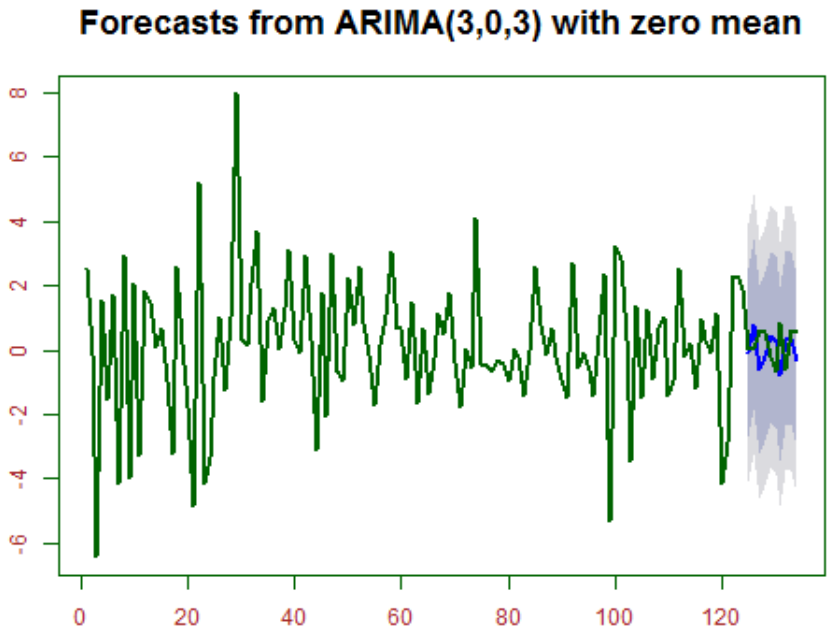

Figure 4.19: Plot of Forecast of Fitted Model on YHOO Stock Returns

```
Summary of Forecast:
    Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
125    -0.08290855 -2.691304 2.525487 -4.072105 3.906288
126     0.80066850 -1.830636 3.431973 -3.223564 4.824901
127    -0.60983579 -3.242757 2.023085 -4.636541 3.416869
128    -0.20338353 -2.837339 2.430572 -4.231671 3.824904
129     0.42551327 -2.224052 3.075079 -3.626647 4.477674
130     0.24324520 -2.433847 2.920338 -3.851015 4.337505
131    -0.76174868 -3.446156 1.922659 -4.867196 3.343698
132     0.34041599 -2.343995 3.024827 -3.765036 4.445868
133     0.38173617 -2.303345 3.066817 -3.724741 4.488214
134    -0.32929448 -3.027951 2.369362 -4.456533 3.797944
```

Figure 4.20: Summary of Forecast of Fitted Model on YHOO Stock Returns

Forecast Errors

```
Forecast errors:
                    ME      RMSE       MAE        MPE     MAPE
Training set 0.1613876 1.9854891 1.5039435   86.86134 164.8643
Test set     0.1566480 0.9086362 0.8094468 -114.76883 465.8526
                  MASE        ACF1
Training set 0.5986001 -0.03934442
Test set     0.3221763          NA
```

Figure 4.21: Errors of Forecast of Fitted Model on YHOO Stock Returns

Using the accuracy() function, we get the errors of forecast of next 10 trading days

(Figure 4.21). The training set errors are the errors faced during model fitting and the test set errors are the forecast errors. We look at MPE or the mean percentage error to determine the accuracy for the forecast. Unfortunately, in this case, the error is huge. We can try checking the other models that we had ignored previously to see if the error gets reduced. If they lead to huge errors as well, the chances of which are fairly high, then other factors or models would need to be incorporated to make a better model. More about this has been discussed in Section 5.1

4.3 System Implementation

Our system is a responsive website which was built using PHP (ver. 5.6.23), HTML 5, CSS, Javascript and MYSQL languages. We used the Xampp software to create a local server in our laptop. The database server we used was MariaDB 10.1.13. As mentioned before, R was used to perform the different analysis and to generate the results.

4.3.1 The Website

The website consists of three pages the Home page, the Analysis page and the Model-Fitting Page. All these pages were developed with the help of the templates provided by [3]. We inserted a stock ticker watch list widget in our website, which was collected from the [6] website. The watch list monitors stock quotes of 100 S&P500 companies.

The Homepage
The homepage gives a brief explanation on what facilities our site provides.

The Analysis Page
In the Analysis page, the user gets to see a form on the left where they are asked to choose a ticker name, a date interval and a differencing order. On the right, we give them an idea about what outputs they would get after the analysis and we also explain how they should interpret those results. Once they give the inputs and submit the form, the outputs are shown on the right with the descriptions at the bottom. We also give them the option to download all the results, which are .png files and .txt file.

The Model-Fitting Page
The functionality of this page is the same as the analysis page except for the fact that they insert additional inputs to fit the ARIMA models. The outputs shown here are also different.

Figure 4.22: The Homepage of the System

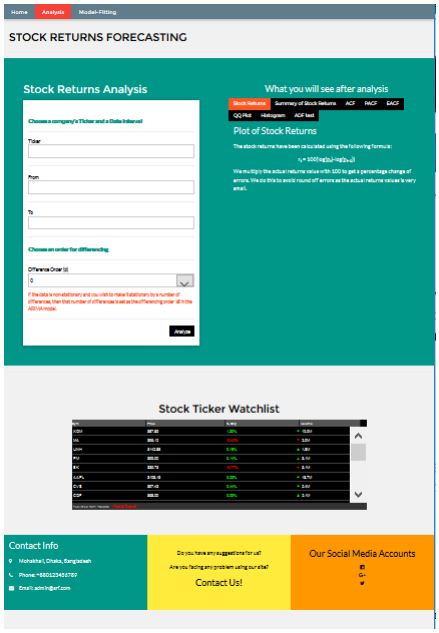

Figure 4.23: The Analysis Page Before Input is Submitted

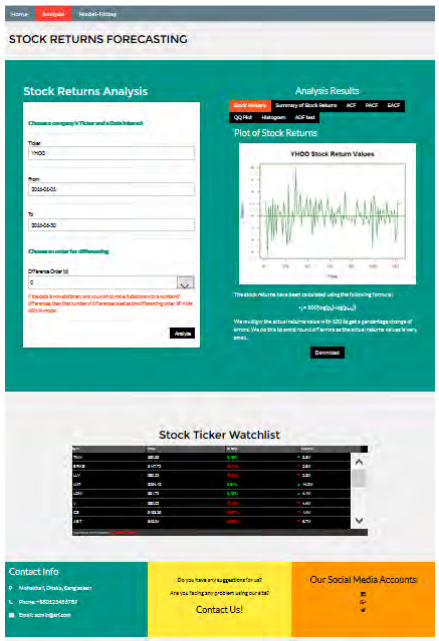

Figure 4.24: The Analysis Page After Input is Submitted

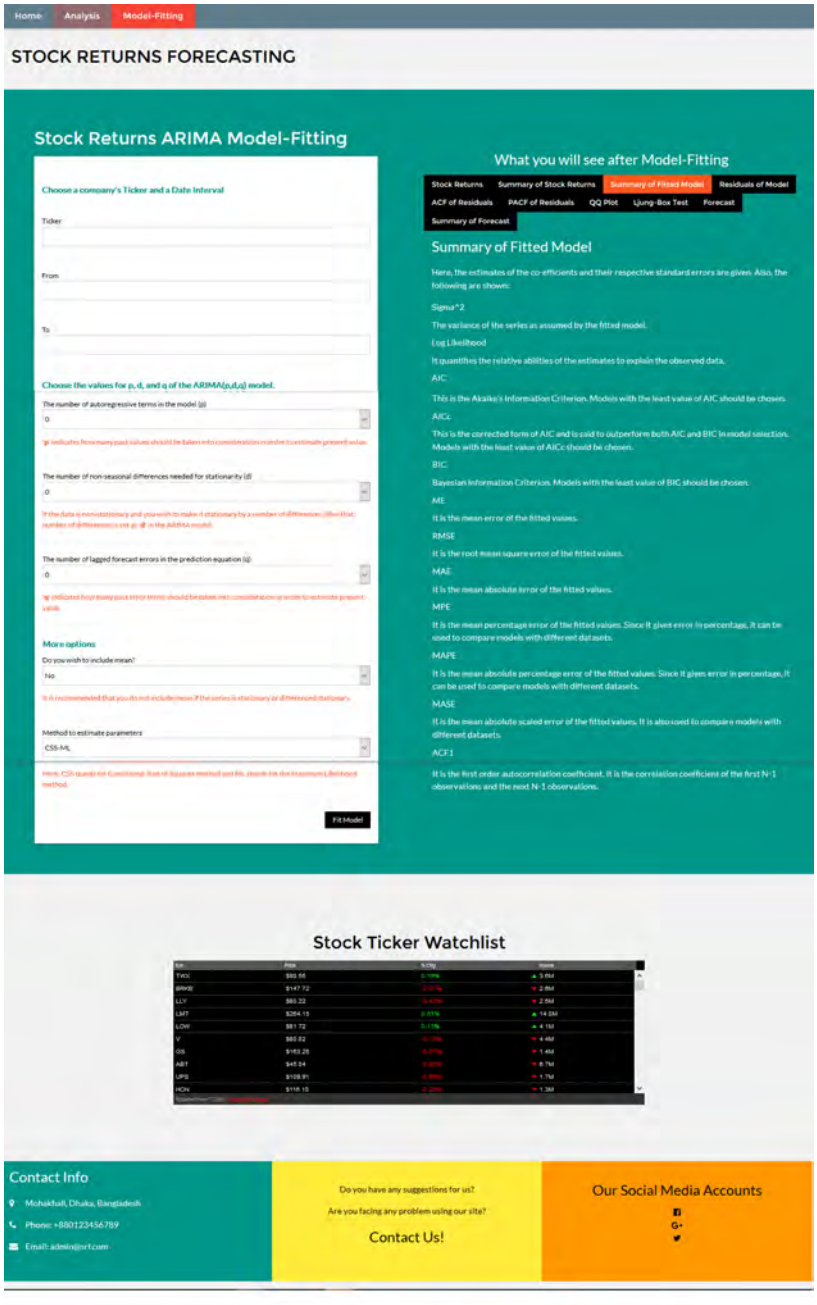

Figure 4.25: The Model-Fitting Page Before Input is Submitted

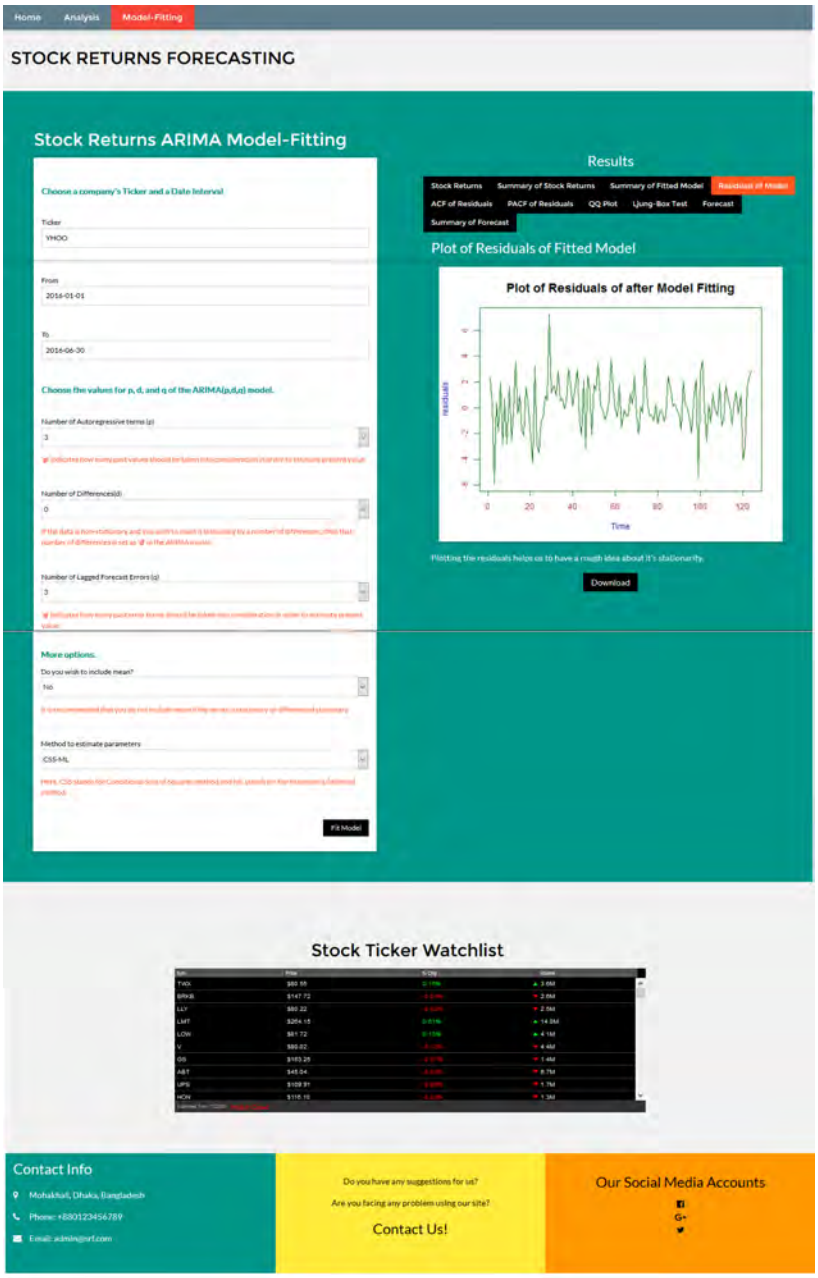

Figure 4.26: The Model-Fitting Page After Input is Submitted

4.3.2 R

R is a programming language and software environment which is widely used for statistical computations. It is not only used by statisticians to perform data analysis, but it also helps in developing statistical software. R scripts (with .R extensions) are simple text files, where we write all the commands that we want R to execute. We made three separate R scripts for our project update.R, model.R and analysis.R. We run update.R manually for now to update the database every day. The model.R and analysis.R scripts are run, depending on which page's form the user submits.
To learn the R language as well as to create the Rscript files for our project, we used the RStudio software which is an IDE for using R.

Data Frames
Data frames in R are used to store data tables. A data frame is actually a list of vectors of equal length. In our project, we used R to collect data from Yahoo Finance, to store the data in the database and to extract required data from database. All these were done by the use of data frames. The data from Yahoo Finance were collected as data frames and then the data frames were modified according to our need. We then appended the data from the data frame to the existing table in the database using R.
The data frames imported from Yahoo Finance had the dates column set as the row_names attribute. The row_names attribute is a character vector having length equal to the number of rows of the data frame. To append the data from the data frame, we used the WriteTable() function. This function creates a table in the database in case the table does not exist. Otherwise, it simple appends the data. So, the database table has the same attribute names as the data frame. We tried to change the name of the row_names attribute to Date but we could not do it. As it did not hamper our work in any way, we decided to leave it as it was.

Packages and Functions Used
R has a lot of useful functions to perform statistical analysis. However, sometimes we require some specialized functions. Rs active user community has built numerous useful specialized and these are available at the CRAN website. We can also install them directly from the RStudio.
The packages that we needed for our project are as follows:

- DBI : This package helps to build the communication between R and the database management system.
  i) *dbConnect()* – Used to build connection with the database.
  ii) *dbSendQuery)* - Used to execute a query on the connected database.
  iii) *fetch()* – Used to fetch records from the previously executed query.
  iv) *dbWriteTable()* - Used to copy dataframes into the database table.
- forecast : This package helps to analyze and display univariate time series forecasts. It also requires the installation of the zoo package.
  i) *Arima()* - Used to fit an ARIMA model to a time series.
  ii) *forecast.Arima()* - Used to make forecast up to specified number of steps.
  iii) *plot.forecast()* - Used to plot the forecasted series.
  iv) *accuracy()* - Used to calculate the accuracy of the fitted model and its forecasts.
- TSA : This package was created by [9]. It includes various functions for time series analysis. It also requires the installation of the leaps, locfit, mgcv and tseries packages.
  i) *eacf()* - Used to compute the sample EACF of the data.
- tseries : This package includes various functions for time series analysis as well as computational finance.

i) *adf.test()* - Used to perform the Augmented Dickey Fuller test on the data.

- quantmod : This package contains tools for downloading financial data, plotting common charts and doing technical analysis. It also requires the TTR and xts packages.

  i) *getSymbols()* – Used to collect the historical prices from Yahoo Finance.

- RMySQL : This package is used to implements DBI Interface to MySQL and MariaDB databases.

  i) *MySQL()* - Used to authenticate and connect to one or more MySQL databases.

A lot more built in functions were used in our project , such as, acf(), pacf(), png(), capture.output(), hist(), Box.test() etc.

4.3.3 Database

In our database, we currently have the historical daily price data of forty S&P500 companies that are listed in NASDAQ and NYSE since January'1980. Some companies' data didn't begin from 1980 however as they were not available. The data was collected as a data frame and then were copied to our database table by R.

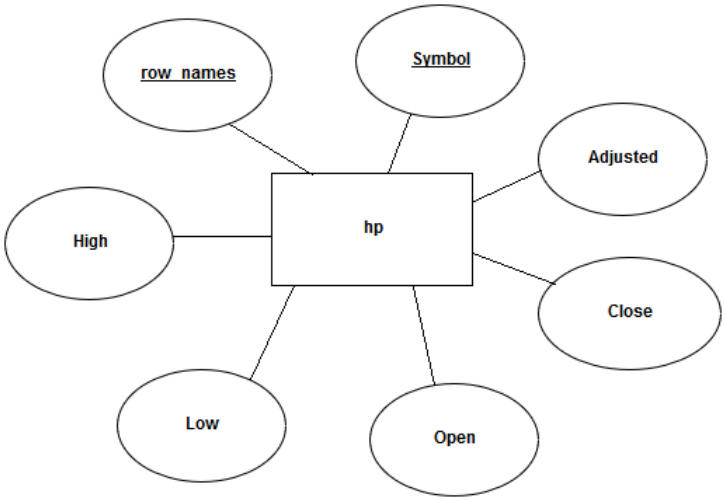

Figure 4.27: ER Diagram for the System

Our database consists of only one table at present, which is the hp table. It stores the historical price data of the different companies. Our table has a composite primary key, consisting of the date and the stock symbol name, i.e. the row_names and Symbol attributes. Although we are working only with the Close prices of the stocks, we chose to keep the other retrieved information as well, so that we could use them in the future.

4.3.4 Integration of PHP and R

When users submit the inputs, PHP executes the Rscript and passes the input variables concatenated as a string to the latter through its exec() command. If the user submits the form of the Analysis page, then the analysis.R Rscript is executed as follows:

```
exec("Rscript --vanilla E:\analysis.R ".\$param)
```

In case they use the Model-Fitting page, the following code is executed:

```
exec("Rscript --vanilla E:\model.R ".\$param)
```

At the beginning of both the Rscripts, there is the following function:

```
commandArgs(trailingOnly=TRUE)
```

This function captures all the arguments supplied by the command line when the R session gets invoked. Setting trailingOnly=TRUE ensures that the only the arguments after --args are captured. Based on these arguments, the Rscript runs the different functions to generate the results. The results are then saved in a directory by R, which are then fetched and shown on the website using PHP and HTML. If the user presses the download button, a download.php file gets called. Then by using the content disposition header, they force the browser to show the dialogue box for saving the 'txt or .png file.

4.4 Challenges

The biggest challenge while working with this project was that there was no initial knowledge regarding Time Series Analysis. There were a lot of things to learn in a very short amount of time. The bulk of the time was spent on understanding the different concepts of time series analysis and forecasting. Still, there are more topics to be covered and incorporated into this project. Moreover, the initial intention was to store the historical prices of all companies listed in NASDAQ, NYSE and AMEX starting from the year 1980. However, importing only the data of NASDAQ caused the database performance to deteriorate. The server would occasionally freeze or would take several minutes when sent a query. Therefore, the decision was taken to use a smaller database for the system, consisting of 40 companies' historical prices.

Chapter 5

Further Exploration

5.1 Theoretical Approach

In our project, we have considered fitting ARIMA models to the stock returns time series which almost all the time demonstrates stationarity. Often, while trying to fit ARIMA models to a time series of stock returns it is seen that the ACF, PACF, EACF plots as well as the Q-Q plots and histograms plots seem to indicate the series is white noise, which would mean that it cannot be used to predict future values. However, if further in-depth analysis of the series is done, we can find more hidden information in it.
Let us consider an example of the stock returns of AAPL for the year 2015. The ACF and PACF plots show no significant correlations among the returns values, as shown in Figure 5.1 and Figure 5.2.

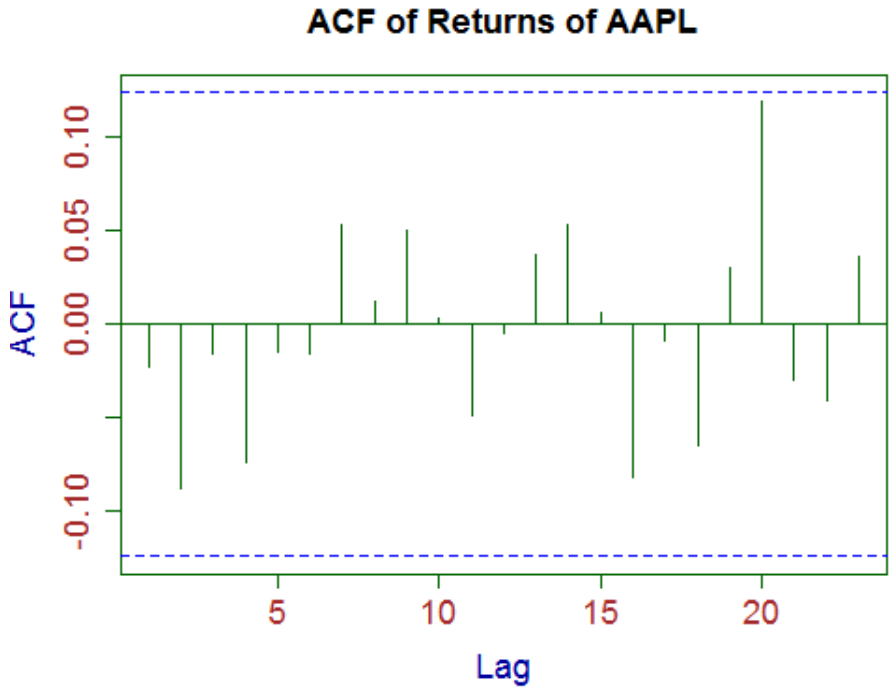

Figure 5.1: ACF Plot of AAPL Stock Returns for 2015

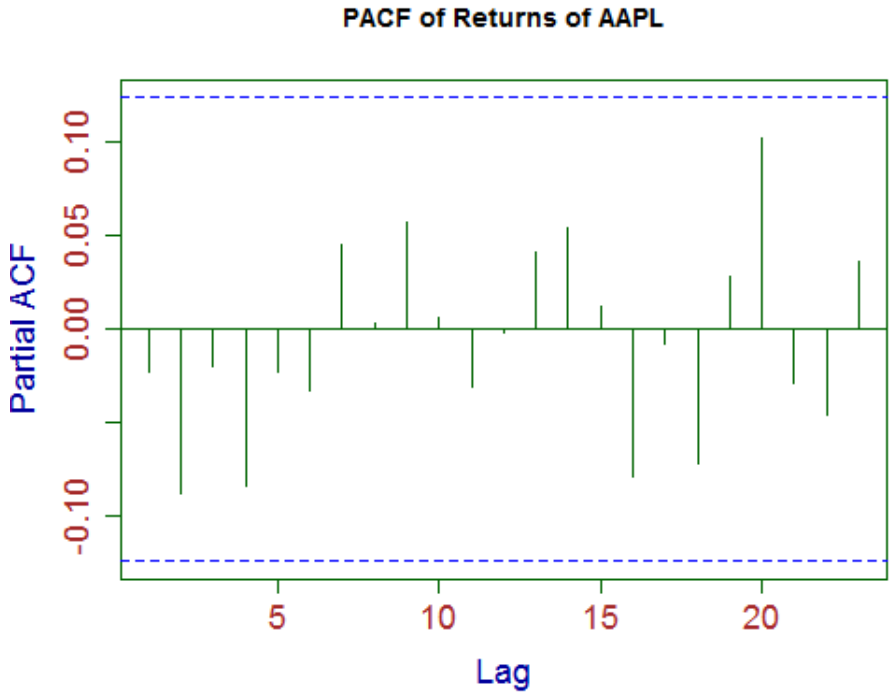

Figure 5.2: PACF Plot of AAPL Stock Returns for 2015

The EACF table (Figure 5.3) hints at an ARMA(0,0) model for the series which is the random walk model.

```
$symbol
  0   1   2   3   4   5   6   7   8   9   10  11  12  13
0 "o" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o"
1 "x" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o"
2 "x" "x" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o"
3 "x" "x" "x" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o"
4 "x" "x" "x" "x" "o" "o" "o" "o" "o" "o" "o" "o" "o" "o"
5 "x" "x" "x" "o" "x" "o" "o" "o" "o" "o" "o" "o" "o" "o"
6 "x" "x" "x" "o" "x" "o" "o" "o" "o" "o" "o" "o" "o" "o"
7 "o" "x" "x" "x" "x" "o" "o" "o" "o" "o" "o" "o" "o" "o"
```

Figure 5.3: EACF of AAPL Stock Returns for 2015

The histogram plot (Figure 5.4) and Q-Q plot(Figure 5.5) indicate slight normality and the ADF test results in a p.value less than 0.1, meaning that the series is stationary.

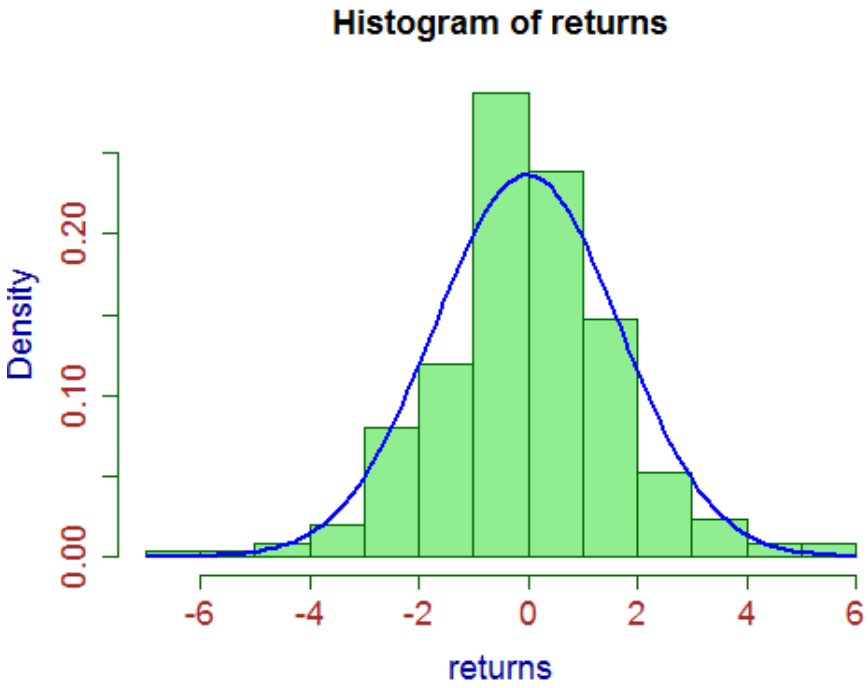

Figure 5.4: Histogram Plot of AAPL Stock Returns for 2015

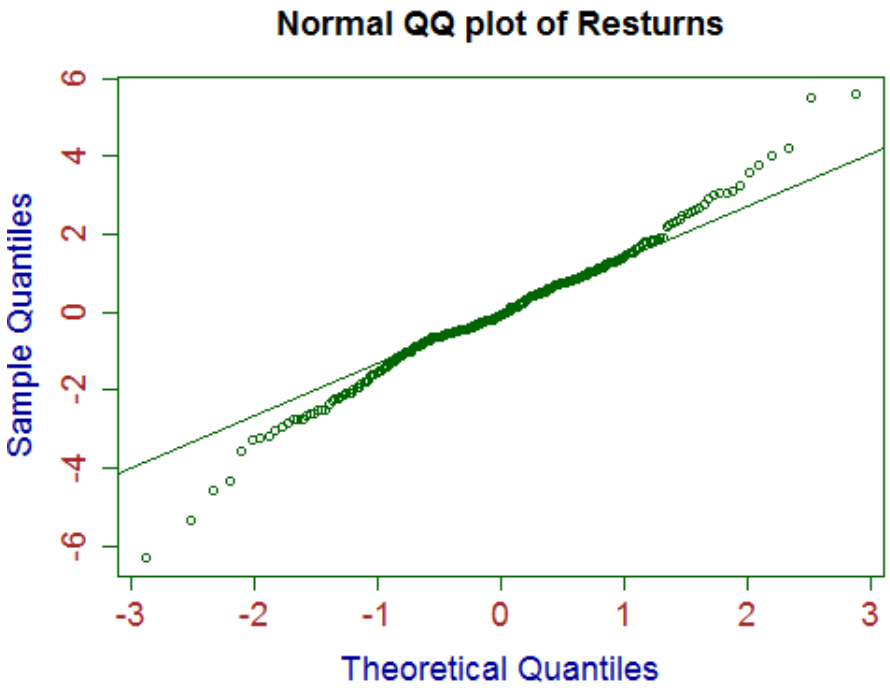

Figure 5.5: Q-Q Plot of AAPL Stock Returns for 2015

All these outputs indicate that the series is white noise, that is, it is independently and identically distributed with constant mean and variance.
If a particular time series is independently and identically distributed, then the absolute

value or square or the logarithm of that time series will also be independently and identically distributed. If we plot the ACF and PACF of the absolute values of the AAPL dataset, then we see significant autocorrelations in the series, as shown in Figure 5.6 and Figure 5.7 respectively.

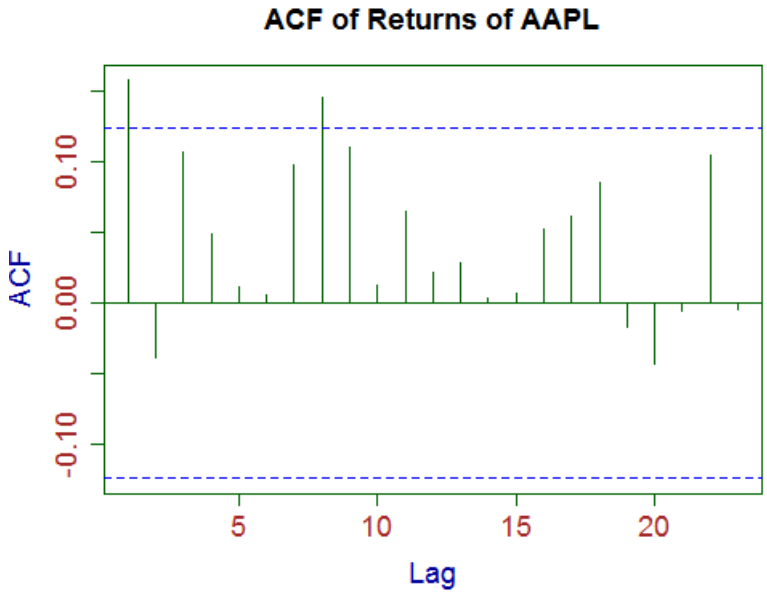

Figure 5.6: ACF Plot of Absolute Values of AAPL Stock Returns for 2015

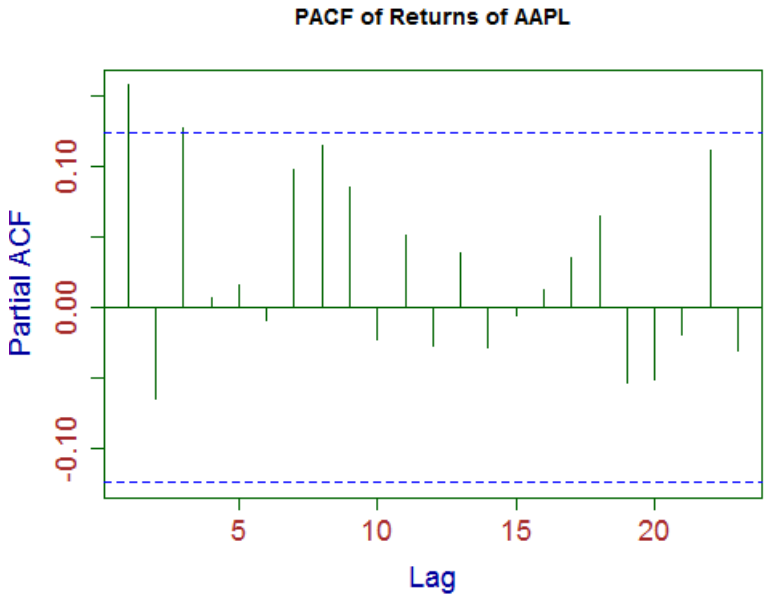

Figure 5.7: PACF Plot of Absolute Values of AAPL Stock Returns for 2015

Thus we can conclude that there are still information present within the returns time series.

5.1.1 ARCH/GARCH Models

There could be higher order dependence such as heavy-tailed distribution or volatility clustering in the returns time series which could not be detected by the simple ARIMA models. The conditional variance of the series might vary over time. This type of characteristic is very common in most financial time series and they are usually addressed with the Generalized Autoregressive Conditional Heteroscedasticity (GARCH) models.

A GARCH(p,q) model is expressed as follows:

$$\sigma^2_{t|t-1} = \omega + \beta_1\sigma^2_{t-1|t-2} + ... + \beta_p\sigma^2_{t-p|t-p-1} + \alpha_1 r^2_{t-1} + \alpha_2 r^2_{t-2} + ... + \alpha_q r^2_{t-q} \tag{5.1}$$

Here,
p = lags of conditional variance
q = order of ARCH
An ARCH(q) model is given by,

$$\sigma^2_{t|t-1} = \omega + \alpha_1 r^2_{t-1} + \alpha_2 r^2_{t-2} + ... + \alpha_q r^2_{t-q} \tag{5.2}$$

By combining the ARIMA models with the ARCH/GARCH models, we could extract more information from the series of AAPL stock returns.

5.1.2 ARIMAX Models

Time series analysis assumes that all observations are taken at equal intervals of time, which is not the case, as the stock market remains closed during the non-trading days and within those days, a lot of information may have been circulated that could cause dramatic changes in the prices on the first trading day. In cases such as these, it would be difficult to fit time series ARIMA models to the data set. In order to make the time series approach of model-fitting and prediction more effective, we have to bring other factors into account. [12] suggested that we have to incorporate secondary variables such as, competition of the company, political events in the company's country, natural disasters, speculations about the company made in the market, etc.

ARIMAX models are simply ARIMA models with additional explanatory variables provided by economic theory. It can be expressed as follows:

$$Y_t = \beta X_t + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} - \theta_1 e_{t-1} - ..... - \theta_q e_{t-q} + e_t \tag{5.3}$$

Here,
X_t = a covariate at time t
β = coefficient of the covariate
By including external variables into our ARIMA model, we could extract more information from the series of AAPL stock returns.

5.2 Real Time Analysis

In this project, we worked with only the daily closing prices of stock market data, whereas stock quotes get updated every minute. The closing prices get updated at the end of the day when the market is closed. Within this time frame, a lot can happen, none of which

gets included in our analysis. If we had minutely data in our hands, we could have been able to fit models more efficiently to the data and thus we could have gotten better predictions. In the future, we intend to collect live data and process them in real time to give outputs to the users.

5.3 System

We could incorporate the following features in the website:

- Build an Android/iOS application
- Forum for discussion
- Video tutorials regarding how to perform the analysis and model fitting
- More interactive during analysis and model-fitting

Chapter 6

Conclusion

This project was undertaken with a view to contributing to the financial technology sector by allowing investors or any interested party to learn to analyze stock market data and also to use that knowledge to build their own models. In order to bring this project to fruition, we started by studying the various methods of stock market prediction. We then chose to work with the time series approach of analysis and forecasting of financial data. We studied the different ARIMA models and built the website, basing on them. While working with ARIMA models, we realized that using only pure ARIMA models cannot lead to an accurate prediction. More factors need to be taken in to account. The next evolution of our project is to incorporate such models with the basic ARIMA models and thus make the system more efficient in terms of analysis and forecasting.

Bibliography

[1] Rob j. hyndman. URL http://robjhyndman.com/tsdldata/roberts/skirts.dat.

[2] Applied time series analysis. URL https://onlinecourses.science.psu.edu/stat510/node/41.

[3] w3schools. URL http://www.w3schools.com/w3css/.

[4] Econometrics academy. URL https://sites.google.com/site/econometricsacademy/econometrics-models/time-series-arima-models.

[5] R-bloggers. URL https://www.r-bloggers.com/.

[6] Tc2000. URL https://widgets.tc2000.com/.

[7] A. O. Adebiyi, A. A.and Adewumi and C. K. Ayo. Stock price prediction using the arima model. *AMSS 16th International Conference on Computer Modeling and Simulation*, pages 105–111, 2014.

[8] C. Chatfield. *The Analysis of Time Series : Theory and Practice*. Springer, 1975.

[9] J. D. Cryer and K. Chan. *Time Series Analysis With Applications in R*. Springer.

[10] T. Ding, V. Fang, and D. Zuo. Stock market prediction based on time series data and market sentiment. URL http://murphy.wot.eecs.northwestern.edu/~pzu918/EECS349/final_dZuo_tDing_vFang.pdf.

[11] E. F. Fama. The behavior of stock-market prices. *Journal of Finance*, 38:34–105, January 1965.

[12] S. Green. Time series analysis of stock prices using the box-jenkins approach. *Electronic Theses & Dissertations*, 2011.

[13] Investopedia. Efficient market hypothesis: Is the stock market efficient? URL http://www.investopedia.com/articles/basics/04/022004.asp.

[14] B. Malkiel. *A Random Walk Down Wall Street*. W. W. Norton & Company, 1973.

[15] P. Mondal, L. Shit, and S. Goswami. Study of effectiveness of time series modeling (arima) in forecasting stock prices. *International Journal of Computer Science, Engineering and Applications (IJCSEA)*, 4:13–29, April 2014.

[16] P. Pai and C. Lin. A hybrid arima and support vector machines model in stock price forecasting. *The International Journal of Management Science*, 33:497–505, 2004.

[17] D. K. Pearce. Stock prices and the economy. *Federal Reserve Bank of Kansas City Economic Review*, pages 7–22, 1983.

[18] V. H. Shah. Machine learning techniques for stock prediction. URL http://www.vatsals.com/.

[19] K. Tseng, O. Kwon, and L. C. Tjung. Time series and neural network forecast of daily stock prices. *Investment Management and Financial Innovations*, 9:32–54, 2012.

[20] R. Weber. Time series, 1999. URL http://www.statslab.cam.ac.uk/~rrw1/timeseries/index.html.

[21] S. Y. Xu. Stock price foreasting using information from yahoo finance and google trend. URL https://www.econ.berkeley.edu/sites/default/files/Selene%20Yue%20Xu.pdf.