Project Dissertation on

Time series forecasting of GDP in India using ARIMA model

Submitted By :

Monika Singh

2K17/MBA/049

Under the Guidance of :

Prof. P.K. Suri

Professor,DSM

DELHI SCHOOL OF MANAGEMENT

Delhi Technological University

Bawana Road Delhi 110042

CERTIFICATE FROM THE INSTITUTE

This is to certify that Monika Singh (2K17/MBA/049) have satisfactorily completed the Project Report entitled “Time series forecasting of GDP in India using ARIMA model” in partial fulfillment of the requirement for the award of degree of Masters of Business Administration from Delhi School of Management, Delhi Technological University, New Delhi during academic year 2018-1019.

Signature of Guide

(Prof. P.K. SURI)

Signature of Head (DSM)

(Dr. Rajan Yadav)

Place:

Seal of Head:

Date:

DECLARATION

I, Monika Singh (2K17/MBA/049), student of Delhi School of Management, Delhi Technological University declare that the work entitled “Time series forecasting of GDP in India using arima model “ is my individual work under the supervision of Professor P.K. Suri.

The findings in this report are not copied from any report and are true to the best of my knowledge.

(Monika Singh)

Place:

Date:

Acknowledgement

With colossal pleasure,I am presenting “Time series forecasting of GDP in India using ARIMA model” report as part of the curriculum of ‘ Master of Business Administration’. I wish to thank all the people who gave me unending support while bringing out this project to its ultimate form.

First and foremost, I would like to express my deep sense of gratitude and gratefulness to my supervisor P.K. Suri, Assistant Professor ,Delhi School of Management ,Delhi Technological University for his invaluable encouragement, suggestions and support from an early stage of this project and providing me extraordinary experiences throughout the work.Above all,his priceless and meticulous supervision at each and every phase of work inspired me in innumerable ways.

Finally, I am thankful to the entire faculty members of Delhi School of Management, Delhi Technological University , New Delhi , my colleagues and my family members for their moral support and constant encouragement while carrying out this project.

-Monika Singh

Executive Summary:

Gross Domestic Product (GDP) of a country is the money value of all final goods and services produced by all the enterprises within the borders of a country in a year. It represents the aggregate statistic of all economic activity. The performance of economy can be measured with the help of GDP. Forecasting future economic outcomes is a vital component of the decision-making process in central banks for all countries. Monetary policy decisions affect the economy with a delay, so, monetary policy authorities must be forward looking, i.e. must know what is likely to happen in the future.Gross domestic product (GDP) is one of the most important indicators of national economic activities for countries. Scientific prediction of the indicator has important theoretical and practical significance on the development of economic development goals.For the forecasting of time series we use models that are based on a methodology that was first developed in Box and Jenkins (1976), known as ARIMA (Auto-Regressive-Integrated-Moving-Average) methodology. This approach was based on the World representation theorem, which states that every stationary time series has an infinite moving average (MA) representation, which actually means that its evolution can be expressed as a function of its past developments (Jovanovic and Petrovska 2010).The rest of the paper is organized as follows: Section 1 describes introduction while in Section 2 and 3, theoretical background and data interpretation and coding in R is given.In Section 4 the empirical results are presented. Section 5 is the forecasting and finally, conclusions.

Table of Contents

1.INTRODUCTION

1.1 About Time series forecasting :

Time series modeling is a dynamic research area which has attracted attentions of researchers community over last few decades. The main aim of time series modeling is to carefully collect and rigorously study the past observations of a time series to develop an appropriate model which describes the inherent structure of the series. This model is then used to generate future values for the series, i.e. to make forecasts. Time series forecasting thus can be termed as the act of predicting the future by understanding the past. Due to the indispensable importance of time series forecasting in numerous practical fields such as business, economics, finance, science and engineering, etc., proper care should be taken to fit an adequate model to the underlying time series. It is obvious that a successful time series forecasting depends on an appropriate model fitting. A lot of efforts have been done by researchers over many years for the development of efficient models to improve the forecasting accuracy. As a result, various important time series forecasting models have been evolved in literature. One of the most popular and frequently used stochastic time series models is the Autoregressive Integrated Moving Average (ARIMA) model. The basic assumption made to implement this model is that the considered time series is linear and follows a particular known statistical distribution, such as the normal distribution. ARIMA model has subclasses of other models, such as the Autoregressive (AR) , Moving Average (MA) and Autoregressive Moving Average (ARMA) models. For seasonal time series forecasting, Box and Jenkins had proposed a quite successful variation of ARIMA model, viz. the Seasonal ARIMA (SARIMA) .The popularity of the ARIMA model is mainly due to its flexibility to represent several varieties of time series with simplicity as well as the associated Box-Jenkins methodology for optimal model building process. But the severe limitation of these models is the pre-assumed linear form of the associated time series which becomes inadequate in many practical situations. To overcome this drawback, various non-linear stochastic models have been proposed in literature , however from implementation point of view these are not so straight-forward and simple as the ARIMA models.

A time series is a sequential set of data points, measured typically over successive times.. The measurements taken during an event in a time series are arranged in a proper chronological order. A time series containing records of a single variable is termed as univariate. But if records of more than one variable are considered, it is termed as multivariate. A time series can be continuous or discrete. In a continuous time series observations are measured at every instance of time, whereas a discrete time series contains observations measured at discrete points of time. For example temperature readings, flow of a river, concentration of a chemical process etc. can be recorded as a continuous time series. On the other hand population of a particular city, production of a company, exchange rates between two different currencies may represent discrete time series. Usually in a discrete time series the consecutive observations are recorded at equally spaced time intervals such as hourly, daily, weekly, monthly or yearly time separations.The variable being observed in a discrete time series is assumed to be measured as a continuous variable using the real number scale. Furthermore a continuous time series can be easily transformed to a discrete one by merging data together over a specified time interval. Components of a Time Series A time series in general is supposed to be affected by four main components, which can be separated from the observed data. These components are: Trend, Cyclical, Seasonal and Irregular components. A brief description of these four components is given here. The general tendency of a time series to increase, decrease or stagnate over a long period of time is termed as Secular Trend or simply Trend. Thus, it can be said that trend is a long term movement in a time series. For example, series relating to population growth, number of houses in a city etc. show upward trend, whereas downward trend can be observed in series relating to mortality rates, epidemics, etc. Seasonal variations in a time series are fluctuations within a year during the season. The important factors causing seasonal variations are: climate and weather conditions, customs, traditional habits, etc. For example sales of ice-cream increase in summer, sales of woolen cloths increase in winter. Seasonal variation is an important factor for businessmen, shopkeeper and producers for making proper future plans. The cyclical variation in a time series describes the medium-term changes in the series, caused by circumstances, which repeat in cycles. The duration of a cycle extends over longer period of time, usually two or more years. Most of the economic and financial time series show some kind of cyclical variation. For example a business cycle consists of four phases, viz. i) Prosperity, ii) Decline, iii) Depression and iv) Recovery.

Irregular or random variations in a time series are caused by unpredictable influences, which are not regular and also do not repeat in a particular pattern. These variations are caused by incidences such as war, strike, earthquake, flood, revolution, etc. There is no defined statistical technique for measuring random fluctuations in a time series. Considering the effects of these four components, two different types of models are generally used for a time series viz. Multiplicative and Additive models. Multiplicative model is based on the assumption that the four components of a time series are not necessarily independent and they can affect one another; whereas in the additive model it is assumed that the four components are independent of each other.Examples of Time Series Data Time series observations are frequently encountered in many domains such as business, economics, industry, engineering and science, etc . Depending on the nature of analysis and practical need, there can be various different kinds of time series. To visualize the basic pattern of the data, usually a time series is represented by a graph, where the observations are plotted against corresponding time. Below we show two time series plots:

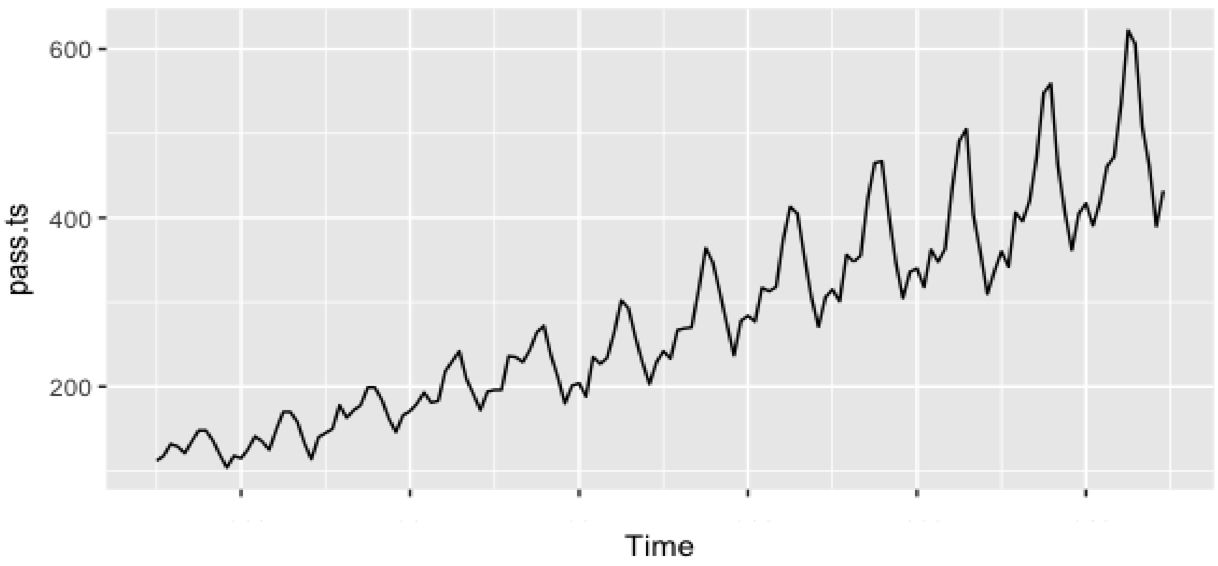

600
400
200
pass.ts
Time

Fig 1.1

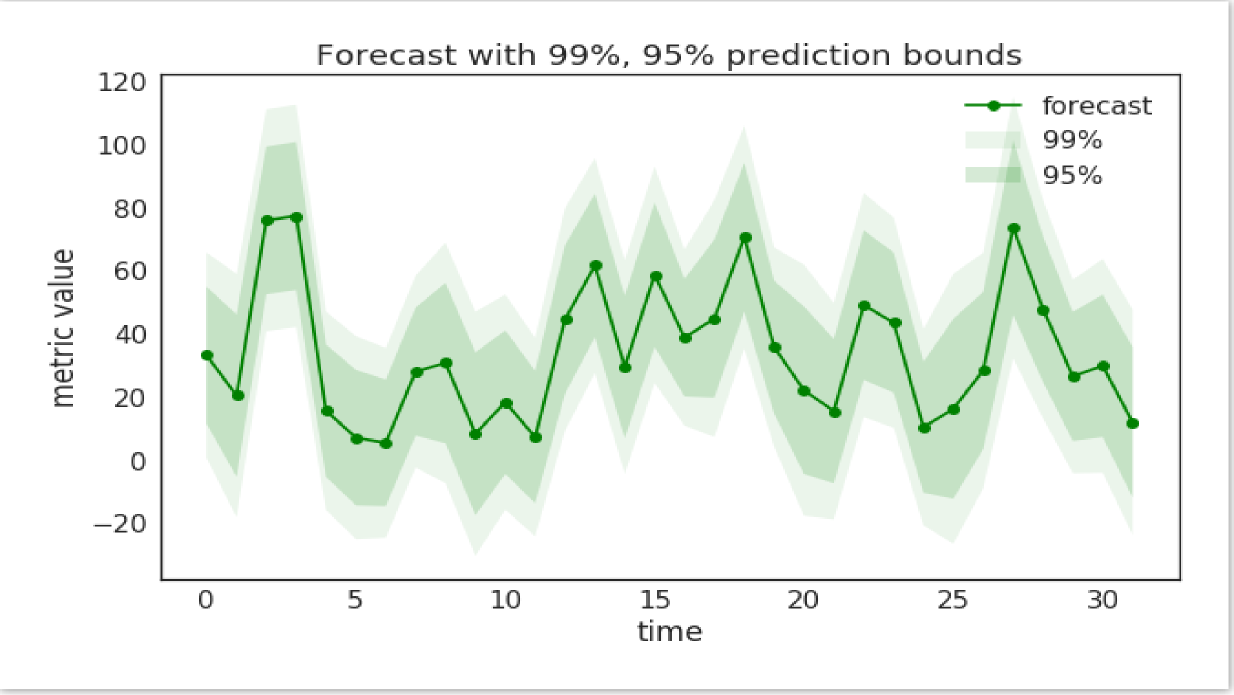
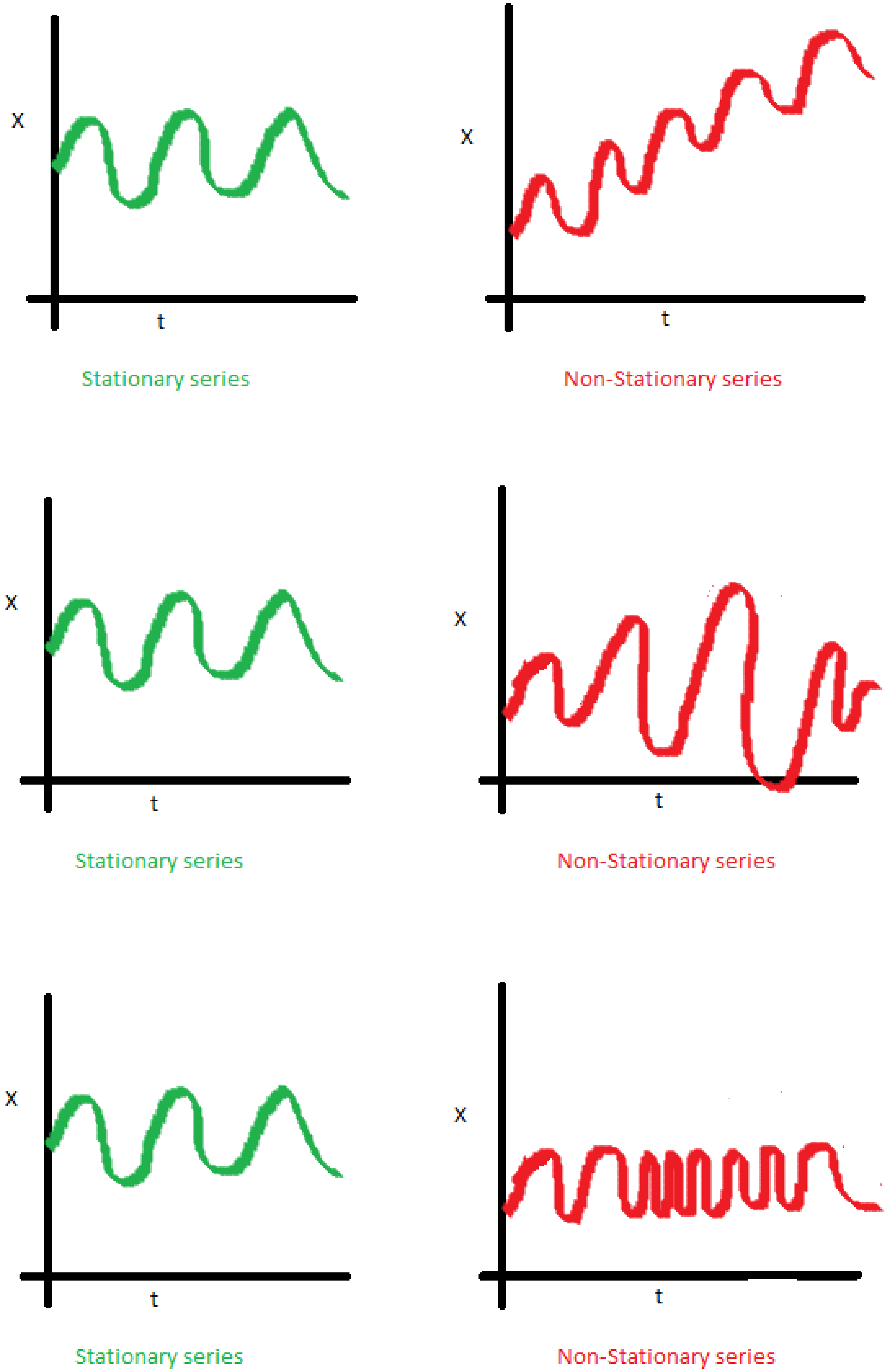

Forecast with 99%, 95% prediction bounds
forecast
99%
95%
120
100
80
60
40
20
0
−20
metric value
0
5
10
15
20
25
30
time

Fig 1.2

Introduction to Time Series Analysis : In practice a suitable model is fitted to a given time series and the corresponding parameters are estimated using the known data values. The procedure of fitting a time series to a proper model is termed as Time Series Analysis. It comprises methods that attempt to understand the nature of the series and is often useful for future forecasting and simulation. In time series forecasting, past observations are collected and analyzed to develop a suitable mathematical model which captures the underlying data generating process for the series .The future events are then predicted using the model. This approach is particularly useful when there is not much knowledge about the statistical pattern followed by the successive observations or when there is a lack of a satisfactory explanatory model. Time series forecasting has important applications in various fields. Often valuable strategic decisions and precautionary measures are taken based on the forecast results. Thus making a good forecast, i.e. fitting an adequate model to a time series is very important. Over the past several decades many efforts have been made by researchers for the development and improvement of suitable time series forecasting models.A time series is non-deterministic in nature, i.e. we cannot predict with certainty what will occur in future. Generally a time series is assumed to follow certain probability model which describes the joint distribution of the random variable.The mathematical expression describing the probability structure of a time series is termed as a stochastic process . Thus the sequence of observations of the series is actually a sample realization of the stochastic process that produced it. A usual assumption is that the time series variables tx are independent and identically distributed following the normal distribution.; They follow more or less some regular pattern in long term. For example if the temperature today of a particular city is extremely high, then it can be reasonably presumed that tomorrow’s temperature will also likely to be high. This is the reason why time series forecasting using a proper technique, yields result close to the actual value.

Concept of Stationarity :

The concept of stationarity of a stochastic process can be visualized as a form of statistical equilibrium .The statistical properties such as mean and variance of a stationary process do not depend upon time. It is a necessary condition for building a time series model that is useful for future forecasting. Further, the mathematical complexity of the fitted model reduces with this assumption. There are two types of stationary processes which are defined below:



Stationary series

Non-Stationary series

Stationary series

Non-Stationary series

Stationary series

Non-Stationary series

Fig 1.3

A process is Strongly Stationary or Strictly Stationary if the joint probability distribution function is independent of t for all .Thus for a strong stationary process the joint distribution of any possible set of random variables from the process is independent of time . However for practical applications, the assumption of strong stationarity is not always needed and so a somewhat weaker form is considered. A stochastic process is said to be Weakly Stationary of order k if the statistical moments of the process up to that order depend only on time differences and not upon the time of occurrences of the data being used to estimate the moments.For example a stochastic process is second order stationary if it has time independent mean and variance and the covariance values.It is important to note that neither strong nor weak stationarity implies the other. However, a weakly stationary process following normal distribution is also strongly stationary.Some mathematical tests like the one given by Dickey and Fuller are generally used to detect stationarity in a time series data. As mentioned in], the concept of stationarity is a mathematical idea constructed to simplify the theoretical and practical development of stochastic processes. To design a proper model, adequate for future forecasting, the underlying time series is expected to be stationary. Unfortunately it is not always the case. Greater the time span of historical observations, the greater is the chance that the time series will exhibit non-stationary characteristics. However for relatively short time span, one can reasonably model the series using a stationary stochastic process. Usually time series, showing trend or seasonal patterns are non-stationary in nature. In such cases, differencing and power transformations are often used to remove the trend and to make the series stationary. In the next chapter we shall discuss about the seasonal differencing technique applied to make a seasonal time series stationary.While building a proper time series model we have to consider the principle of parsimony .According to this principle, always the model with smallest possible number of parameters is to be selected so as to provide an adequate representation of the underlying time series data .Out of a number of suitable models, one should consider the simplest one, still maintaining an accurate description of inherent properties of the time series. The idea of model parsimony is similar to the famous Occam’s razor principle. One aspect of this principle is that when face with a number of competing and adequate explanations, pick the most simple one. The Occam’s razor provides considerable inherent informations, when applied to logical analysis. Moreover, the more complicated the model, the more possibilities will arise for departure from the actual model assumptions. With the increase of model parameters, the risk of overfitting also subsequently increases. An over fitted time series model may describe the training data very well, but it may not be suitable for future forecasting. As potential overfitting affects the ability of a model to forecast well, parsimony is often used as a guiding principle to overcome this issue. Thus in summary it can be said that, while making time series forecasts, genuine attention should be given to select the most parsimonious model among all other possibilities.

The selection of a proper model is extremely important as it reflects the underlying structure of the series and this fitted model in turn is used for future forecasting. A time series model is said to be linear or non-linear depending on whether the current value of the series is a linear or non-linear function of past observations. In general models for time series data can have many forms and represent different stochastic processes. There are two widely used linear time series models in literature, viz. Autoregressive (AR) and Moving Average (MA) models. Combining these two, the Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA) models have been proposed in literature. The Autoregressive Fractionally Integrated Moving Average (ARFIMA) model generalizes ARMA and ARIMA models. For seasonal time series forecasting, a variation of ARIMA, viz. the Seasonal Autoregressive Integrated Moving Average (SARIMA) model is used. ARIMA model and its different variations are based on the famous Box-Jenkins principle and so these are also broadly known as the Box-Jenkins models. Linear models have drawn much attention due to their relative simplicity in understanding and implementation. However many practical time series show non-linear patterns. For example,non-linear models are appropriate for predicting volatility changes in economic and financial time series. Considering these facts, various non-linear models have been suggested in literature. Some of them are the famous Autoregressive Conditional Heteroskedasticity (ARCH) model and its variations like Generalized ARCH (GARCH) , Exponential Generalized ARCH (EGARCH) etc., the Threshold Autoregressive (TAR)] model, the Non-linear Autoregressive (NAR) model, the Non-linear Moving Average (NMA) model, etc.

The Autoregressive Moving Average (ARMA) Models :

An ARMA(p, q) model is a combination of AR(p) and MA(q) models and is suitable for univariate time series modeling. In an AR(p) model the future value of a variable is assumed to be a linear combination of p past observations and a random error together with a constant term. Usually For estimating parameters of an AR process using the given time series, the Yule-Walker equations are used. Just as an AR(p) model regress against past values of the series, an MA(q) model uses past errors as the explanatory variables.The random shocks are assumed to be a white noise process, i.e. a sequence of independent and identically distributed random variables with zero mean and a constant variance .Generally, the random shocks are assumed to follow the typical normal distribution. Thus conceptually a moving average model is a linear regression of the current observation of the time series against the random shocks of one or more prior observations. Fitting an MA model to a time series is more complicated than fitting an AR model because in the former one the random error terms are not fore-seeable. Autoregressive (AR) and moving average (MA) models can be effectively combined together to form a general and useful class of time series models, known as the ARMA models. Here the model orders qp, refer to p autoregressive and q moving average terms. Usually ARMA models are manipulated using the lag operator notation.Polynomials of lag operator or lag polynomials are used to represent ARMA models.

Stationarity Analysis: It is proved by Box and Jenkins that a necessary and sufficient condition for the AR(p) process to be stationary is that all the roots of the characteristic equation must fall outside the unit circle. Hipel and McLeod mentioned another simple algorithm (by Schur and Pagano) for determining stationarity of an AR process.An MA(q) process is always stationary, irrespective of the values the MA parameters . The conditions regarding stationarity and invertibility of AR and MA processes also hold for an ARMA process. An ARMA(p, q) process is stationary if all the roots of the characteristic equation lie outside the unit circle. Similarly, if all the roots of the lag equation lie outside the unit circle, then the ARMA(p, q) process is invertible and can be expressed as a pure AR process.

Autocorrelation and Partial Autocorrelation Functions (ACF and PACF) :

To determine a proper model for a given time series data, it is necessary to carry out the ACF and PACF analysis. These statistical measures reflect how the observations in a time series are related to each other. For modeling and forecasting purpose it is often useful to plot the ACF and PACF against consecutive time lags. These plots help in determining the order of AR and MA terms.
The autocovariance at lag zero i.e. 0γ is the variance of the time series. From the definition it is clear that the autocorrelation coefficient kρ is dimensionless and so is independent of the scale of measurement Another measure, known as the Partial Autucorrelation Function (PACF) is used to measure the correlation between an observation k period ago and the current observation, after controlling for observations at intermediate lags (i.e. at lags k<). At lag 1, PACF(1) is same as ACF(1). Normally, the stochastic process governing a time series is unknown and so it is not possible to determine the actual or theoretical ACF and PACF values. Rather these values are to be estimated from the training data, i.e. the known time series at hand. The estimated ACF and PACF values from the training data are respectively termed as sample ACF and PACF. As explained by Box and Jenkins ,the sample ACF plot is useful in determining the type of model to fit to a time series of length N. Since ACF is symmetrical about lag zero, it is only required to plot the sample ACF for positive lags, from lag one onwards to a maximum lag of about N/4. The sample PACF plot helps in identifying the maximum order of an AR process.

Autoregressive Integrated Moving Average (ARIMA): Models The ARMA models, described above can only be used for stationary time series data. However in practice many time series such as those related to socio-economic business show non-stationary behavior. Time series, which contain trend and seasonal patterns, are also non-stationary in nature.Thus from application view point , ARMA models are inadequate to properly describe non-stationary time series, which are frequently encountered in practice. For this reason the ARIMA model is proposed, which is a generalization of an ARMA model to include the case of non-stationarity as well. In ARIMA models a non-stationary time series is made stationary by applying finite differencing of the data points.An ARIMA(p,0,0) is nothing but the AR(p) model and ARIMA(0,0,q) is the MA(q) model. ARIMA(0,1,0), is a special one and known as the Random Walk model .It is widely used for non-stationary data, like economic and stock price series. A useful generalization of ARIMA models is the Autoregressive Fractionally Integrated Moving Average (ARFIMA) model, which allows non-integer values of the differencing parameter d. ARFIMA has useful application in modeling time series with long memory. In this model the expansion of the term is to be done by using the general binomial theorem. Various contributions have been made by researchers towards the estimation of the general ARFIMA parameters. Box and Jenkinshave generalized this model to deal with seasonality. Their proposed model is known as the Seasonal ARIMA (SARIMA) model. In this model seasonal differencing of appropriate order is used to remove non-stationarity from the series. A first order seasonal difference is the difference between an observation and the corresponding observation from the previous year.

1.2 Research Objectives:

- To test the stationarity in the data of GDP over the period.
- To study Autocorrelation in the observed series of GDP
- To Forecast the GDP using appropriate ARIMA Model.
- To test the Model fitness using Information Criterion and goodness of fit model.

2.Literature Review:

India’s growth along with sustainability in the next decade majorly depends on the growth in its market and economy as a whole. In the present study researcher has attempted to forecast the GDP growth for the country. Out of a variety of forecasting models ARIMA (1,2,2) model has been applied to forecast the GDP over a period of ten years ranging from 2020 to 2030.The results indicate the fitness of AR (1) I (2) MA(2) parameters for making the future predictions. It is concluded that the GDP of India would be rising continuously over the estimated period. Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) were studied to study the AR and MA terms. Augmented Dickey Fuller (ADF) Unit root test was used to test the stationarity of the data and identification of Integration order. Akaike Information Criterion (AIC), Root Mean Squared Error, Mean Absolute Percentage Error were applied to study the model fitness.

Keywords: Auto Regressive Integrated Moving Average (ARIMA), Autocorrelation Function (ACF), Akaike Information Criterion (AIC), Forecast, GDP, Mean Absolute Percentage Error, Partial Autocorrelation Function (PACF), Root Mean Squared Error

3.Research Methodology: -

3.1 R Programming:

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

One of R’s strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

R is available as Free Software under the terms of the Free Software Foundation’s GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

The R environment

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display either on-screen or on hardcopy, and
- a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

The term “environment” is intended to characterise it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.

R, like S, is designed around a true computer language, and it allows users to add additional functionality by defining new functions. Much of the system is itself written in the R dialect of S, which makes it easy for users to follow the algorithmic choices made. For computationallyintensive tasks, C, C++ and Fortran code can be linked and called at run time. Advanced users can write C code to manipulate R objects directly.

Many users think of R as a statistics system. We prefer to think of it of an environment within which statistical techniques are implemented. R can be extended (easily) via *packages*. There are about eight packages supplied with the R distribution and many more are available through the CRAN family of Internet sites covering a very wide range of modern statistics.

R has its own LaTeX-like documentation format, which is used to supply comprehensive documentation, both on-line in a number of formats and in hardcopy.

R package:

The primary uses of R is and will always be, statistic, visualization, and machine learning. The picture below shows which R package got the most questions in Stack Overflow. In the top 10, most of them are related to the workflow of a data scientist: data preparation and communicate the results. All the libraries of R, almost 12k, are stored in CRAN. CRAN is a free and open source. You can download and use the numerous libraries to perform Machine Learning or time series analysis. Why use R?

Data science is shaping the way companies run their businesses. Without a doubt, staying away from Artificial Intelligence and Machine will lead the company to fail. The big question is which tool/language should you use?

They are plenty of tools available in the market to perform data analysis. Learning a new language requires some time investment. The picture below depicts the learning curve compared to the business capability a language offers. The negative relationship implies that there is no free lunch. If you want to give the best insight from the data, then you need to spend some time learning the appropriate tool, which is R.

In a nutshell, R is a great tool to explore and investigate the data. Elaborate analysis like clustering, correlation, and data reduction are done with R. This is the most crucial part, without a good feature engineering and model, the deployment of the machine learning will not give meaningful results.

4.DATA ANALYSIS AND INTERPRETATION

The data of GDP (Nominal Value) was collected over the time period from 2000 to 2018. It had been collected from the website publications of Reserve Bank of India (RBI). The following figure exhibits the GDP Curve over the observed period of 18 years.

Conceptual Framework of ARIMA Model:

Popularly known as the Box-Jenkins (BJ) methodology, and technically known as the ARIMA methodology, the emphasis of these methods is not on constructing single equation or simultaneous equation models but on analyzing the probabilistic or stochastic properties of economic time series on their own under the philosophy of letting the data speak for themselves. Unlike the regression models, in which *Yt* is explained by *k* regressors, *X*1, *X*2, *X*3...*Xk*, the BJ type time series models allow *Yt* to be explained by past or lagged valued of *Yt* itself and stochastic error terms. For this reason, ARIMA models are sometimes called a-theoretic models because these are not derived from any economic theory, while economic theories are often the basis of simultaneous equation models. Let Yt be a time series sequence for t = 1, 2 ….t ,then we can say that *Yt* follows a first order Autoregressive (AR)(1). Here the value of *Y* at time *t* depends on its value in the previous time period and a random term. In other words, this model says that the forecast value of *Y* at time *t* is simply some proportion of its value at time (*t* – 1) plus a random shock or disturbance at time *t,* again the values are expressed around their mean values. Economic variables with time series data are usually non-stationary, since these are integrated. These need first order differencing for attaining stationarity. If a time series is integrated of order 1, its first differences are *I* (0), and it is stationary. Similarly, if a time series is *I* (2), its second difference is *I* (0). In general, if a time series is *I* (*d*), after differencing it *d* times, we obtain an I (0) series. Therefore , if we have to difference a time series *d* times to make it stationary and then apply an ARIMA time series, where *p* denotes the number of AR terms, *d* represents the number of times, the series has to be differenced before it becomes stationary, and *q* is the number of MA terms.

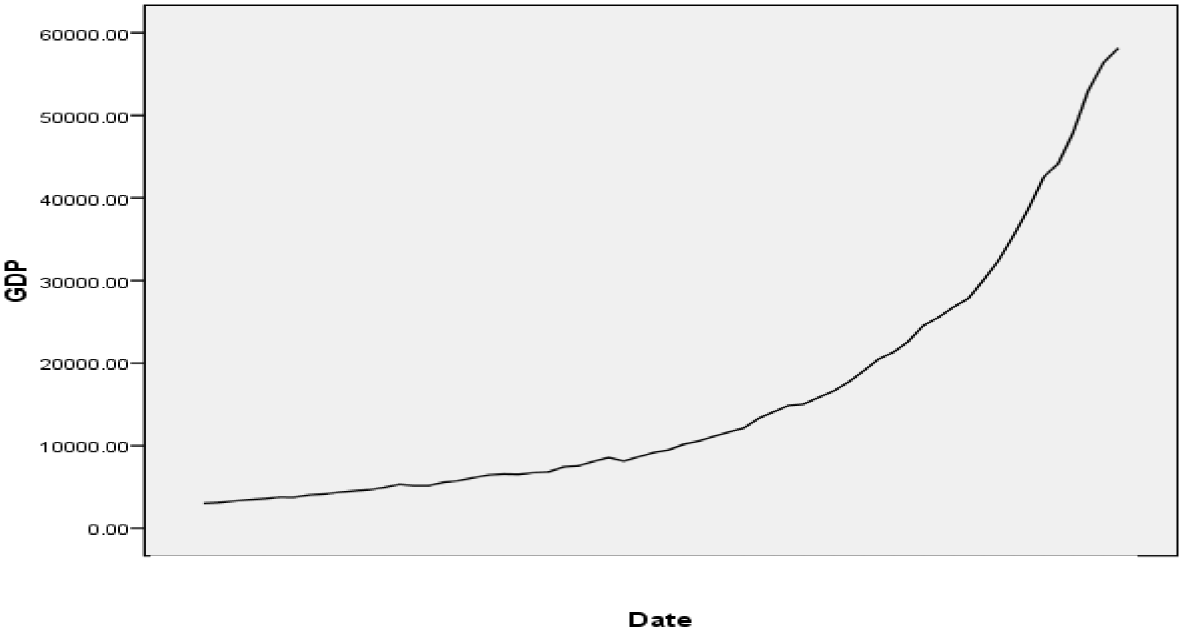


Fig 4.1

Dickey Fuller Test of Stationarity

Here is a small tweak which is made for our equation to convert it to a Dickey Fuller test:

X(t) = Rho * X(t-1) + Er(t)
=> X(t) - X(t-1) = (Rho - 1) X(t - 1) + Er(t)

We have to test if Rho – 1 is significantly different than zero or not. If the null hypothesis gets rejected, we'll get a stationary time series.

Stationary testing and converting a series into a stationary series are the most critical processes in a time series modelling. You need to memorize each and every detail of this concept to move on to the next step of time series modelling.

4.2 Coding in R:

```
> data(GDP)
 > class(GDP)
 [1] "ts"
#This tells you that the data series is in a time series format
 > start(GDP)
 [1] 2000 1
#This is the start of the time series
> end(GDP)
```

```
 [1] 2018
#This is the end of the time series
> plot(GDP)
#This will plot the time series
>abline(reg=lm(GDP~time(GDP)))
# This will fit in a line
```

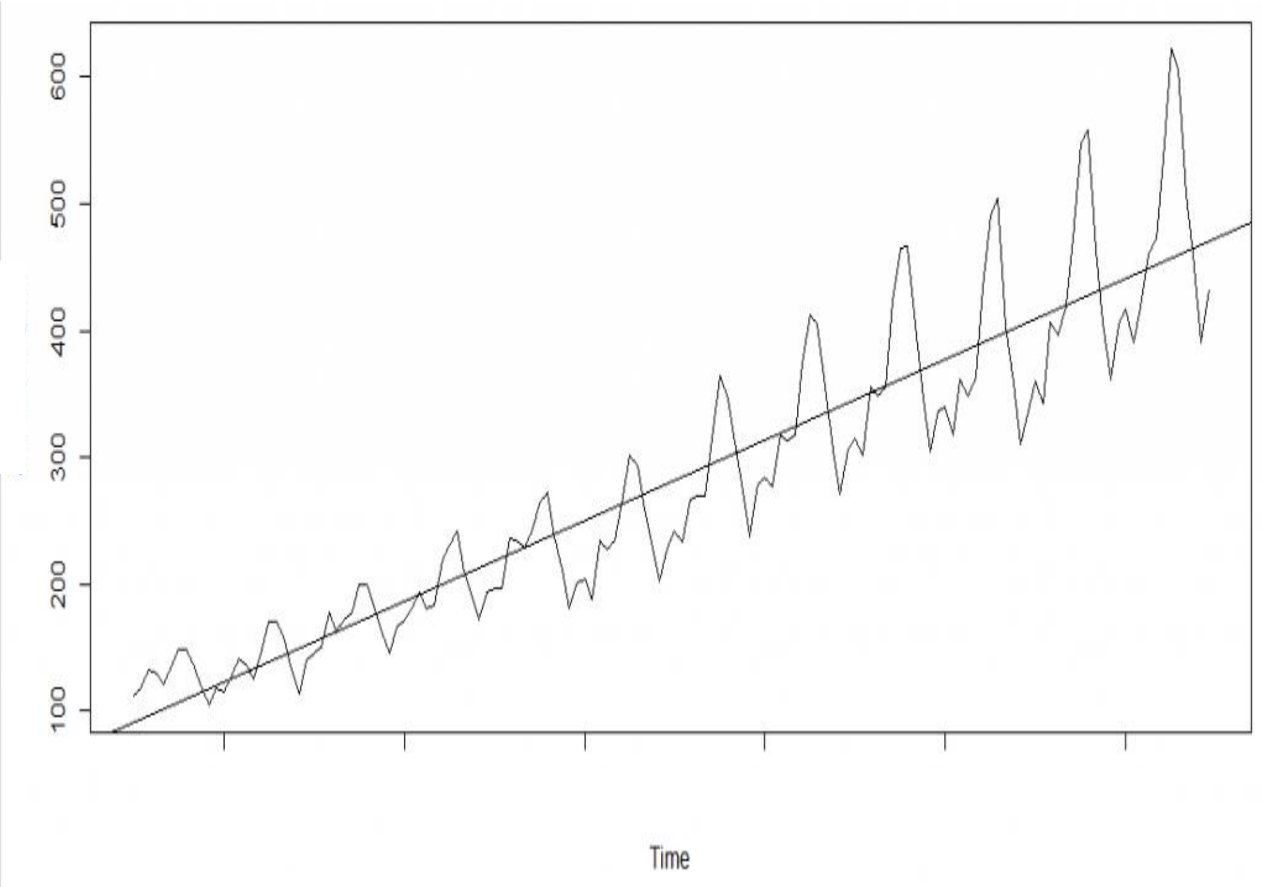


Fig4.2

Here are a few more operations you can do:

```
> cycle(GDP)
#This will print the cycle across years.
>plot(aggregate(GDP,FUN=mean))
#This will aggregate the cycles and display a year on year trend
> boxplot(GDP~cycle(GDP))
#Box plot across months will give us a sense on seasonal effect
```

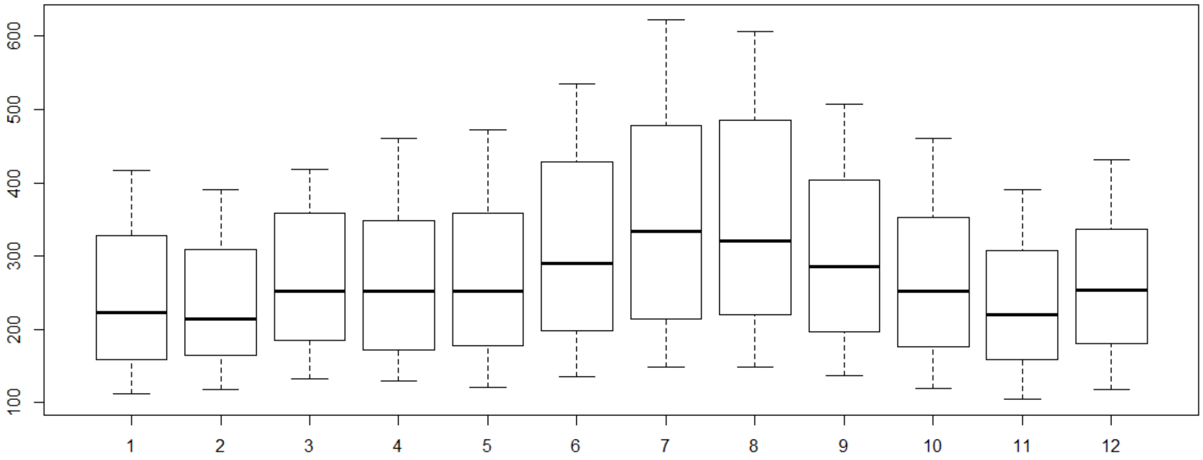


Fig 4.3

```
adf.test(diff(log(GP)), alternative="stationary", k=0)
data: diff(log(GDP))
 Dickey-Fuller = -9.6003, Lag order = 0,
 p-value = 0.05
 alternative hypothesis: stationary
acf(log(GDP))
```

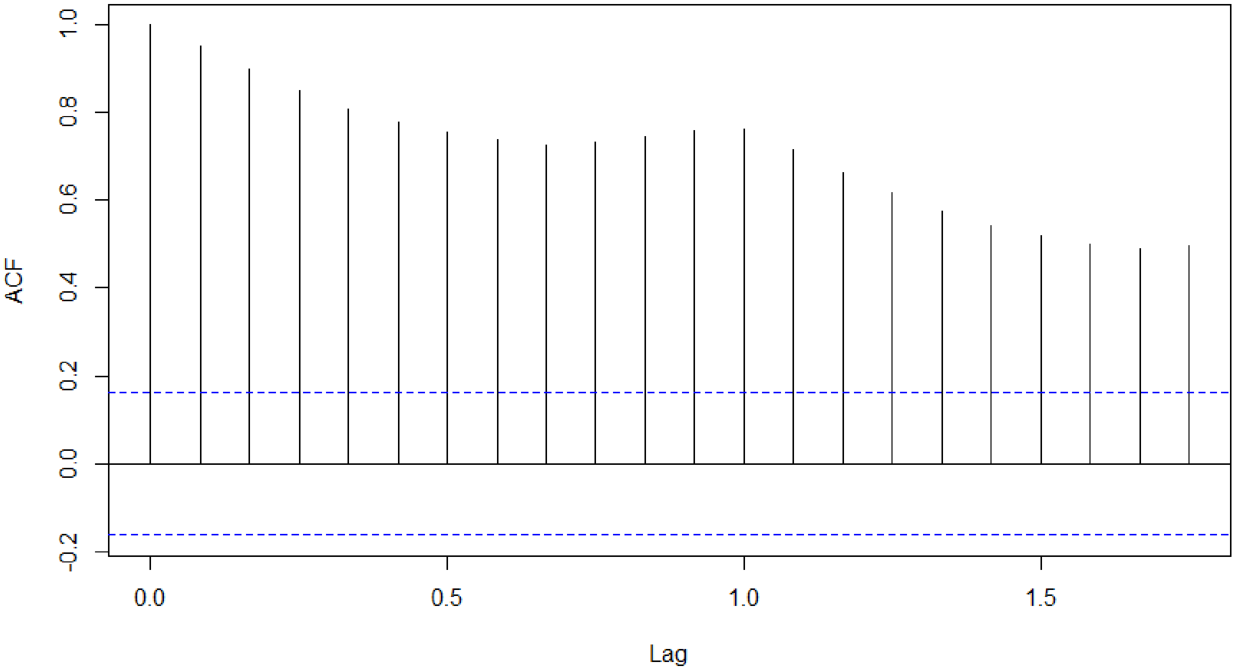


Fig 4.4

```
acf(diff(log(GDP)))
```

z

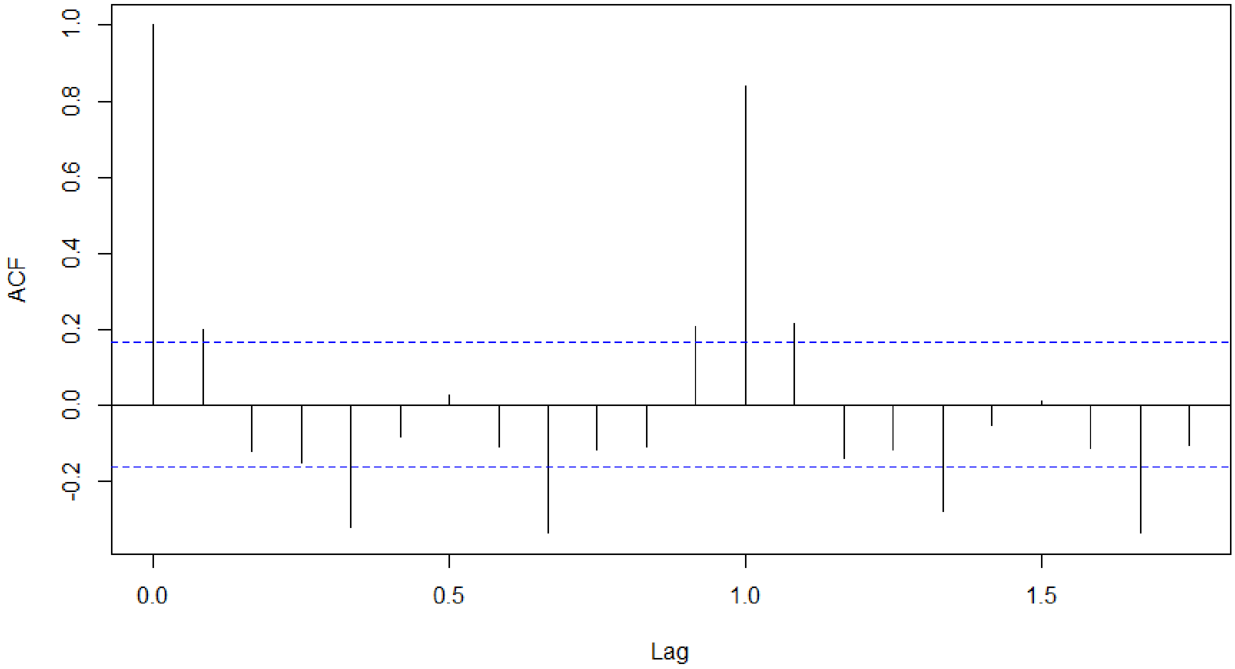


Fig 4.5

Clearly, ACF plot cuts off after the first lag. Hence, we understood that value of p should be 0 as the ACF is the curve getting a cut off. While value of q should be 1 or 2. After a few iterations, we found that (0,1,1) as (p,d,q) comes out to be the combination with least AIC and BIC.

Let's fit an ARIMA model and predict the future 10 years. Also, we will try fitting in a seasonal component in the ARIMA formulation. Then, we will visualize the prediction along with the training data. You can use the following code to do the same

```
(fit <- arima(log(GDP), c(0, 1, 1),seasonal = list(order = c(0, 1, 1), period = 12)))
pred <- predict(fit, n.ahead = 10)
```

4.3 Important Inferences

1. The year on year trend clearly shows that the GDP have been increasing without fail.
2. The variance and the mean value in July quarter 2018 is much higher than rest of the quarters.
3. Even though the mean value of each year is quite different their variance is small.
4. ARMA models are commonly used in time series modeling. In ARMA model, AR stands for auto-regression and MA stands for moving average.
5. In case you get a non-stationary series, you first need to stationarize the series (by taking difference / transformation) and then choose from the available time series models.

Auto-Regressive Time Series Model:

The current GDP of a country say x(t) is dependent on the last year's GDP i.e. x(t – 1). The hypothesis being that the total cost of production of products & services in a country in a fiscal year (known as GDP) is dependent on the set up of manufacturing plants / services in the previous year and the newly set up industries / plants / services in the current year. But the primary component of the GDP is the former one.

Hence, we can formally write the equation of GDP as:

$$x(t) = \text{alpha} * x(t-1) + \text{error}(t)$$

This equation is known as *AR(1) formulation*. The numeral one (1) denotes that the next instance is solely dependent on the previous instance. The alpha is a coefficient which we seek so as to minimize the error function. Notice that x(t- 1) is indeed linked to x(t-2) in the same fashion. Hence, any shock to x(t) will gradually fade off in future.

For instance, let's say x(t) is the number of juice bottles sold in a city on a particular day. During winters, very few vendors purchased juice bottles. Suddenly, on a particular day, the temperature rose and the demand of juice bottles soared to 1000. However, after a few days, the climate became cold again. But, knowing that the people got used to drinking juice during the hot days, there were 50% of the people still drinking juice during the cold days. In following days, the proportion went down to 25% (50% of 50%) and then gradually to a small number after significant number of days. The following graph explains the inertia property of AR series:

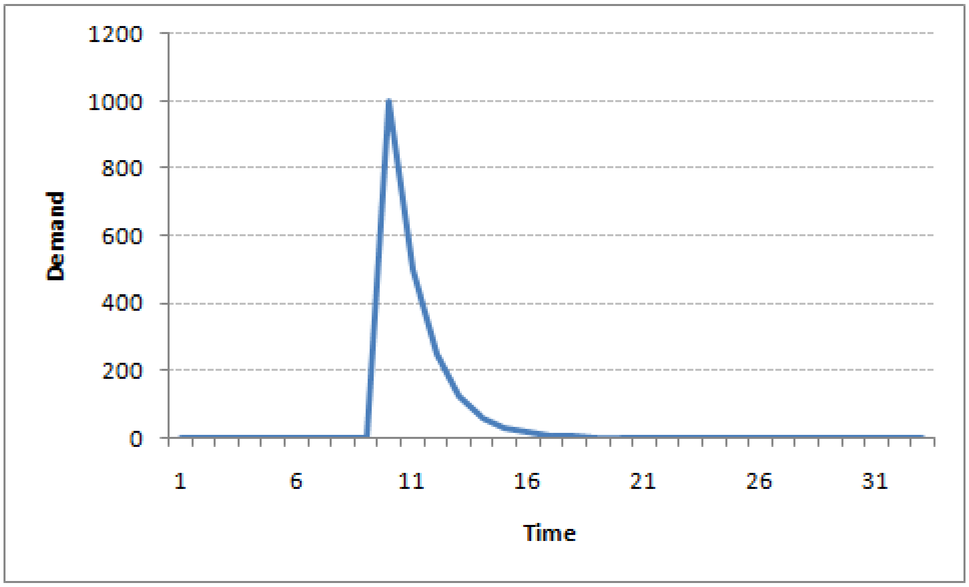


Fig 4.5

Moving Average Time Series Model:

Example-A manufacturer produces a certain type of bag, which was readily available in the market. Being a competitive market, the sale of the bag stood at zero for many days. So, one day he did some experiment with the design and produced a different type of bag. This type of bag was not available anywhere in the market. Thus, he was able to sell the entire stock of 1000 bags (lets call this as x(t)). The demand got so high that the bag ran out of stock. As a result, some 100 odd customers couldn't purchase this bag. Lets call this gap as the error at that time point. With time, the bag had lost its woo factor. But still few customers were left who went empty handed the previous day. Following is a simple formulation to depict the scenario :

x(t) = beta * error(t-1) + error (t)

If we try plotting this graph, it will look something like this :

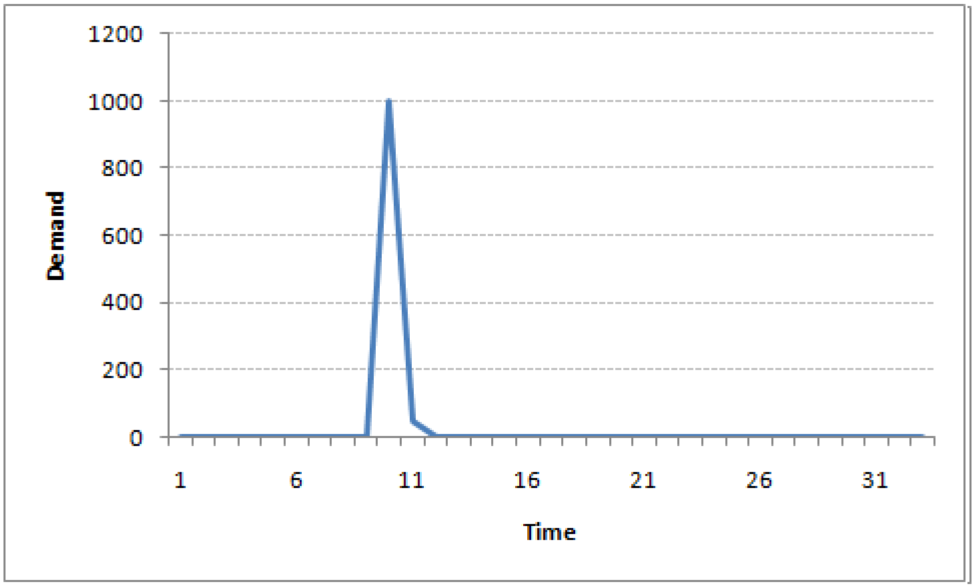


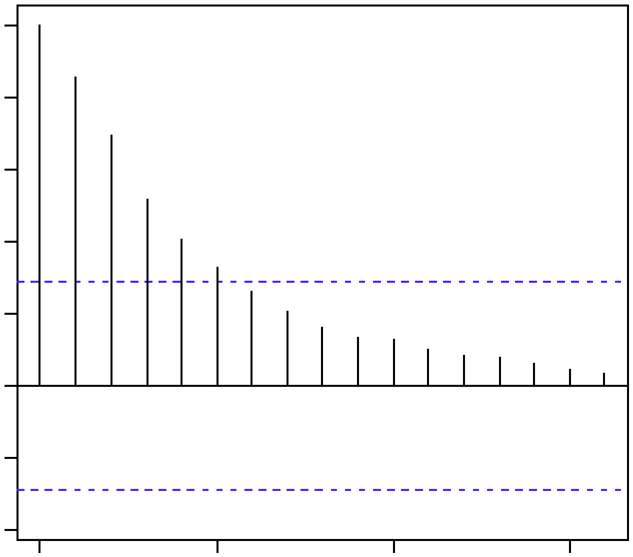
Fig 4.6

? In MA model, noise / shock quickly vanishes with time. The AR model has a much lasting effect of the shock.

Difference between AR and MA models:

The primary difference between an AR and MA model is based on the correlation between time series objects at different time points. The correlation between x(t) and x(t-n) for n > order of MA is always zero. This directly flows from the fact that covariance between x(t) and x(t-n) is zero for MA. However, the correlation of x(t) and x(t-n) gradually declines with n becoming larger in the AR model. This difference gets exploited irrespective of having the AR model or MA model. The correlation plot can give us the order of MA model.

ACF



PACF

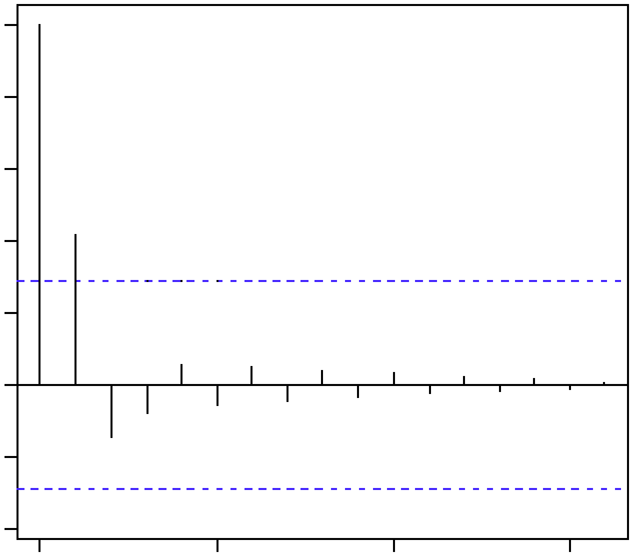


Fig 4.7

The blue line above shows significantly different values than zero. Clearly, the graph above has a cut off on PACF curve after 2nd lag which means this is mostly an AR(2) process.

This framework(shown below) specifies the step by step approach on ‘How to do a Time Series Analysis‘:

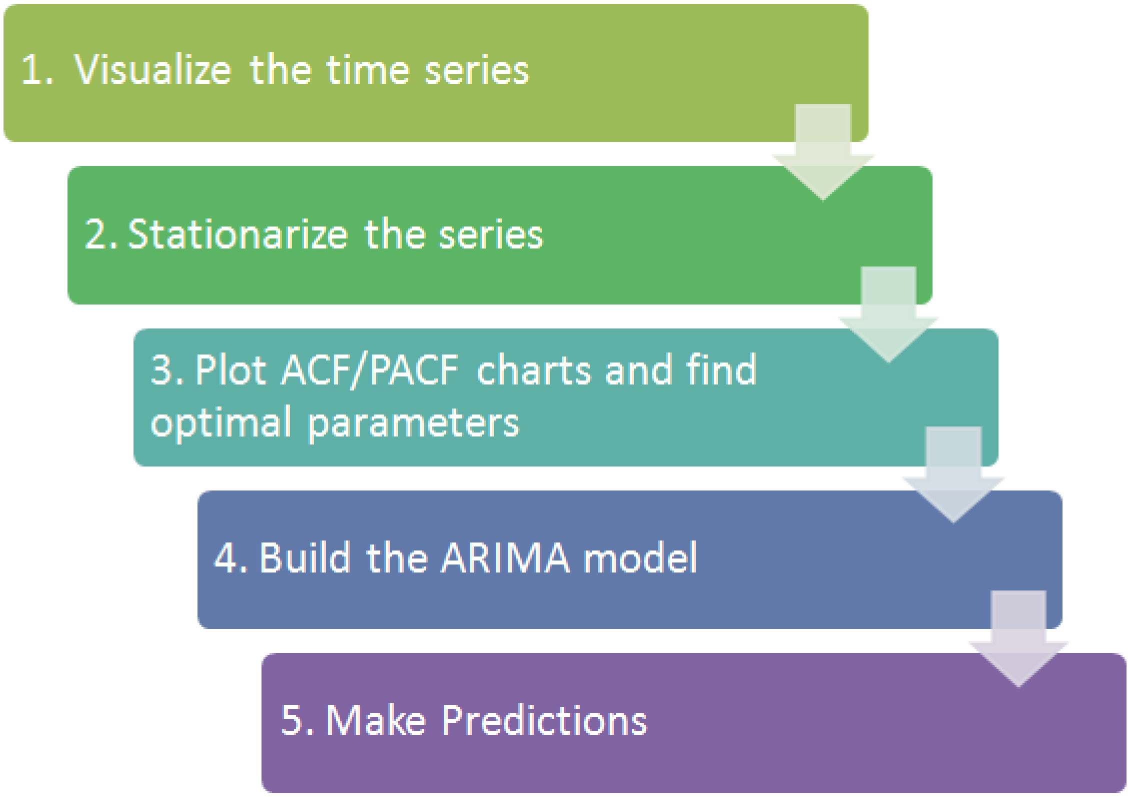

1. Visualize the time series
2. Stationarize the series
3. Plot ACF/PACF charts and find optimal parameters
4. Build the ARIMA model
5. Make Predictions

Fig 4.8

4.4 Results:

Identification of I (d) term:

The GDP series was differences twice as ADF test justified the presence of unit root in the data. Thus, the stationarity was achieved by differencing the GDP series twice. It thus, validated the integration order at two lags I(2).

Table-1: ADF Unit Root Test

t-Stat istic Pro b.*

Augmented Dickey-Fuller te st statistic 4.41 7378 1.0 000

Test critical values: 1% level -3.568308

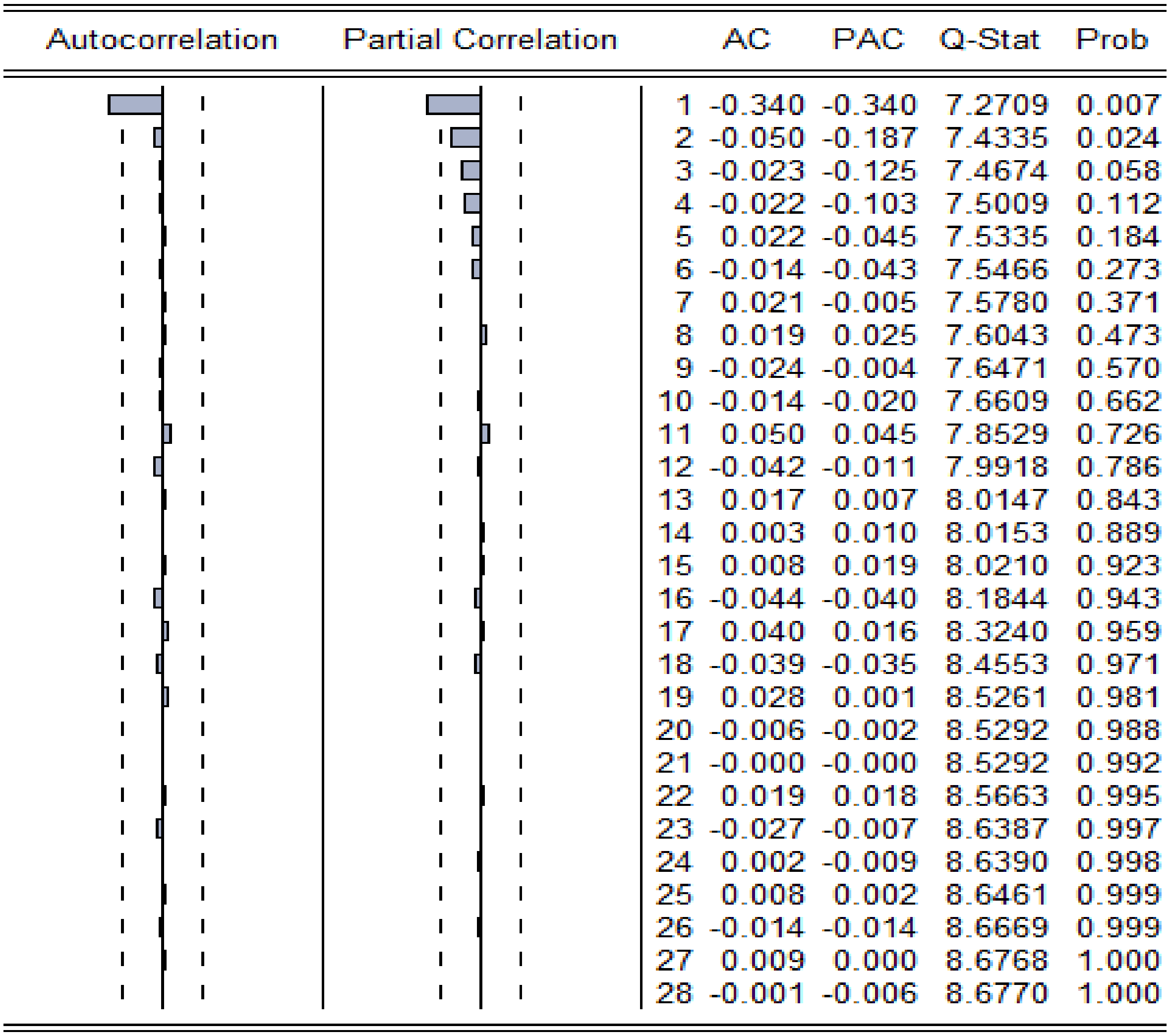
5% level -2.921175

10% level -2.598551

*MacKinnon - one-sided p-values .

Identification of AR (p) and MA (q) terms:

After making the GDP series stationary , the autocorrelation and partial autocorrelation functions were studied. By observing the PACF values and term AR was found to be fit for predictions. Similarly, MA term were justified by observing the ACF values. The MA (1) was rejected as it was found to be insignificant. Thus, the ARMA (1, 2) parameters were identified using Autocorrelation and Partial Autocorrelation Functions.



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.340	-0.340	7.2709	0.007
		2	-0.050	-0.187	7.4335	0.024
		3	-0.023	-0.125	7.4674	0.058
		4	-0.022	-0.103	7.5009	0.112
		5	0.022	-0.045	7.5335	0.184
		6	-0.014	-0.043	7.5466	0.273
		7	0.021	-0.005	7.5780	0.371
		8	0.019	0.025	7.6043	0.473
		9	-0.024	-0.004	7.6471	0.570
		10	-0.014	-0.020	7.6609	0.662
		11	0.050	0.045	7.8529	0.726
		12	-0.042	-0.011	7.9918	0.786
		13	0.017	0.007	8.0147	0.843
		14	0.003	0.010	8.0153	0.889
		15	0.008	0.019	8.0210	0.923
		16	-0.044	-0.040	8.1844	0.943
		17	0.040	0.016	8.3240	0.959
		18	-0.039	-0.035	8.4553	0.971
		19	0.028	0.001	8.5261	0.981
		20	-0.006	-0.002	8.5292	0.988
		21	-0.000	-0.000	8.5292	0.992
		22	0.019	0.018	8.5663	0.995
		23	-0.027	-0.007	8.6387	0.997
		24	0.002	-0.009	8.6390	0.998
		25	0.008	0.002	8.6461	0.999
		26	-0.014	-0.014	8.6669	0.999
		27	0.009	0.000	8.6768	1.000
		28	-0.001	-0.006	8.6770	1.000

The ARIMA (1, 2, 2) parameters were found significant at 5% level ofsignificance. The coefficient of AR (1) was estimated as 0.54 and that of MA (2) as 0.49. Thet-test confirms the significance of these coefficients for predicting the GDP. A model ARIMA (1, 1, 1) was rejected due to insignificance of its AR (1) term while testing. The model fitness was confirmed by lower AIC values and lower values of root mean squared error. The R-square is merely 24%.

Model Validity:

The correlogram of ACF of residuals (Fig 3) suggests that there is no substantial spike has been observed as the case of correlogram of PACF of residuals. In turn, it has been concluded that the error terms become white noise. It thus validates the model as no further information is available.

Figure: Residual ACF and PACF

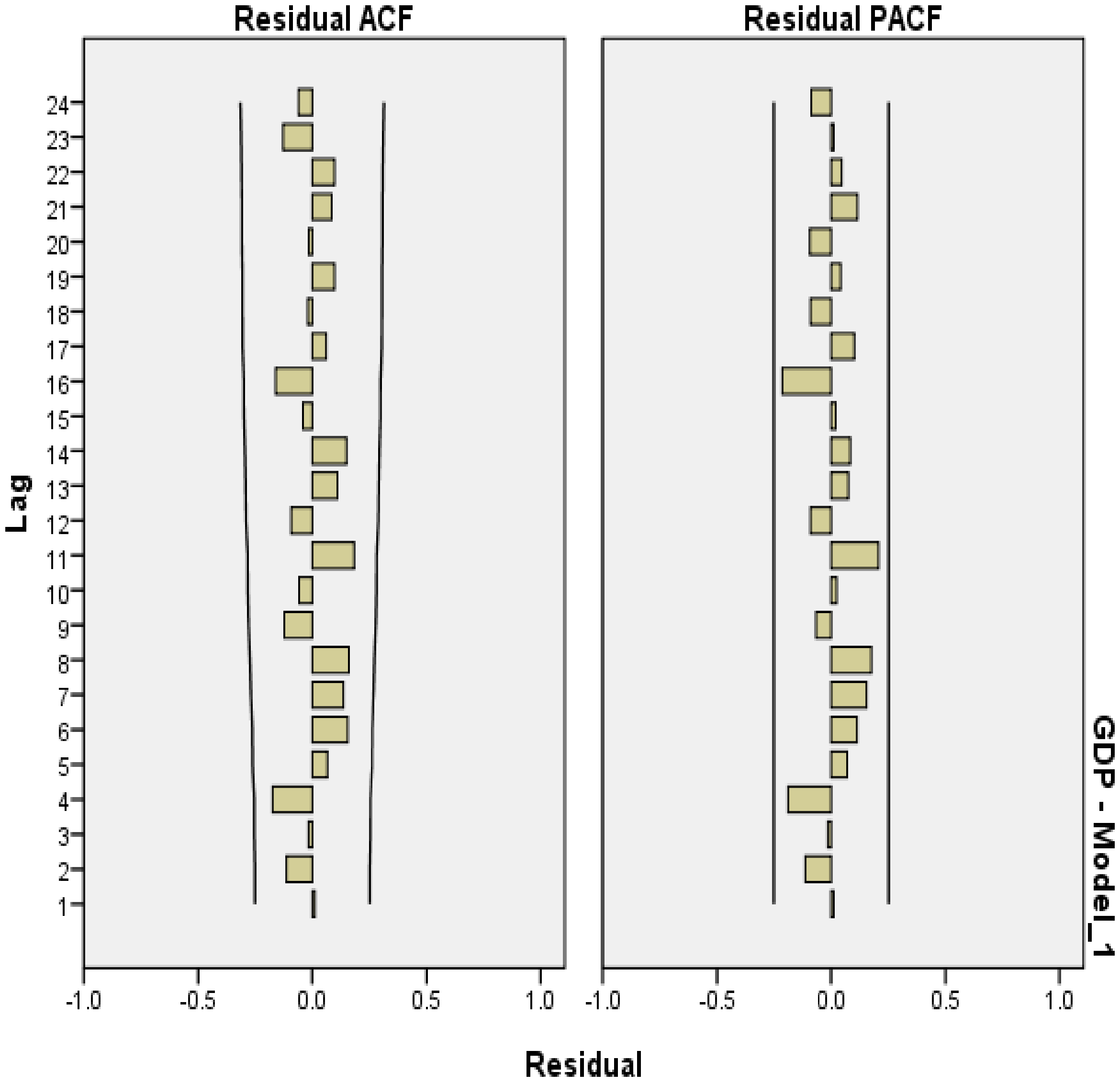


Fig 4.9

4.5 Forecasting:

```
 Output: Time Series:
 Start=21
 End=30
 Frequency = 1
 [1] 134170.5   281196.3   468367.7   746227.4   1111836.9   1541294.7   2061617   2685230.4
340003.6  [10] 4213741.4
```

4.6 Limitations of the study :

- Overdifferencing in ARIMA model– A series that has been differenced one too many times will show very strong negative autocorrelation and a strong MA signature, probably with a unit root in MA coefficients.

- Sometimes AR, ARCH. GARCH, ARIMA, etc. do not seem to be helpful in forecasting in the case of coming of crisis.

- In time-series forecasting, you rarely know the true shape of the distribution with which you are working. Workers in other areas often assume normal distributions while knowing that assumption is not fully accurate.

.

4.7 Conclusion:

The GDP forecast for the next decade is increasing over the period of time.Also the RBI should take necessary steps as and when needed to regulate different rates to keep the indian economy stable.The ARIMA (1, 2, 2) model was found to be a better fit model in forecasting India’s GDP.

4.8 Implications of the Study:

The results of this study would be very useful for policy makers and managers dealing with macro variables such as foreign direct investment (FDI), foreign institutional investment (FII), etc. The findings of the study will be helpful for formulation of better policies. Managers who are planning to invest in the expansion of existing business or in the new project will be benefitted greatly as the findings will provide them a picture of the economic conditions of India well in advance. Further, the findings suffer from some limitations since the researchers have not taken into consideration the models such as Regression Analysis, VAR, ECM etc. to forecast GDP and its growth rates in India.

5.References:

- Ansley, C.F. (1979), An algorithm for the exact Likelihood of a mixed autoregressivemoving average process, Biometrika,.
- Ard, H.J, Den, R. (2010), Macroeconomic Forecasting using Business Cycle leading indicators, Stockholm
- Bipasha, M. and Chatterjee, B. (2012), Forecasting GDP Growth Rates of India: An Empirical Study. International Journal of Economics and Management Sciences
- Box, G.E.P. and Jenkins, G.M. (1970), Time Series Analysis: Forecasting and Control, San Francisco: Holden-Day.
- Box, G.P. and Jenkins, Time Series Forecasting and Control
- Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994), Time Series Analysis: Forecasting and Control