FORECASTING KENYA GROSS DOMESTIC PRODUCT (GDP) USING ARIMA MODEL

A research project submitted in partial fulfillment of the requirements for the award of the Bachelors degree in Applied Statistics of Masinde Muliro University of Science and Technology

DECLARATION

This research project is our original work and has never been submitted for the award of degree or any other certificate in any other university or institution of higher learning.

BRIAN OTIENO OLOO
REG NO: SST/B/01-01827/2017
Signature.............................
Date..................................

ROONEY ODHIAMBO ODAGO
REG NO: SST/B/01-04083/2017
Signature.............................
Date.....................................

MANYARA ERICK OIRERE
REG NO:SST/B/01-01828/2017
Signature................................
Date.....................................

RAHMA ADAN IBRAHIM
REG NO:SST/B/01-01796/2017
Signature................................
Date.....................................

ABIGAEL MWANISA LUGAYE
REG NO:SST/B/01-01800/2017
Signature................................
Date.....................................

APPROVAL

This research project has been presented for examination with my approval as the supervisor duly appointed by the university.

DR . DRINOLD MBETE
Signature................................
Date.....................................

ACKNOWLEDGMENT

We thank God for giving us the grace to complete this research project. We equally thank our lecturer, Mr. Drinold Mbete, for leading and guiding us through the drafting of this project to the completion. To all our lecturers for teaching and equipping us with relevant knowledge and skills.

ABSTRACT

Gross Domestic Product (GDP) is the value of all goods and services that are produced within the countrys national borders in a year. Our study aims to estimate and predict Kenyan GDP using time series data from the year 2020 to the year 2029. In this paper, the Box-Jenkins approach has been used to build the appropriate Autoregressive-Integrated Moving-Average (ARIMA) model for the Kenyan GDP data. Kenyas annual GDP data was obtained from the World Bank database, one of the most credible source of worlds microeconomic data for the years 1960 to 2019. We find that the appropriate statistical model for Kenya GDP is ARIMA (1, 2, 1). We selected the appropriate statistical ARIMA model for the Kenyan GDP with the help of R Software employing Box-Jenkins approach using the principle of parsimony and use the fitted ARIMA model to forecast the GDP of Kenya for the next ten years.

LIST OF ABBREVIATION AND ACRONYMS

1. GDP -Gross Domestic Product.

2. $AR_{(p)}$ - Autoregressive process of order p.

3. $MA_{(q)}$-Moving Average process of order q.

4. mad-In statistics, the median absolute deviation (MAD) is a robust measure of the variability of a univariate sample of quantitative data. It can also refer to the population parameter that is estimated by the MAD calculated from a sample.

5. Trimmed- Mean expressed in percentages. The percentage tells you what percentage of data to remove. For example, with a 5% trimmed mean, the lowest 5% and highest 5% of the data are excluded.

6. sd- standard deviation

7. mean-Average of a set of values

8. AIC- AKaikes Information Criterion

9. min-minimum value in the dataset

10. max- maximum value in the dataset

11. range- Difference between maximum value and minimum value

12. kurtosis-Is a statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution

13. skew -Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data.

14. se - Standard error.

Contents

List of Figures

List of Tables

CHAPTER ONE

INTRODUCTION

1.1 Background Information

The Gross Domestic Product (GDP) is the total monetary or market value of all the finished goods and services produced within a country's borders in a specific time period usually in quarterly or yearly basis. It includes the personal consumptions, government purchases, private inventories, paid-in construction costs and foreign trade balance (exports minus imports).

In Kenya, just like in any other country, the need for a more consistent and accurate GDP forecast for the conduct of forward-looking monetary policy is unstoppable. This could be attributed to the fact that the availability of real-time data is very important especially in determining the initial conditions of economic activity on latent variables such as the output gap to make more realistic policy recommendations.

The aim of this research was to model and predict future Kenya's GDP, using a time series data for the period from 2020 till 2029. This study offered a valuable understanding for the Kenya's expected GDP. Prediction of GDP involved application of statistical and mathematical models to predict upcoming developments in the economy. It allowed to review previous economic movements and predict how current economic changes amend the patterns of future trend.Therefore, a more accurate prediction provide a significant help to the government in setting up economic development goals, strategies and policies. Consequently, an accurate GDP prediction presents a leading insight and an understanding for future economic trend. The GDP is highly affected by economic factors, but not limited to Inflation level and Exports and Imports.

1.1.1 Inflation

Inflation refers to the rise in the prices of goods and services of daily use such as food, clothing,recreation,transport and consumers staples.It is also the decline of purchasing power of a given currency over time. Some of the factors that causes inflation are:Demand pull inflation,rising wages,imports prices,raw material prices,profit push inflation and declining productivity.

The conventional view on inflation holds that inflation should not be too high, but should be moderate and stable in order to enhance economic growth. Lucas (1973) posits that inflation should be low in order to propel economic growth by making prices and wages more flexible. Sidrauski (1967) posits that inflation has no effect on growth because money is neutral. In his paper, money is introduced in the utility function. Tobin (1965) believes that money and capital are perfect substitutes; hence, inflation will have a long run positive effect on growth. On the other hand, the cash in advance model of Stockman (1981) argues that money and capital are complementary. Their paper examined the effect of anticipated inflation on the steady state capital stock in an economy, where money is introduced through a cash in advance constraint rather than through the utility functions of individuals.
They assert that there is a negative long-term relationship between growth and inflation.
However, the real effect of money will be different if money serves as a transitionary through a shopping time technology. Inflation represents a tax on real balances; the real effects of altering that tax depend on what we assume about the role and nature of money (Dornbusch and Frenkel 1973: 141). Feldstein (1982) believes that the relationship between inflation and the tax system could affect the lending decisions of consumers and, ultimately, affect the cost of capital and dampen investment, leading to a decline in economic growth. Fischer (1993), Barro (1995; 1996), and De Gre gorio (1993) found evidence for a negative link between inflation and growth. The most recent inflation growth theory postulated is the non-linear effect of inflation

on growth, which is explained through money demand elasticity (Gillman and Ke jak 2005). In the endogenous model, the relationship between inflation and growth is introduced through the marginal product of capital (physical or human).

Most of the empirical studies have conrmed the negative and nonlinear impact of inflation, especially beyond a certain threshold level (Sarel 1996; Ghosh and Philips 1998; Bruno and Easterly 1998; Khan and Senhadji 2001; Gillman and Kejak 2005). The marginal effect of inflation on growth is stronger when the level is at lower rates (Ghosh and Philips 1998). The ingratiation relationship can also be affected by other macroeconomic variables (e.g., trade openness, and degree of financial development, and public expenditure). For example, Eggoh and Khan (2014) observed that macroeconomic factors like trade openness with an excess demand gap can lower the cyclical movement of inflation and output growth in a competitive economy.

1.1.2 Exports and Imports

Exports are defined as movable good produced within the boundaries of one country,which are traded with another country.The sales of these goods generates foreign currency earnings in the country that produces them thus boost its economic growth. Whereas,Imports are defined as goods produced outside the boundaries of one country,which are then purchased by that country. Demands for imports depend on economic conditions in the buying country,as well as exchange rate and relative prices.

Several econometric studies have relied on co-integration theory and the error correction model, developed also by Engle and Granger (1987), to test for causality. Dutt and Ghosh (1996) used a bi-variate vector error correction model (VECM) to examine the causal relations between exports and GDP growth rates. Their sample included several countries for the period 1953-1991. For some countries the authors found no statistical evidence of causality between GDP growth and exports; for some other countries it was found that exports Granger caused GDP growth (export-led growth). In a few countries reverse causality was detected, that is, GDP growth Granger caused exports. In a few other countries the data indicated bi-directional causality.

Ghartley (1993) examined causal relations in Taiwan, US, and Japan, within the framework of a trivariate VECM. The three endogenous variables in this model were GDP growth, exports growth, and the capital stock or the terms of trade. For the US, Ghartley found that economic growth caused growth in exports; for Taiwan growth of exports caused GDP growth and for Japan the data revealed bi-directional causality between GDP and exports growth rates. Tao and Zestos (1999) employed a trivariate VECM in the GDP, exports and imports for Japan and Korea, and found stronger causal links for Korea, which is a more trade dependent country than Japan.

In a similar study based on a trivariate VECM involving U.S. and Canada data, Zestos and Tao (2002) examined Granger causality among exports, imports and GDP growth. Their results support stronger causal relations between GDP and the foreign sector for Canada than for the U.S.

1.2 Statement of the Problem

Kenya has made significant political, structural and economic reforms that have largely driven sustained economic growth, social development and political gains over the past decade. However, its key development challenges still include poverty, inequality, climate change, continued weak private sector investment and the vulnerability of the economy to internal and external shocks. The recent economic expansion has been boosted by a stable macroeconomic environment, positive investor confidence and a resilient services sector. Currently the Kenyan budget is running on a budget deficit meaning that the total expenditure on development, social protection, education and healthcare is much greater than the government revenue this is because government revenue collect persistently hit below the target. The government have increased tax amount to the major economy builders i.e. manufacturing sectors ICT and Financial sectors which only contribute to 14.3% on the country production, this has led to contraction of the companies. The country total debt stands at ksh.7.12 Trillion increase in debt has led to arise of a Debt-to-GDP Ratio. This ratio compares how much a country can produce and how much it owe, this can be used to determine how many years a country will take to completely settle down its debts if it uses its entire GDP for repayment.

1.2.1 Main Objective of the Study.

The study aimed at modelling and forecasting of Gross Domestic Product (GDP) of Kenya using time series.

1.2.2 Specic objectives.

a) To identify the best ARIMA model.
b) To estimate the parameters of the ARIMA model.
c) To fit the GDP data to the best ARIMA model.

1.3 Significance of the study

1. Studying the Gross Domestic Product (GDP) will enable the ministry of finance to determine the health of the economy and whether the economy is growing..

2. The information gotten from forecasting the GDP helps determine whether our country Kenya will be able to pay off her loans, that is, when the GDP will surpass the debts.

3. Its also handy in determining the impacts of Exports and Imports to the economy of Kenya which are the key elements that greatly influences the level of GDP

4. The study will also assist economic policy makers to fine-tune policies to achieve the desired economic level in Kenya.

1.4 Justification of the Study

The aim of this research is to estimate and predict future Kenyas GDP, using a time series data for the period from 2020 till 2029. This study is supposed to bring forth a valuable understanding for the Kenyas expected GDP. Prediction of GDP involves application of statistical and mathematical models to predict upcoming developments in the economy. It allows to review previous economic movements and predict how current economic changes will amend the patterns of previous trend, therefore, a more accurate prediction would provide a significance help to the government in setting up economic development goals, strategies and policies. Consequently, an accurate GDP prediction presents a leading insight and an understanding of future economy trend.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

Economic size is measured by its output, the most widely-used measure of economic output is the GDP, it is generally measured using one of a three approaches, these approaches of measuring GDP should result in the same number, with some possible discrepancies caused by usual mathematical and statistical figures rounding.

Predicting GDP is a vital issue if it is able to understand and capture the future developments of the economy, so it is meaningful to review the economic trends and predict the effect of current economic circumstances on the future GDP trend. This can be done through using time series data of GDP, which consists of observations consecutively generated over time. Such data are ordered with respect to the time, which shows the trend related to the time period observed. The trend may be increasing, decreasing, constant or having a cyclical fluctuation an ups and downs pattern the over time. Also, the data may show that the underlying process has periodic fluctuation of constant length, which is seasonal behavior. therefore, Modeling would capture this underlying process using the observed time series data so that it could be possible to forecast what would likely be realized at a specific point in time in the future. In predicting macroeconomic time series variables like GDP, there are many possible types of models that are used in literature, in our study we will use ARIMA (Litterman,1986, Stockton and Glassman,1987).
In predicting a time series data of GDP as a macroeconomic variable, ARIMA model has been proven to be reliable and an accurate model.

2.2 ARIMA Models

(Tsay and Tiao,1985) used ARIMA modeling, they fitted a non-seasonal data by identifying auto regressive and moving average terms with the help of partial auto correlation and auto-correlation functions (Box and Jenkins,1970). Also (Topolewski, et al. 1995) deployed automatic methods were developed to identify as well as estimate the parameters of ARIMA model by utilizing time-series data for a single variable. Furthermore (Mait and Chatterjee,2012) used similar methodology to model macroeconomic variable like GDP. however, both the studies were limited to only non-seasonal time series and such modeling needs a long time-series data on the macroeconomic variable in question.

Wabomba, Musundi (2016) studied modeling and forecasting Kenyan Gross Domestic Product using ARIMA. GDP data was obtained from the Kenya National Bureau of Statistics for the years 1960 to 2012.
R statistical software was used to build ARIMA model using theBox-Jenkins method to model the GDP.ARIMA(2,2,2)time series model was established as the best for modeling the Kenyan GDP according to the recognition rules and stationary test of time series under the AIC criterion. Finally, they used the model to forecast the GDP of Kenya for the next five years.

Wali et. al (2013) conducted a study on forecasting the cultivated area and production of cotton in India using Auto-regressive Integrated Moving Average (ARIMA) model. Time series data covering the period of 1950-2010 was used for the study. The study revealed that ARIMA (0,1,0), ARIMA (1,1,4) and ARIMA (0,1,1) are the best fitted model for forecasting of cotton area, production and yield in India respectively.
Burnham and Anderson (2004) shows that the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) time series analysis are based on the analysis of historical data. These analyses support the assumption that past patterns in data can be used to forecast future

data points.

Rahman (2010) applied ARIMA model to forecast Boro rice production in Bangladesh from 2008-09 to 2012-13. It appeared from the study that the local, modern and total Boro time series are first order homogeneous stationary. It was found from the study that the ARIMA (0,1,0), ARIMA (0,1,3) and ARIMA (0,1,2) are the best for local, modern and total Boro rice production respectively .

Awal, M. A and Siddique M.A. B (2011) modeled rice production in Bangladesh using ARIMA model. The study was carried out to examine the best ARIMA model to efficiently forecast Aus, Aman and Boro rice production in Bangladesh. ARIMA (4,1,1), ARIMA (2,1,1) and ARIMA (2,2,3) to Aus, Aman and Boro rice production respectively .

Banhi G. et al (2016), studied the application of ARIMA time series model to forecast the future Gold price in Indian browser based on past data from November 2003 to January 2014 to mitigate the risk in purchase of Gold. ARIMA (1,1,1) was chosen from six different model parameters as it provided the best model which satisfied all the criterion of fit statistics.

Mondal et al (2014), conducted a study on the effectiveness of ARIMA model on fifty six Indian stocks from different sectors. The model was chosen because of its simplicity and wide acceptability of the model. The comparison and parametrization of the ARIMA model have been done using Akaike information criterion (AIC).

Ashwini D. et. al (2017) studied forecasting of common Paddy prices in India. The time series data on monthly average prices of Paddy from January 2006 to December 2016 was collected. ARIMA model was employed to predict the future prices of Paddy. Model parameters were estimated using the R programming software. The ARIMA model was the most representative model for the price forecast of Paddy in overall India [5]. Chinn, M.D., et. al (2005) studied the predictive content of energy futures. They examined the relationship between spot and future prices for energy commodities. An ARIMA (1,1,1) was used for crude oil prices forecas.

(Mei, et. al,2011) constructed a multi-factor dynamic system VAR prediction model of GDP by selecting six main economic indicators, using time series data extracted from the Shanghai region in China, also (Wang and Wang ,2011) deployed ARIMA for predicting the GDP of China based on time series data. They set up an ARIMA model of the GDP of China from 1978 to 2006. They then choose the best ARIMA model based on statistical tests and predict the GDP from 2007 to 2011. The result shows that the error between the actual value and the predicted value is insignificant which shows that the ARIMA model is highly accurate. (Wei, Bian and Yuan ,2010) used ARIMA model for GDP data series from 1952 to 2007 to predict the GDP of the Shanxi province in China, they found that error between the real GDP value and the predicted value is within 5 .

A large variety of linear and nonlinear models are now available for modeling and forecasting macroeconomic time series data. A large dataset gathered from China's Shaanxi GDP from 1952 to 2007 forecasted the countrys GDP for the next 6 years. Scrutinized the forecasting of GDP growth for India. They applied an ARIMA (1,2,2) model. Their study displayed that predicted figures follow a growing trend for the succeeding years.

Zhao Ying used an ARIMA model with time-series data of actual GDP from 1954 to 2004 in China to analyze and predict the national GDP growth pattern attempted to construct a time series model that was utilized to forecast the gross domestic product of China up to the first quarter of 2009. This paper was based on figures collected from secondary sources from the years1962 to 2008, Lu got ARIMA (4,1,0), which he applied for forecasting purposes.

In developing a time series model to forecast the GDP of manufacturing and industries test and forecast GDP for Albania using quarterly data from 2003 to 2013. They applied two important time-series model: ARIMA and VAR. Their paper shows that the group of VAR models provides better results on GDP forecasting than the ARIMA model. In summary, these studies motivated me to carry out this research which analyzes the potential for GDP growth rates in Kenya.

In a study, Tsay and Tiao (1984, 1985) used ARIMA model, which is in fact fitted on nonseasonal data by identifying autoregressive and moving average terms with the help of partial auto correlation and auto correlation functions (Box and Jenkins 1970:1976,Pankratz 1991).However, in the case of seasonal data, a number of studies used filtering approach, which in fact very helpful in case of weekly, monthly, quarterly and semiannual data to estimate a model to forecast any macro variable (Liu 1989; Liu and Hudak 1992; Liu 1999).

In another research, Reynolds et al. (1995) developed automatic methods to identify as well as estimate the parameters of ARIMA model by utilizing time-series data for a single variable. In another study, Reilly (1980) used similar methodology to model macroeconomic variable like GDP. However, both the studies confined themselves only on non-seasonal time series data and restrained to predict the variable in future. However, the above-mentioned methods need a long time-series data on the macroeconomic variable in question. To estimate the model for prediction of a macro variable, a number of studies imply analytical neural network techniques, which is very effective in the case of seasonal data (Chiu et al. 1995; Cook and Chiu 1997; Geo et al. 1997; Saad et al. 1998). These types of models have got pace since the seminal paper of Granger and Joyeux (1980) and Hosking (1981). However, this neural networking approach is very difficult to applying in real life situation by the policy makers /managers due to difficult network design, training and testing are required to build the model as well as to estimate the parameters.

CHAPTER THREE

METHODS

3.1 Introduction

In univariate time series analysis, the ARIMA models are flexible and widely used. The ARIMA model is the combination of three processes:

i Autoregressive (AR) process.

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

ii Moving Average (MA) process.

$$Y_t = \mu + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - ... - \theta_q \epsilon_{t-q} \tag{2}$$

iii Differencing process.

These processes are known in statistical literature as main univariate time series models, and are commonly used in many applications.The general ARIMA model is given by:

$$Y_t = \mu + \phi_1 Y_{-1} + \phi_2 Y_{-2} + ... + \phi_p Y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - ... - \theta_q \epsilon_{t-q}, \tag{3}$$

3.1.1 Assumptions

1. The data are stationary. A stationary process has the property that the mean, variance and autocorrelation structure do not change over time.
2. The residual component $\varepsilon_t \sim N(0, 1)$

3.2 Box-Jinken's Approach

In time series analysis, the Box-Jenkins (1970) approach, named after the statisticians George Box and Gwilym Jenkins, applies ARIMA models to find the best fit of a time series model to past values of a time series.The four stages in modelling in the Box-Jenkins iterative approach are:

3.2.1 Model identification.

This involves making sure that the variables are stationary, identifying seasonality in the series, and using the plots of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) of the series to identification which autoregressive or moving average component should be used in the model. However, by the capability of R software, we employed both the algorithm of ACF and PACF to determine the class our model and apply the "$auto.arima()$" function from "$forecast$" package that automatically estimates the coefficients of the ARIMA model as well as class of ARIMA model for the GDP data.

3.2.2 Model estimation

. Using computation algorithms to arrive at coefficients that best fit the selected ARIMA model. The most common methods use Maximum Likelihood Estimation (MLE) or non-linear least-squares estimation

Autoregressive Estimation

The Autoregressive equation is given by

$$Y_t = \mu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \epsilon_t \tag{4}$$

To estimate the parameter using the Least Square Estimation, let us first consider AQR_1 model

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + \epsilon_t. \tag{5}$$

which can regarded as regression with predictor Y_{t-1} and response variable Y_t.
We solve by minimizing by sum of squares of differences :

$$\epsilon_t = (Y_t - \mu) - (\phi Y_{t-1} - \mu) \tag{6}$$

and summing from $t-1...t-n$ since we only have $Y_1, Y_2, ..., Y_n$ observations;
Let

$$S_\epsilon(\phi, \mu) = \sum_{t=2}^{n}((Y_t - \mu) - (\phi Y_{t-1} - \mu))^2 \tag{7}$$

be the conditional sum of squares.
We minimize $S_\epsilon(\phi, \mu$ given $Y_1, Y_2, ..., Y_n$ to estimateϕ and ν
Consider equation (7) with respect to μ

$$\partial\frac{S_\epsilon}{\partial\mu} = \sum_{t=2}^{n}((Y_t - \mu) - (\phi Y_{t-1} - \mu))(-1 + \phi) = 0 \tag{8}$$

$$\mu = \frac{1}{(n-1)(1-\phi)}(\sum_{t=2}^{n}(Y_t - \phi\sum_{t=2}^{n}Y_{t-1}) \tag{9}$$

For large values of n,

$$\frac{1}{n-1}\sum_{t=2}^{n}Y_t \approx \frac{1}{n-1}\sum_{t=2}^{n}Y_{t-1} \approx \bar{Y} \tag{10}$$

which reduces to

$$\hat{\mu} = \frac{1}{1-\phi}(\bar{Y} - \phi\bar{Y}) = \bar{Y} \tag{11}$$

$$\hat{\mu} = \bar{Y} \tag{12}$$

Consider equation (7), substitute equation (12 and we minimize with respect to ϕ

$$\partial\frac{(\phi, \bar{Y})}{\partial\phi} = \sum_{t=2}^{n}((Y_t - \bar{Y}) - (\phi Y_{t-1} - \bar{Y}))(Y_{t-1}\bar{Y})) = 0 \tag{13}$$

$$\hat{\phi} = \frac{\sum_{t=2}^{n}((Y_t - \bar{Y})(Y_{t-1} - \bar{Y})}{\sum_{t=2}^{n}(Y_t - \bar{Y})^2} \tag{14}$$

To estimate the ϕ for $AR_{(p)}$ model, we need to consider the second order $AR_{(2)}$ model and substitute μ with $\bar{Y}$.

$$S(\phi_1, \phi_2, \bar{Y}) = \sum_{t=3}^{n}[(Y_t - \bar{Y}) - \phi_1 Y_{t-1} - \bar{Y}) - \phi_2(Y_{t-2} - \bar{Y}]^2 \tag{15}$$

Setting $\partial\frac{S\epsilon}{\partial\phi_1} = 0$, we have

$$-2\sum_{t=3}^{n}[(Y_t - \bar{Y}) - \phi_1(Y_{t-1} - \bar{Y}) - \phi_2(Y_{t-2} - \bar{Y})][Y_{t-1} - \bar{Y}] = 0 \tag{16}$$

Re-writing equation (16) to

$$\sum_{t=3}^{n}(Y_t - \bar{Y})(Y_{t-1} - \bar{Y}) = \phi_1[\sum_{t=3}^{n}(Y_{t-1} - \bar{Y})^2] + \phi_2[\sum_{t=3}^{n}(Y_t - \bar{Y})(Y_{t-2} - \bar{Y}] \tag{17}$$

Upon dividing both sides by $\sum_{t=3}^{n}(Y_t - \bar{Y})^2$. We obtain

$$r_1 = \phi_1 + r_1\phi_1 \tag{18}$$

and also taking

$$\partial\frac{S\epsilon}{\partial\phi_2}\sum_{t=3}^{n}[(Y_t - \bar{Y}) - \phi_1 Y_{t-1} - \bar{Y}) - \phi_2(Y_{t-2} - \bar{Y}][Y_{t-2} - \bar{Y}] = 0 \tag{19}$$

Upon re-writing the equation (19)

$$\sum_{t=3}^{n}(Y_t - \bar{Y})(Y_{t-2} - \bar{Y}) = \phi_1[\sum_{t=3}^{n}(Y_t - \bar{Y})(Y_{t-2} - \bar{Y}] + \phi_2[\sum_{t=3}^{n}(Y_{t-1} - \bar{Y})^2] \tag{20}$$

Dividing equation (20) by$\sum_{t=3}^{n}(Y_{t-1} - \bar{Y})^2$, we have

$$r_2 = r_1\phi_1 + \phi_2 \tag{21}$$

Upon solving equation (18) and (21), we have

$$\hat{\phi_1} = \frac{r_1(1 - r_2)}{1 - {r_1}^2} \tag{22}$$

and

$$\hat{\phi_2} = \frac{r_2 - {r_1}^2}{1 - {r_1}^2} \tag{23}$$

Differencing process

A process Y_t is said to be ARIMA(p,d,q) if the stationary series obtained after differencing is an ARMA(p,q) given as:

$$Z_t = \triangle^d Y_t = (1 - B)^d Y_t \tag{24}$$

In ARIMA model, the random error term ϵ_t is assumed to be a $white - noise$ which is identically and independently distributed, with mean 0 and common variance, σ^2 that is,
$\epsilon_t \sim iidN(0, \sigma^2)$
The general d^th order difference can be expressed as

$$\triangle^d Y_t = (1 - B)^d Y_t \tag{25}$$

The general ARIMA(p,d,q) with an AR of order p, a MA part of order q and with a d order of differencing is given by:

$$[1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p][(1 - B^d)Y_t] = [1 - \theta B - \theta_2 B^2 - ... - \theta_q B^q \epsilon_t] \tag{26}$$

$$\phi(B)(1 - B)^d Y_t = \sigma + \theta(B)\epsilon_t \tag{27}$$

or

$$Z_t = \sum_{i=1}^{p}\phi_i Z_{t-i} - \sum_{i=1}^{q}\theta_i\epsilon_{t-i} + \sigma + \epsilon_t \tag{28}$$

Where $\phi^{'s}$ and $\theta^{'s}$ are coefficients of AR and MA process.
p and q are the number of past values of Y_t and the error term.

Moving Average Process

The MA assumes that the time series value Y_t can be expressed as a function of q past values of the random error terms in the series.

The MA_q process is always stationary regardless of the values of the weight. The general MA_q model is:

$$Y_t = \mu_t\epsilon_t - \theta_1\epsilon_{t-1} - \theta_2\epsilon_{t-2} - ... - \theta_q\epsilon_{t-q} = \theta B\epsilon_t + \mu; \tag{29}$$

$$Y_t = \mu + \epsilon_t - \sum{}_{i=1}^{q}\theta_1\epsilon_{t-1} \tag{30}$$

Or using the back-shifting operator :

$$Y_t - \mu = \epsilon_t - \theta_1\epsilon_{t-1} - \theta_2\epsilon_{t-2} - ... - \theta_q\epsilon_{t-q}; \tag{31}$$

$$Y_t - \mu = \epsilon_t - \theta_1(B)\epsilon_t - \theta_2(B^2)\epsilon_t - .. - \theta_q(B^q)\epsilon_t; \tag{32}$$

$$Y_t - \mu = (1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q)\epsilon_t; \tag{33}$$

$$Y_t - \mu = \theta(B)\epsilon_t; \tag{34}$$

Where $(Y_t - \mu)$ is a white-noise.
μ is the mean of the time series.
$\theta(B)$ is the Moving Average Operator of order q define by:

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q \tag{35}$$

Also,given that an MA_1 made $Y_t = \epsilon_t - \theta\epsilon_{t-1}$.
The invertible MA_1 can be expressed :

$$Y_t = \epsilon_t - \theta Y_{t-1} - \theta^2 Y_{t-2} - ... \tag{36}$$

of infinite order.
The method of least squares can be carried out by by choosing θ that minimizes:

$$S_\epsilon(\theta) = \sum{}_{i=1}^{\infty}{\epsilon_t}^2 = \sum{}_{i=1}^{\infty}[Y_t + \theta Y_{t-1} + \theta^2 Y_{t-2} + ...]^2 \tag{37}$$

Where $\epsilon_t = \epsilon_t(\theta)$ is a function of the observed series and unknown parameter θ
For Moving Averages, we result to numerical optimization, because:
First, the least square is non-linear in the parameter and its not possible to minimize $S_\epsilon(\theta)$ by taking a derivative with respect to θ and setting it to zero. We thus consider evaluating $S_\epsilon(\theta)$ for a single value of θ for our series.
We could rewrite the MA_1 model to:

$$\epsilon_t = Y_t + \theta\epsilon_{t-1} \tag{38}$$

To calculate $\epsilon_1, \epsilon_2, ..., \epsilon_n$ recursively if we have the initial value ϵ_o. A common approach is to set $\epsilon_o = 0$. The expected value for the mean of the random part. Thus;

$$\epsilon_1 = Y_1, \epsilon_2 = Y_2 + \theta(\epsilon_i...\epsilon_n) = Y_n + \theta\epsilon_{n-1}. \tag{39}$$

We then calculate $S_\epsilon(\theta) = \sum{}_{i=1}^{\infty}\epsilon_t^2$
Conditioning $\epsilon_o = 0$ for higher order MA model, we compute,

$$\epsilon_t = \epsilon_t(\theta_1, \theta_2, ...\theta_q) \tag{40}$$

Recursively from

$$\epsilon_t = Y_t + \theta\epsilon_{t-1} + ... + \theta_q\epsilon_{t-q} : \tag{41}$$

with $\epsilon_o = \epsilon_{-1} = ... = \epsilon_{-q-q} = 0$
The sum of squares is minimized jointly in $\theta_1, \theta_2, .., \theta_q$ in a multivariate methods.

3.2.3 Model checking

This includes testing whether the estimated model conforms to the specifications of a stationary univariate process. In particular, the residuals should be independent of each other and constant in mean and variance over time. plotting the ACF and PACF of the residuals are helpful to identify misspecification. If the estimation is inadequate, we have to return to step one and attempt to build a better model. However, utilizing the capability of R software, we were able to visualize the residual plots at once and arrive at a conclusion.

3.2.4 Forecasting

This comes in when the auto-fitted ARIMA model conforms to the specifications of a stationary univariate process, then we can use this model for forecasting for the next 10 years,that is from the year 2020 to2029.

CHAPTER FOUR

DATA ANALYSIS AND RESULTS

The source of data used in this paper is World Bank (WDI) database. Refer from the Appendix for the complete data set.

4.1 Descriptive Statistic

n	mean	sd	median	trimmed	mad	min	max
60	17961386194	23540051859	8061149768	12821173726	8885755304	791265458	95503088538

range	skew	kurtosis	se
94711823080	1.75	2.11	3039007627

The variable used in the analysis is the GDP in US$,cover the period from 1960 to 2019 as shown in Figure 1

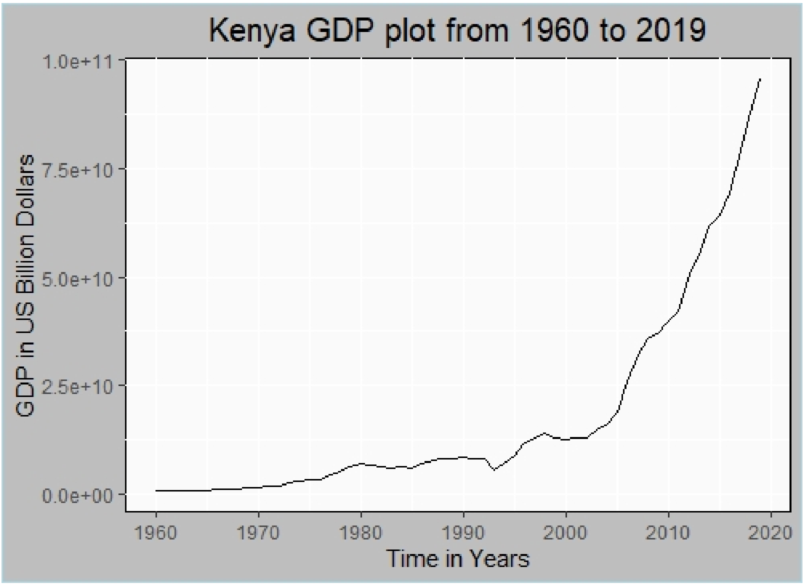


Figure 1: Kenyan GDP plot

The ARIMA approach is an iterative four-stage process of stationary, identification, estimation and testing.

4.2 Testing for Stationary

Modelling and fitting a time series model to a data requires the data to be stationary. Having a look at the plot on Figure 1, we can visualize at glance an increasing trend.

4.2.1 Augmented Dickey-Fuller Test

$AugmentedDickey - FullerTest$
$data : TSDTA$
$Dickey - Fuller = 4.0291, Lagorder = 3, p - value = 0.99$
$alternativehypothesis : stationary$
The Dickey-Fuller state Hypothesis:
H_o: non stationary
H_1: stationary.
Conclusion: The data is not stationary since P-value > 0.05 hence we fail to reject H_o

4.2.2 ACF and PACF plots

Visualizing Figure 2 below , we can conclude that the coefficients of autocorrelation (ACF) starts with a high value and exponentially declines slowly as number of lags increases, indicating that the series is non-stationary. Therefore before proceeding with analysis, we have to make the data stationary .Thus the series must be configured in first or second differences. .

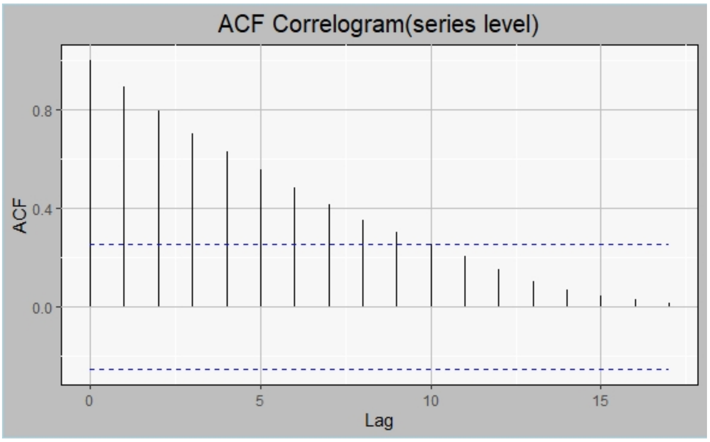


Figure 2: ACF plot

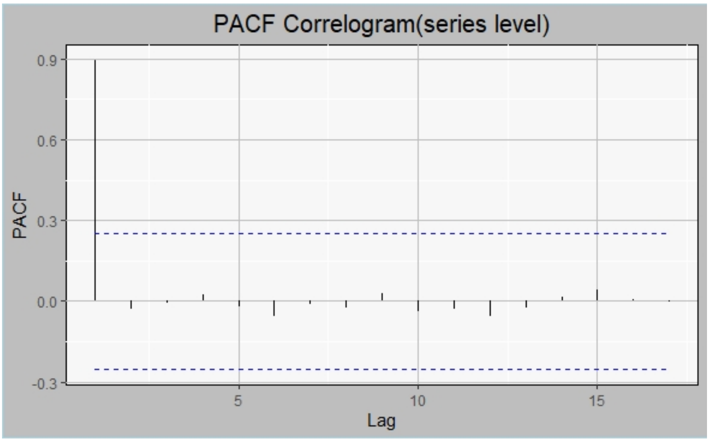


Figure 3: PACF plot

4.2.3 Making series Stationary

Figure 4 shows that after taking first difference after log transformation, we see that the statistical property like mean and variance are not constant, therefore the data is not yet stationary. We take the second difference

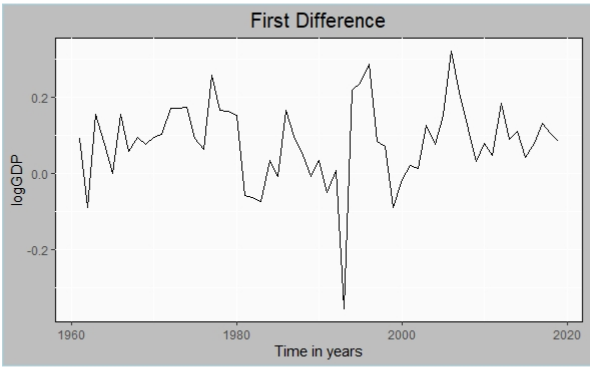


Figure 4: First differencing log transformation

Figure 5 shows that after taking the second order difference(period to period change), the time series seem to have been made stationary; there is no clear upward and downward trend. As a consequence, we can aplly this data for determining ARIMA(p,d, q) model. Hence our difference $I = 2$. However, taking higher order differences like 3, 4 or more may lead to having wrong data and brings a little to no change. Therefore, its best to use difference of order 2

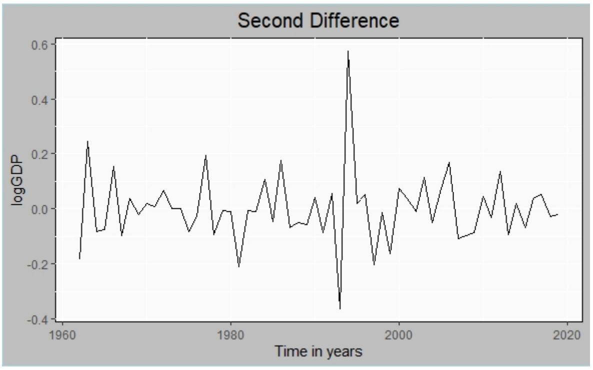


Figure 5: Second differencing log transformation

4.2.4 Unit Root Test

The hypothesis is that
H_o : Non Stationary(Inverse root > 1)
H_1 : Stationary(Inverse root < 1)
The inverse root test plot in Figure 6 below shows that the Time series data is indeed stationary by since the inverse roots are within the unit circle. This means rhat we reject H_o

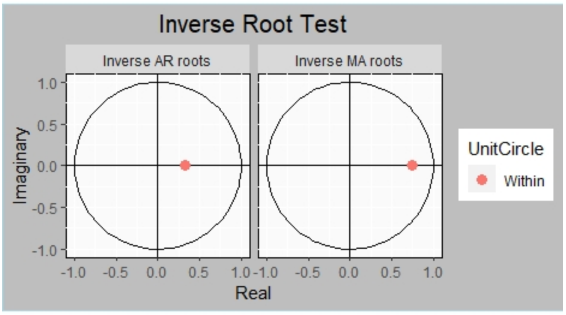


Figure 6: Inverse Root Test

4.3 Model Selection

Previous attempts to achieve stationary time series data shows that the order of difference , $I = 2$. Since our modelling employs ARIMA(p,d,q), thus $d = 2$. Keeping d constant, and iteratively fitting the stationary time series data for 8 ARIMA models, we find that the ARIMA(1,2,1) has the least Akaike's Information Criterion value of -89.6 among the eight fitted models as shown in Table 1 below.

ARIMA(p,d,q)	**AIC**
(0, 2, 1)	-86.57
1, 2, 0	-79.65
1, 2, 1	-89.6
2, 2, 0	-77.73
2, 2, 1	-88.07
2, 2, 2	-87.91
2, 2, 0	-77.73
0, 2, 2	-88.36

Table 1: Tested ARIMA(p,d,q) models

Consequently, determining the order q of $MA_{(q)}$ and p of $AR_{(p)}$ can be arrived at by visualizing the plots of ACF and PACF of stationary series data and determining the two orders, respectively from these plots. The first significant spikes at lag 1 in both Figure 7 and 8 determines the values of p and q, ignoring the significant spike at lag zero

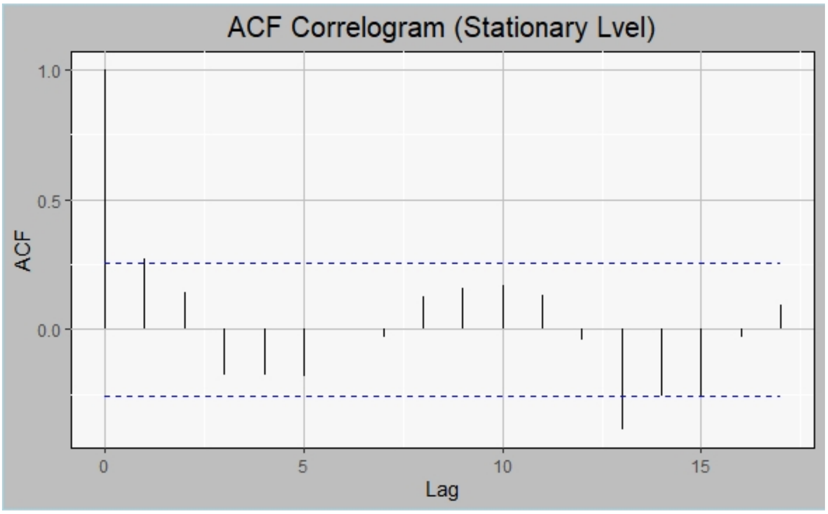


Figure 7: ACF Correlogram for Stationary Series

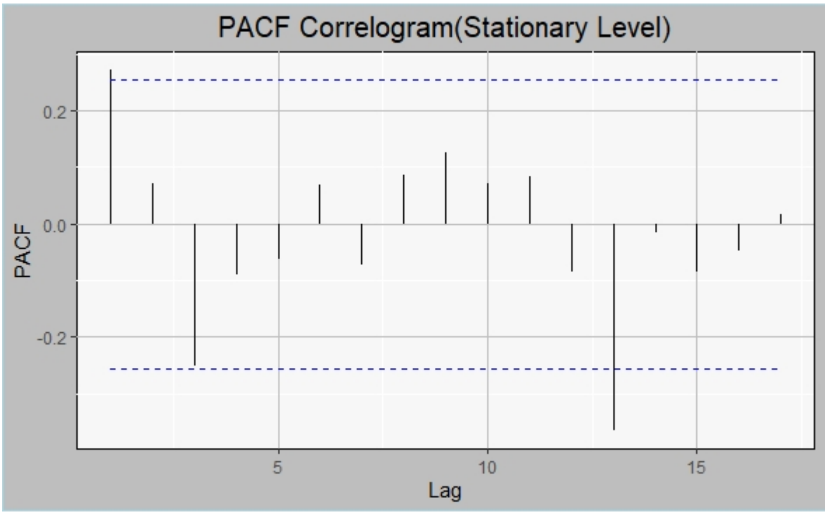
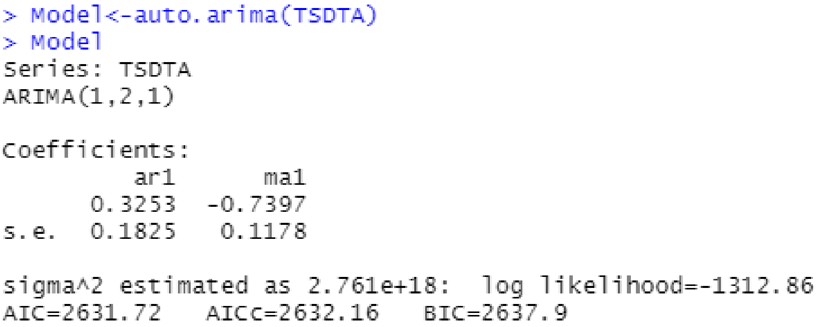


Figure 8: PACF Correlogram for Stationary Series

Hence; $q = 1$ and $p = 1$

4.4 Model Estimation

The estimation of model parameters involves using maximum likelihood or method of least squares. Our study employed t the capability of *auto.arima*() from *forecast* package to arrive at the parameters as shown below.



```
> Model<-auto.arima(TSDTA)
> Model
Series: TSDTA
ARIMA(1,2,1)

Coefficients:
         ar1      ma1
      0.3253  -0.7397
s.e.  0.1825   0.1178

sigma^2 estimated as 2.761e+18:  log likelihood=-1312.86
AIC=2631.72   AICc=2632.16   BIC=2637.9
```

This leads us into creation of full ARIMA(1,2,1) given as:

$$Y_t = 0.3253X_{t-1} - 0.7397\epsilon_{t-1} + \epsilon_t; \tag{42}$$

where $\epsilon_t \backsim N(0,1)$

4.5 Diagnostic Checking

The ARIMA(1, 2, 1) model was then subjected to diagnostic checking to test if the residuals are white noise. The Figure 8 below of residual plots shows that residual appear to be centered around 0 as noise, with no pattern. The ARIMA(1,2,1) model is a good fit for the data.

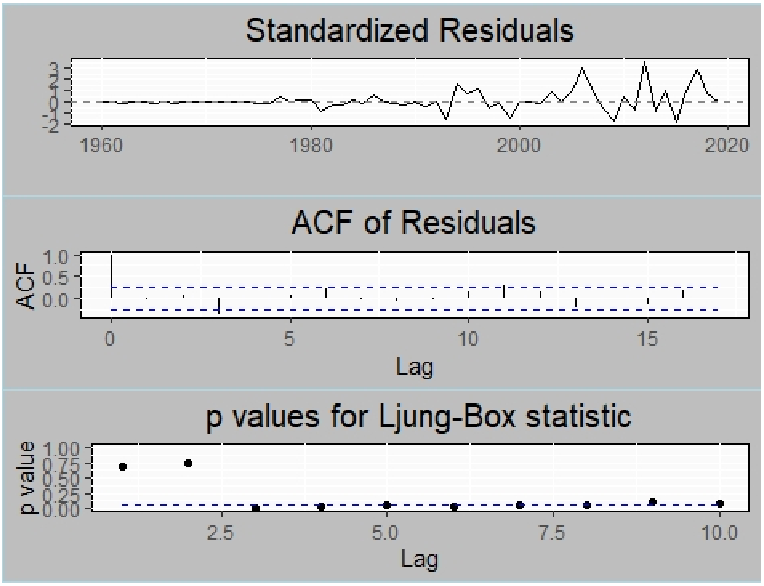


Figure 9: Diagnostic Check plot

The Figure 10 below shows how that the data fits the model. Hence our choice of ARIM(1,2,1) is significantly a good fit.

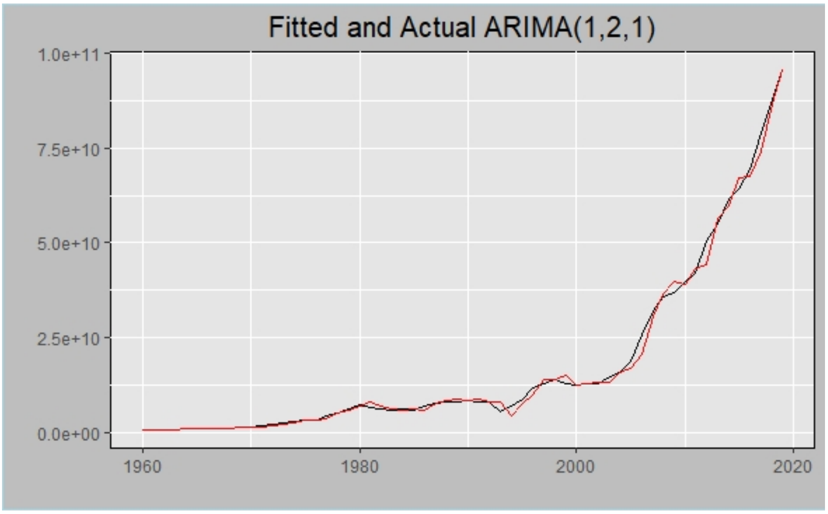


Figure 10: Fitted - Actual ARIMA(1,2,1) plot

4.6 Forecasting

In the final stage of Box-Jenkins, we make forecast of the Kenyan GDP by using the ARIMA(1,2,1) as shown in the table below.

Year	*Predicted GDP(US$)*
2020	102779462, 534
2021	109910038865
2022	116993180558
2023	124060889598
2024	131123577692
2025	138184632242
2026	145245155325
2027	152305505499
2028	159365799417
2029	166426075032

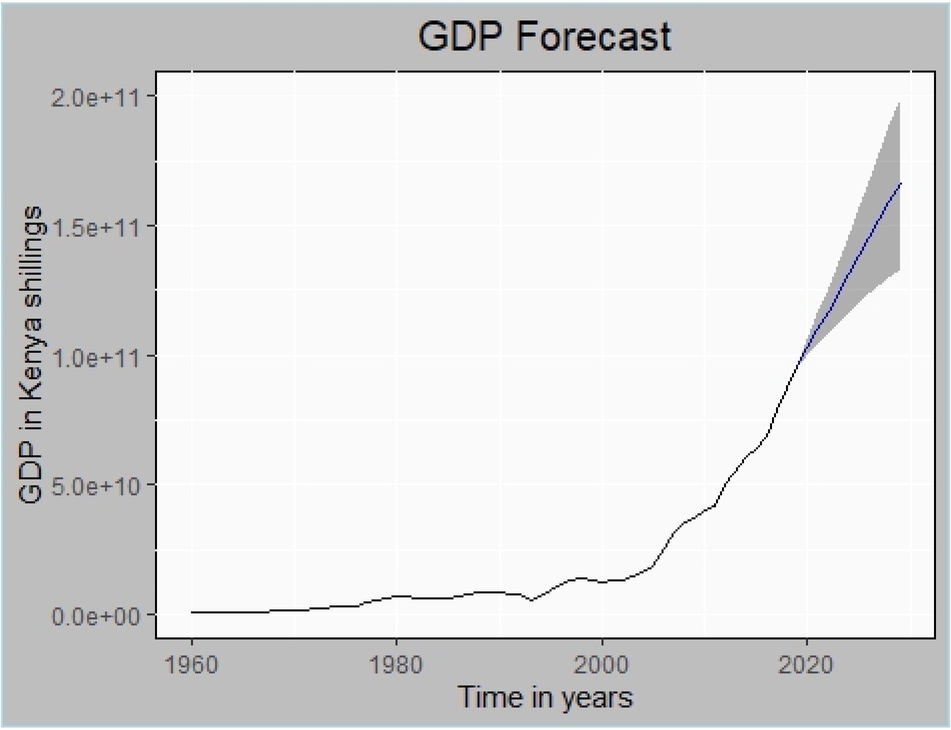


Figure 11: Forecast GDP (US$) from 2020 to 2029

In the above Figure 11, the GDP is plotted against the respective years. Here, we showed the previous trend and future trend together. The GDP shows an upward trend and from the above graph, we see that the past values(data) fits very well, that is, the estimated line represents the data very closely.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 Summary

The aim of the study was to model and forecast Kenyan GDP based on the Box-Jenkins approach based on the annual data (from 1960 to 2019). The four stages of the Box-Jenkins approach are conducted to obtain an appropriate ARIMA model for the Kenyan GDP and forecast for the next ten year. Time series and correlogram plots were used in testing the stationarity of the data.Also, the non-linear least square was used in estimating the model. Using the different goodness of fit measures (MSE,AIC and BIC),the various ARIMA models with different order of Autoregresive and Moving Average terms were compared. We find the best that the best model is ARIMA(1,2,1),because it had the minimum values of MSE, AIC, BIC. Moreover, we expect that the Kenyan GDP will continue to rise exponentially according to the forecasted value of our model.

5.2 Conclusion

Forecasting micro-economic indicators present a dear picture of what the situation of the economy will be in the future. Having the relevant model to forecast these micro-economic variables is significant for policy-makers and the government in allocating resources efficiently to formulate better policies. This paper forecast increasing growth in Kenya's GDP beginning in the year 2020 to 2029. Precisely, the manager who is planning to invest in the expansion of existing business or in new projects will benefit tremendously since the finding will help them to portrait the picture of economic condition of Kenya. Taking into account the values of GDP forecasted, we can see that the GDP exponentially will increase with time until the year 2029. Hence, we expect Kenya to have a GDP value of about $166, 426, 075, 032$ Billion US dollars.

5.3 Recommendation

1. The use of ARIMA model is a highly flexible tool to forecast Kenyan GDP rate if there is no government's intervention that will change this trend.

2. Many empirical studies have been done regarding the effectiveness of ARIMA model in economic forecasting and as a result its an essential component of all forecasting techniques. The proper evaluation of ARIMA model is necessary to study and carry out the forecast process.

3. The properly selected model using the ARIMA method enhances the predictability of the models and assist the players in making sound government policy such as imposing different level of tax to control production and importation or exportation of a certain product ,proper decision making for example whether to implement a business idea and strategic plans like Vision 2030 that seek to enhance general level of development in Kenya.

REFERENCES

Box, G. E. P.and Jenkins G. M. (1976). *Time series analysis. Forecasting and Control.*

C. Dritsaki, (2015). *Forecasting real GDP rate through econometric models: An empirical study from Greece, Journal of International Business and Economics, 3, 13-19.* https://doi.org/10.15640/jibe.v3n1a2

Dickey, D. A. and Fuller W. A. (1979).Distribution of the Estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association, 74(366), 427431. The European System of Accounts ESA 1995, Eurostat, 1996.*

Maity, B., and ChatterjeeB. (2012). *Forecasting GDP growth rates of India: An empirical study. International Journal of Economics and Management Sciences, 1(9), 52-58.*

12M. S. Wabomba, M. P. Mutwiri and M. Fredrick (2016). *Modeling and Forecasting Kenyan GDP Using Autoregressive Integrated Moving Average (ARIMA) Models, Science Journal of Applied Mathematics and Statistics 4, 64-73.*,https://doi.org/10.11648/j_sjams
Box,Gand, JenkinsG.(1970).*Time series analysis: Forecasting and control,San Francisco Holden-Day.*

Box, G.E.P et. al(1994).*Time series Analysis:Forecasting and control,*Third Edition, Engelwood *Cli,*NJ Prentice Hall.

Burnhamand Anderson(2004).Multimodel inference: *Understanding AIC and BIC in model selection.*

APPENDIX

Year	Kenyan GDP(US\$$)
1960	791265458
1961	868111400
1962	792959472
1963	926589348
1964	998759333
1965	997919320
1966	1164519673
1967	1232559506
1968	1353295458
1969	1458379415
1970	1603447357
1971	1778391289
1972	2107279157
1973	2502142444
1974	2973309272
1975	3259344936
1976	3474542392
1977	4494378855
1978	5303734883
1979	6234390975
1980	7265315332
1981	6854491454
1982	6431579357
1983	5979198464
1984	6191437070
1985	6135034338
1986	7239126717
1987	7970820531
1988	8355380879
1989	8283114648
1990	8572359163
1991	8151479004
1992	8209129172
1993	5751789915
1994	7148145376
1995	9046326060
1996	12045858436
1997	13115773738
1998	14093998844
1999	12896013577
2000	12705357103
2001	12986007426
2002	13147743911
2003	14904517650
2004	16095337094
2005	18737897745
2006	25825524821
2007	31958195182
2008	35895153328
2009	37021512049
2010	40000088347
2011	41953433591
2012	50412754861
2013	55096728048
2014	61448046802
2015	64007750169
2016	69188755364
2017	78965004656
2018	87778582964
2019	95503088538

Table 2: Gross Domestic Product data.