Mount Kenya University

FORECASTING MONTHLY CONSUMER PRICE INDEX (CPI) FOR NAIROBI COUNTY USING ARMA MODELS.

BY

JEREMIAH MAINA

A project submitted to the school of pure and applied sciences studies in partial fulfillment of the requirements for the award of Bachelor science in Applied Statistics at Mount Kenya University

Table of Contents

Abbreviations

CPI –Consumer Price Index

ARMA-Autoregressive Moving Average

AR-Auto-regression

MA-Moving Average

KNBS- Kenya National Bureau of Statistics

E(x) –Expectation of X

Var (X) –Variance of X

CHAPTER ONE

Introduction

1.1 Background information

Inflation is a critical indicator that affects the overall performance of an economy, one of the most popular measures of inflation is Consumer Price Index (CPI), (Karabell, 2015). In Kenya there has been a persistent problem, with monthly consumer price index(CPI) changing from 5.6% in 2015 to 5.9% in 2020. Accurate forecasting of CPI is essential for policy makers, businesses and investors to make informed decision. This study aims to forecast monthly consumer price index (CPI) for Nairobi county using ARMA models (Corazza & Pizzi, 2014).

There is several definition of CPI, according to Wikipedia Consumer Price Index (CPI), is the price of weighted average market basket of consumer goods and services purchased by households while the Bureau of Labor Statistics (BLS) they defines CPI as a weighted average of price for basket of goods and services. The basket is updated periodically to reflect changes in consumers spending habit (Bresnahan & Gordon, 2008). The prices of goods and services in the basket are collected monthly from a sample of retail and services establishment. The prices are then adjusted for changes in quality or features use.

According to the Kenya National Bureau of Statistics (KNBS), the CPI for Nairobi has been increasing steady over the past few years, with an average annual increase of 5.6% from 2015 to 5.9% in 2020(KNBS,2020). This trend is expected to continue in the future, making it an important to develop an accurate forecasting model for the CPI. (Government Publishing Office, 2010) Additionally, previous studies by (Ghysels & Marcellino, 2018) have shown that ARMA models are effective in forecasting economics variables such as inflation.

Forecasting model for monthly CPI in Kenya using combination of ARIMA and regression analysis (Andrle et al., 2013). The authors found that their model produced accurate forecast for monthly CPI in Kenya, with a mean absolute percentage error of 2.15%. The authors suggested that their model can be used to inform monetary policy decision in Kenya.

Developing forecasted model for monthly CPI in Kenya using a combination of ARIMA and neural network model (Thampi et al.2021). The author found that their model produced accurate forecast for monthly CPI in Kenya, with a mean absolute percentage error of 1.6%. And suggested that their model can be used to inform monetary policy decision in Kenya.

The role of consumer's price index in economic management in Kenya (Schnepf, 2011). This article explores the significance of CPI as an economic indicator in Kenya. This discusses how CPI affect economic policy, consumer behavior and inflation rates and highlights the important of accurate CPI measurement in economic management (Blinder, 2013)

The importance of consumer price index in Kenya economic development discusses the role of CPI in Kenya's development, with focus on Nairobi County (Wakiuru et al., 2023). (Bordo & Orphanides, 2013) explain how CPI data is used to measure inflation, adjust wages and calculating economic growth, emphasizes the importance of accurate CPI measurement for economic stability.

1.2 Problem Statement

Several methods such as regression analysis have been used to forecast CPI for Nairobi County by (Dierckx et al., 2021), and found that this method had less precision. The ability to forecast CPI for Nairobi County has resulted effective economic policies and decisions in Kenya (International Monetary Fund, 2016). CPI is used to measure the average prices of goods and services consumed by households (Boyes & Melvin, 2015). This proposal aims to develop an accurate forecasting model for the monthly CPI in Nairobi County using ARMA model.

The current CPI forecasting methods used in Kenya rely heavily on traditional time series model such as ARIMA which have been shown to less accurate than other models such as ARMA (Berg & Portillo, 2018). ARMA model have the advantage of being more flexible and able to capture more complex pattern in the data which can results in more accurate forecast. (Yao et al., 2022).

The accuracy of CPI forecasts is particularly important in developing countries like Kenya where high inflation rates can have significant impact on the economy and the wellbeing of the population (Rocha et al., 2022). Accurate CPI forecasts can help policymakers make informed decision about monetary policy which can help stabilize the economy and reduce the negative impact of inflation on the population (Chrystal, 2020).

The current CPI forecasting used in Kenya do not take into account the effect of external factor such as global economic condition and political instability, which can have a significant impact on inflation rate in Kenya (Fund., 2015). ARIMA model can be adapted to include external factors, which can result in more accurate factors that take into account the broader economic environment (Morley, 2022).

Previous studies have concentrated forecasting CPI in Kenya generally by using different models such as ARIMA and they have shown mixed results with some studies reporting high accuracy while others reporting poor performance (Zopounidis et al,2018.). Therefore, there is need for further research to explore the effectiveness of time series model in forecasting CPI by concentrating county level.

This study aims to feel this gap by applying ARMA to forecast monthly CPI of Nairobi County. This research will provide valuable insight into the behavior of Consumer Price overtime by using ARMA model we will identify trends, patterns in the data, estimate future CPI values.

1.3 Justification

The purpose this project is to predict monthly CPI of Nairobi County. This study aims to provide accurate and reliable method for predicting future CPI trends which can help policy makers and businesses make informed decisions. Forecasting accuracy has been proven to be effective in forecasting time series data ,such as CPI .By using ARMA model ,this proposal can improve the accuracy and reliability of CPI forecast for Nairobi county ,which can benefit policy makers ,businesses and consumers .Monthly CPI data for Nairobi county is readily available in Kenya bureau of statistics (KNBS) .By using this data ,the proposal can provide valuable insight for the inflation trends and patterns in the economy of Nairobi County which can inform future research and policymaking. The resulting model will be useful for policy makers, business and consumers to make informed decisions based on predicted inflation trends.

1.4 Objective

1.4.1 General objective

To forecast monthly Consumer price index (CPI) for Nairobi County using ARMA model

1.4.2 Specific objectives

- To test for stationarity and Normality
- To fit an approximate ARMA model
- To forecast monthly CPI using fitted model

1.5 Research Structure

In the first chapter an introduction of the study presented giving explanation on the inflation giving background of CPI in Nairobi County. On problem statement, it explains the problem encountered in Nairobi county and giving out a solution by developing an accurate forecasting model for the monthly CPI in Nairobi county using the ARMA model. It also explains the purpose of the proposal on predicting of future CPI. General and Specific objectives are also given here.

The second chapter on literature review; it involves reviewing existing research on ARMA models and their application in forecasting consumers price index (CPI) in different context.

Hypothesis: The null hypothesis of this proposal is that there is no relationship between past values of CPI and future value of CPI in Nairobi County.

Third Chapter on methodology: This involves specifying the type of ARMA model to be used, the data sources and data collection procedures. The methodology will also outline the steps for estimating the model parameters, testing the model assumptions, and evaluating the forecasting accuracy.

Forth chapter on Data collection and analysis: This will involve gathering monthly CPI data for Nairobi County over a specific period, as well as any other relevant data that may be needed to estimate the ARMA model. Data analysis involve fitting the ARMA model to the data, testing the model assumption, and evaluating the forecasting accuracy using various performance metrics.

Fifth chapter on Results, conclusion and recommendation: The results of the study will include the estimated ARMA model parameters, the test of the model assumption, and the evaluation of the forecasting accuracy. The conclusion will summarize the findings of the study, discuss their implications, and offer recommendations for future research. The recommendations may include suggestions for improving the ARMA model, refining the data collection procedure, or exploring alternative forecasting method.

CHAPTER TWO: LITERATURE RIVIEW

Introduction

Consumer price index (CPI) is widely used as economic indicator to measure the average change in price of goods and services consumed by households over a given period of time for policy makers, businesses, and consumers. Accurate and timely forecasting of CPI is crucial because it gives insight into inflationary trends and informs decision related to monetary policy and investment. In this literature review we aim to explore the use of ARMA model for forecasting the monthly CPI in Nairobi County. ARMA model as in time series models captures the linier dependency between past observations and future value of time series. We seek to evaluate the effectiveness of ARMA model in forecasting CPI in Nairobi County by analyzing previous research and to identify the factors that affect the accuracy of this model. Again we aim to identify the gaps in existing literature, and suggest area for future research. In these sections, we provide a critical analysis of the literature on CPI forecasting ARMA models, and the application of ARMA models in CPI forecasting in the different regions and states. We finally summarize our findings and discuss their implications for future research and practice.

2.1 Theoretical Review

2.1.1 Consumer price index (CPI)

Many authors have studied the dependency between CPI from other economic variables and time models (Kryvinska & Greguš, 2022). (Xiaolian & Chen, 2013) used univariable ARMA models and found that it provided accurate forecast CPI. Additionally, (Géron, 2022) used a seasonal ARMA model and found that they provided accurate forecasting of CPI. According to (Cord, 2023) used multivariate ARMA model and found that it also provided accurate forecast of CPI. Also (Sud et al., 2020) have presented a forecast model for CPI in china for a period (1995-2008) and show that ARMA model has a forecast accuracy relatively high.

Changes in prices of goods and services

According to (EREN, 2018) who used ARMA model to forecast monthly CPI in Taiwan, found that ARMA model provided accurate forecast for CPI and was able to capture the effect of changes in prices for goods and services on CPI. (Beckers & Beidas-Strom, 2015) used ARMA to forecast monthly CPI in Iran and found that the effect of changes in prices for goods and services on CPI. In the work of (Abdalla et al., 2023) used ARMA to examine the relationship between CPI and food prices in Pakistan found a significant relationship between food prices and CPI. (Ben Cheikh et al., 2018) found a significant relationship between oil prices and CPI, suggesting that changes in oil prices have strong impact on CPI. According to work of (Erlich Ron & Gindi, 2023) who examined the relationship between CPI and house prices in China. He found that significant positive relationship between housing prices and CPI, suggesting that changes in housing prices have a strong impact on CPI. Another study by the University of Nairobi found that “changes in prices of goods and services in Nairobi County are also influenced by changes in the level of economic activity, as measured by factors like GDP and industrial output” (UN,2018). These factors can be incorporated into an ARMA model to forecast future CPI values.

Changes in exchange rate

This is another factor that influence CPI. It is a term used to describe the value of one currency in relation with another. It is an important factor that can influence CPI since changes in exchange rate can affect the price of imported goods. In corporation exchange data in ARMA model can help capture the effect of exchange rate, changes on CPI and enhances the accuracy of CPI forecast (Enríquez-Díaz et al., 2021). Considering the work of (Information Resources Management Association, 2022) used ARMA model to forecast monthly CPI in Kenya, found that the exchange rate has a significant impact on CPI. (Ganesh, 2012) used ARMA to forecast monthly CPI in Nigeria and found that the ARMA model was able to capture the effect of changes in exchange rate.

Unemployment

Unemployment is a factor in CPI. when unemployment is high people have less money to spend which can lead to lower demand for goods and services (Higgins, 2018). This can cause prices to fall, which can lower the CPI. However, there are also cases where higher unemployment can lead to higher inflation such as wages rise and labor shortages (Daniel Jay Richards et al., 2016). According to study by the bank of Canada, there is a positive relationship between unemployment and inflation in the short run, but this relationship weakens over time and becoming statistically insignificant after about two years (Menzies & Hague, 2016). while unemployment can be factor in CPI, the effect may not be last-longing (Abdih et al., 2016).

Government spending, monetary policy and Interest rate.

According to a study by the central bank of Kenya.” the prices of goods and services in Nairobi County are influenced by a variety of factors including government spending, monetary policy and interest rates” (CBK,2020). Another study of African Development Bank found that “changes in the price of goods and services can be influenced by changes in the level of Government spending, as well as changes in monetary policy and interest rates” (AfDB,2019).

2.1.2 ARMA Model

ARMA model involve combination of autoregressive (AR) terms and Moving average (MA) terms to model the time series data (Durmus, 2023). The AR term capture the effect of past value of the series in its current value, while the MA term capture the effect of past error on the current value (Rosin et al., 2019). By fitting the ARMA model to the data, we can make predictions about future values of the series. Previous studies have shown mixed results in the application of ARMA to forecast CPI. A study by (Marwala, 2013) ARMA models outperformed other models in forecasting CPI in united states, while a study by (Mandal & De, 2022) found that ARMA model performed poorly in forecasting CPI in Ghana. These differences may be due to difference in data sources and model specification and estimation methods

2.1.3 Time series analysis

Time series analysis is statistical method used to analysis the model time dependent data (Wei, 2019). It involves studying the pattern, trends, and cycles in the data to make predictions about future values. It can be used to model and forecast future values of CPI based on its past values. its important tool in many fields including economic, finance and engineering (Management Association, Information Resources, 2015). Time series analysis involves collection of data. The data can be collected in several ways including through survey and other data collection methods (Patel, 2022). Ones the data is collected, it is important to clean and reprocess the data to remove any outlier or errors. Time series data in characterized by four components: Trend, seasonality, cyclicity and randomness (McCuen, 2016).

2.2 Conceptual Framework

The conceptual frame work of this project is based on the following key variables: consumer price index (CPI), time series data, and ARMA models. The relationship of these variables is explained using the flow diagram below.

Relationship between ARMA models and CPI

The relationship between the Consumer Price Index (CPI) and ARMA models is fundamental for understanding and predicting inflation. The CPI tracks changes in the prices of goods and services over time, while ARMA models analyze and forecast these movements based on historical data patterns. By applying ARMA models to CPI data, economists can identify trends, assess inflationary pressures, and make informed decisions regarding monetary policy, investments, and economic planning.

The consumer price index (CPI) is the depended variable in this study and represents the overall price level of goods and services in Nairobi County. The CPI is measured on a monthly basis, and the data used in this study is time series data.

ARMA models are the independent variables in this study and are used to model the relationship between past and present values of the CPI. The ARMA models are used to identify the patterns, trends and cycles in the CPI data and to make predictions about future values.

The conceptual framework provides a basis for understanding the relationship between the key variables in the study and help to guide the development of research methods and data analysis techniques. The frame work also helps to identify potential areas for further research, as the use for the alternative modeling techniques or the inclusion of additional variables in the analysis

Assumptions

The conceptual frame work is based on the following assumption: The CPI is stationary, and the ARMA models are appropriate for modeling the data. The framework also assumes that the data is accurate and reliable, and that the ARMA models are correctly specified.

CHAPTER THREE: METHODOLOGY

Introduction

In this chapter we will discuss the concept of time series analysis in the application of ARMA models in forecasting CPI, by examining the stationarity and normality of the data. We will use ARMA model to model the stationary series and obtain forecast to the future CPI values.

3.1 Data source

The data source for our analysis is Kenya National Bureau of Statistics (KNBS) which provides monthly CPI data from January 2012 to December 2022.

3.2 Time series analysis

This time series represent a sequence of observations taken over a time or space, the time can be months, year or even at regular times intervals such as seconds. Time series is a bivariate distribution of variable time (t) and observations (x). Time series used data points that are collected at regular intervals over time. The data points can be used to analyze patterns and trends in the data as well as forecast future values.

Time series analysis involves using statistical methods to analyze and model time series data with the goal of analyzing patterns and relationships in the data to make predictions about future values. MA, AR, ARIMA, GARCH, ARMA are examples of time series models but we will mainly focus on the MA, AR and the ARMA models.

3.3 Lag

This is the time interval between two related variables. A lag of one period means that the value of one variable is correlated with the value of the other variable one period earlier. For example, lag 1 is the time interval between X_t , X_{t-1} and lag 2 is the difference between X_t and X_{t-2}.

3.4 Differencing

This is a special type of a filter applied to a given time series until it becomes Stationary. It involves subtracting the current values of the series from the previous one. That is, from the first differencing of a time series is the series of changes from one period to the next. If y_t denotes the value of the time series Y at period t, then the first difference of Y at period t is equal to

$\Delta Y_t = Y_t - Y_{t-1}$. This shows that the first differencing filter removes a linear trend. The second differencing filter which is $\Delta^2 Y = Y_t^{(1)} - Y_{t-1}^{(1)}$,it is useful for converting a non-stationary time series to a stationary form of time series. For this both the mean and dispersion may change over time. The mean may be raising and the variability may be growing. If the mean is changing the pattern is eliminated by differencing ones or twice. if the variability changes, algorithm transformation may be used to make the process stationary.

3.5 Stationary and non-stationary

A time series is stationary is if it has no systematic change in mean (there is no trend). No systematic change in variance and strictly there is no periodic variation. A non-stationary series, on the other hand has mean, variance and co-variance that change over time. A non-stationary series contains trend, seasonality, cyclic effects or combination of this components. Therefore, to change a non-stationary process to stationary one of the components causing the non-stationarity must be removed. A non-stationary process with a deterministic trend becomes stationary after trend is removed; A process known as detrending. $Y_t = t + t$, for example, is transformed into stationary process by subtracting the trend t_n: $Y_{t-t} = +t$ .When detrending is used to convert a non-stationary process to stationary one no observation is lost. There are two main forms of stationarity.

Stationarity in the strong or strict sense

The series for which the probability distribution of $X_{t1}, X_{t2} \quad \ldots\ldots X_{tn}$ is the same as the probability distribution of $X_{t1+h}, X_{t2+h}, X_{t3+h}, \ldots\ldots, X_{n+h}$ for any finite set of integers $t_1, t_2, t_3, \ldots\ldots, t_n$ and for all h is called stationary in the strong sense, that is,

F(X $X_{t1}, X_{t2} \ldots\ldots\ldots\ldots X_{tn}$,X,……..X) =F($X_{t1+h}, X_{t2+h}, \ldots\ldots, X_{tn+h}$) $\forall$ { $t_1, t_2, \ldots.., t_n$ }

If n=1(only one random variable),F(X_t)=F(X_{t+h}) which implies that the probability distribution of the process does not change with time ,E(X_t)=μ and Var(X_t)= σ^2 both constants which do not depend on t.

Stationarity in the weak sense

A process is said to be stationary in the weak sense if its mean is constant and its auto covariance function depends only on the lag. Stationarity in the weak sense has four cases, that is,

a) E(X_t) is constant and Var (X_t) is not constant, here the predictions are not reliable or accurate.

b) E(X_t) is not constant and Var (X_t) is constant, here the time series is weakly stationary in variance but not in mean. Predictions here are unreliable.

c) Both E(X_t) and Var (X_t) are not constant, the series is not stationary at all.

d) Both E(X_t) and Var (X_t) are constant, the predictions here are accurate.

Stationarity test

Augmented Dickey-Fuller (ADF)

We use Augmented Dickey-Fuller (ADF) to test for stationarity. Augmented Dickey Fuller (ADF) is a statistical test used to determine whether a unit root is present in a time series dataset. The ADF test is based on the following autoregressive model.

$$\Delta y_t = \alpha + \beta_t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \delta_2 \Delta y_{t-2} + ... + \delta_p \Delta y_{t-p} + \varepsilon_t$$

Where

- y_t is the value of the time series at time t.
- Δy_t represents the differenced series (i.e. $y_t - y_{t-1}$).
- α is a constant.
- β is the coefficient associated with a time trend .
- γ is the coefficient of the lagged level of the series .
- $\delta_1 \delta_2 ... \delta_p$ are coefficients associated with the lagged differences of the series .
- p is the lag order chosen based on criteria like AIC .

- ε_t is the residual term or error term.

The null hypothesis of the ADF test is that $\gamma = 0$ (there is a unit root, and the time series is non-stationary)

The alternative hypothesis is that $\gamma < 0$ (the time series is stationary)

The ADF statistic is then computed from the regression output, and the significance level is used to determine whether to reject the null hypothesis.

The expression for the ADF statistic depends on the version of the test being used (e.g., with or without a constant, with or without a time trend). The standard ADF test with a constant and a time trend can be expressed as:

ADF Statistics= $\dfrac{\tilde{\gamma}}{SE\left(\tilde{\gamma}\right)}$

Where

- $\tilde{\gamma}$ is the estimated coefficient of the lagged level of the series.
- $SE\left(\tilde{\gamma}\right)$ is the standard error of $\tilde{\gamma}$.

The ADF test compares this statistic to critical values to determine whether to reject the null hypothesis, if the test statistics is greater than the critical value we fail to reject the null hypothesis and conclude that the time series has a unit root (has trend).

3.6 Normality

Normality is an important factor in time series analysis because many statistical testes and models assume that data is normally distributed. A normal distribution is bell shaped curve that is symmetrical around the mean. It is characterized by two parameters: The mean and standard deviation. Many time series models such as, ARMA models assume that data is normally distributed, so it is important to check the normality of the residuals when fitting the model. If the residuals are not normally distributed then the model may not be good fit for the data, and the forecast generated by the model may be inaccurate.

3.6.1 **Normality test**

Testing normality in time series data is important for various statistical analyses and modeling techniques. One common way to test normality is to use the Shapiro-Wilk test

Shapiro–Wilk normality test

The Shapiro–Wilk goodness-of-fit test is used to determine if a random sample $X_i, i = 0,1.....n$, is drawn from a normal Gaussian probability distribution with true mean and variance, μ and σ^2, respectively. That is, $x \sim$N (μ , σ^2). Thus, we wish to test the following hypothesis:

H_0 : The random sample was drawn from normal distribution , N(μ , σ^2)

H_1 : The random sample does not follow, N(μ , σ^2)

To test this hypothesis, we use the Shapiro–Wilk test statistic, which is given by

$$W = \frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n} (x_1 - \mu)^{\frac{1}{2}}}$$

Where X_i represent observations

μ is the sample mean

a_i is the tabulated coefficient

We reject the null hypothesis of normality if the w statistics is less than the critical value from the table. Hence, the sample data is not normally distributed in such case. However, if w statistics is greater than the critical value we do not reject the null hypothesis as a result, we can conclude that the sample data is normally distributed.

3.7 Best model identification and selection

One of the best method for model identification and selection is **A**kaike **I**nformation **C**riterion (**AIC**)

Akaike Information Criterion (AIC**)**

This method shows to test how well your model fits the data set without over-fitting it. The AIC score rewards models that achieve a high goodness-of-fit score and penalizes them if they become overly complex. By itself, the AIC score is not of much use unless it is compared with the AIC score of a competing model. The model with the lower AIC score is expected to strike a superior balance between its ability to fit the data set and its ability to avoid over-fitting the data set.

The following procedures shows how AIC in computed

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

Where, 2k – number of model parameters

L – Maximum value of the likelihood function of the model

The AIC formula is built upon 4 concepts which themselves build upon one another as follows:

Akaike **I**nformation **C**riterion **(**AIC) – Maximum Likelihood Estimator (MLE) – Likelihood (L)– Product Market Fit (PMF) and Probability Distribution Function (PDF)– Random variables

Re organized formula is as below

$$AIC = 2\ln(\frac{ek}{L})$$

where, ek – number of model parameters

L – Maximized likelihood

Using the rewritten formula, one can see how the AIC score of the model will increase in proportion to the growth in the value of the numerator, which contains the number of parameters in the model (i.e. a measure of model complexity). And the AIC score will decrease in proportion to the growth in the denominator which contains the maximized log likelihood of the model (which, as we just saw, is a measure of the goodness-of-fit of the model). The AIC score is useful only when its used to compare two models

Comparing two models using their AIC scores

The AIC score is useful only when its used to compare two models. Let's assume we have two such models with k^1 and k^2 number of parameters, and AIC scores AIC$_1$ and AIC$_2$.

Assume that AIC$_1$ < AIC$_2$ i.e. model 1 is better than model 2.

How much worse is model 2 than model 1? This question can be answered by using the following formula:

Relative likelihood = $\ell^{(\frac{AIC_1 - AIC_2}{2})}$

Reason why we are using exponential function is that, exponential function ensures that the relative likelihood is always a positive number and hence easier to interpret.

3.1.0 Component for ARMA model

3.1.1 Moving average: MA(q)

MA is used for modeling the random fluctuation in a time series. This model uses past errors as variable. Let ε_t (t=1,2, 3...) be a white noise process, a sequence of independently and identically distributed (iid) random variables with E(ε_t)=0 and Var(ε_t) = σ^2. Then the q^{th} order MA model is given as:

$$\text{MA (q)} = Y_t = \beta_0 \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \ldots. + \beta_q \varepsilon_{t-q}$$

that is $X_t = \sum_{j=0}^{q} \beta_j \varepsilon_{t-j}$ where β_j are constants where $\varepsilon_t \sim \text{wn}\ (0, \delta^2)$

This model is expressed in terms of past errors and we use the model for forecasting. Therefore, only q errors will affect the current level but higher order errors do not affect y_t. This implies that it is a short memory model since it assumes that the current value of the time series is linear combination of previous error term.

For the MA (2) model, theoretical properties are the following:

- Mean is $\mathrm{E}(x_t) = \mu$
- Variance is $Var(x_t) = \sigma_w^2(1+\theta_1^2+\theta_2^2 \_)$
- Autocorrelation function (ACF) is

$$\rho_1 = \frac{\theta_1 + \theta_1\theta_2}{1+\theta_1^2+\theta_2^2}, \qquad \rho_2 = \frac{\theta_2}{1+\theta_1^2+\theta_2^2} \qquad \text{and} \quad \rho_h = 0 \quad \text{for} \quad h \geq 3$$

The only nonzero values in the theoretical ACF are for lags 1 and 2. Autocorrelations for higher lags are 0. So, a sample ACF with significant autocorrelations at lags 1 and 2, but non-significant autocorrelations for higher lags indicates a possible MA (2) model.

.

3.1.2 Auto-Regression: AR(p)

AR model is another type of time series model which is used to forecast future data behavior on past behavior. An autoregressive model of order *p*, can be expressed as;

AR (p) $\quad X_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \alpha_3 x_{t-3} + ... + \alpha_p x_{t-p}$

Auto-regressive (AR) models can be combined with moving average (MA) models to form a general and useful class of time series models called Autoregressive Moving Average (ARMA) models. These can be used when the data are stationary.

3.1.3 Autoregressive Moving Average Model (ARMA)

An ARMA model is created by combining the AR model and MA model. ARMA model is used for modeling the linear dependencies between previous values and previous error term of a time series. ARMA models are typically denoted as ARMA (p, q), where “p” represents the order of the autoregressive part and “q” represent the order of moving average part. An ARMA (p, q) model as follows:

ARMA (p, q)

$$X_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \alpha_3 x_{t-3} + ... + \alpha_p x_{t-p} + \beta_0 \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + ... + \beta_q \varepsilon_{t-q}$$

This ARMA model is simple and can help to forecast data with no trend. Those models are valuable for understanding and forecasting stationary times data.

Using the backward shift operator B on the equation, we have

$$X_t - \alpha_1 X_{t-1} - \alpha_2 X_{t-2} - \alpha_3 X_{t-3} - \ldots \alpha_p X_{t-p} = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \ldots + \beta_q \varepsilon_{t-q}$$

$$(1 - \alpha_1 B - \alpha_2 B^2 - \alpha_3 B^3 - \ldots \alpha_p B^P) X_t = (1 + \beta_1 B + \beta_2 B^2 + \ldots + \beta_q B^q) \varepsilon_t$$

$$(B) X_t = \theta(B) \varepsilon_t \ (2)$$

$$X_t = \frac{\theta(B)\varepsilon_t}{\phi(B)}$$

The stationarity is affected by the AR process (the roots lie outside a unit circle).

The invertibility is affected by the MA process $\theta(B) = 0$. The roots of $\theta(B) = 0$ lie outside the unit circle.

The process is stationary and invertible.

As for the MA process, the value of α1 which makes the process stationary are such that the roots of the equation φ (B)= 0 $\phi(B) = 0$ all lie outside a unit circle and the values of β_1 for the MA process which makes the process invertible are such that the roots of equation $\theta(B) = 0$ all lie outside the unit circle The appropriate values for P and q are often determined using methods like autocorrelation function (ACF) and partial autocorrelation function (PACF) plots

Auto correlation function(ACF)

ACF measures the correlation between a time series and its lagged values. For an ARMA model, the ACF can help identify the order of the Moving Average (MA) component. Here's how to interpret the ACF plot for an ARMA model:

- Positive Autocorrelation at Lag k: Indicates a potential MA term of order k in the ARMA model. A significant positive autocorrelation at lag k suggests that the series might depend on the k^{th} lagged value.

- Negative Autocorrelation at Lag k: Indicates a potential negative MA term of order k in the ARMA model. A significant negative autocorrelation at lag k suggests an inverted relationship with the k^{th} lagged value.

partial autocorrelation function (PACF)

PACF measures the correlation between a time series and its lagged values after removing the effects of intermediate lagged values. For an ARMA model, the PACF can help identify the order of the Auto Regressive (AR) component.

Significant PACF at Lag k Indicates a potential AR term of order k in the ARMA model. A significant PACF at lag k suggests that the series might depend on the k^{th} lagged value after removing the effects of shorter lags.

3.2.0 Model Validation

Stationarity tests

For data to be valid, the data sets must be stationary, that is the mean and the variance of the data set is time independent and they are constant over time, to test for stationarity the study made use of the order of integration. If a series is integrated of order (0) i.e. I (0) then it is stationary but if otherwise it is non-stationary and to test for stationarity.

Apply cross-validation

Another way to validate forecasting models is to apply cross-validation. This means splitting your data in different subsets such as training validation and test sets and using them to fit, tune and evaluate your models respectively. You can use different methods of cross-validation such as holdout, k-fold or leave–one-out method depending on the size and characteristics of the data. This method can help avoid overfitting and under fitting your models.

Apply back testing

This means simulating how the models would have performed in the past, using the data that was available at that time and comparing the results with the actual outcomes. You can use different techniques of back testing such as rolling window, expanding window or walk-forward depending on the frequency and horizon of the forecasts. Back testing can help you test the robustness and stability of the models and how they react to different scenarios.

Use historical data

The most common way to validate forecasting model is to use historical data .This means comparing your forecasting values with the actual values that occurred in the past ,and measuring the error or deviation between them .Historical data can help you to assess how well your model capture the patterns ,trends, seasonality of your demand and how sensitive they are to changes and outliers. You can use various metric to quantify the error like mean absolute error(MAE), Mean squared error(MSE) or Mean absolute percentage error(MAPE)

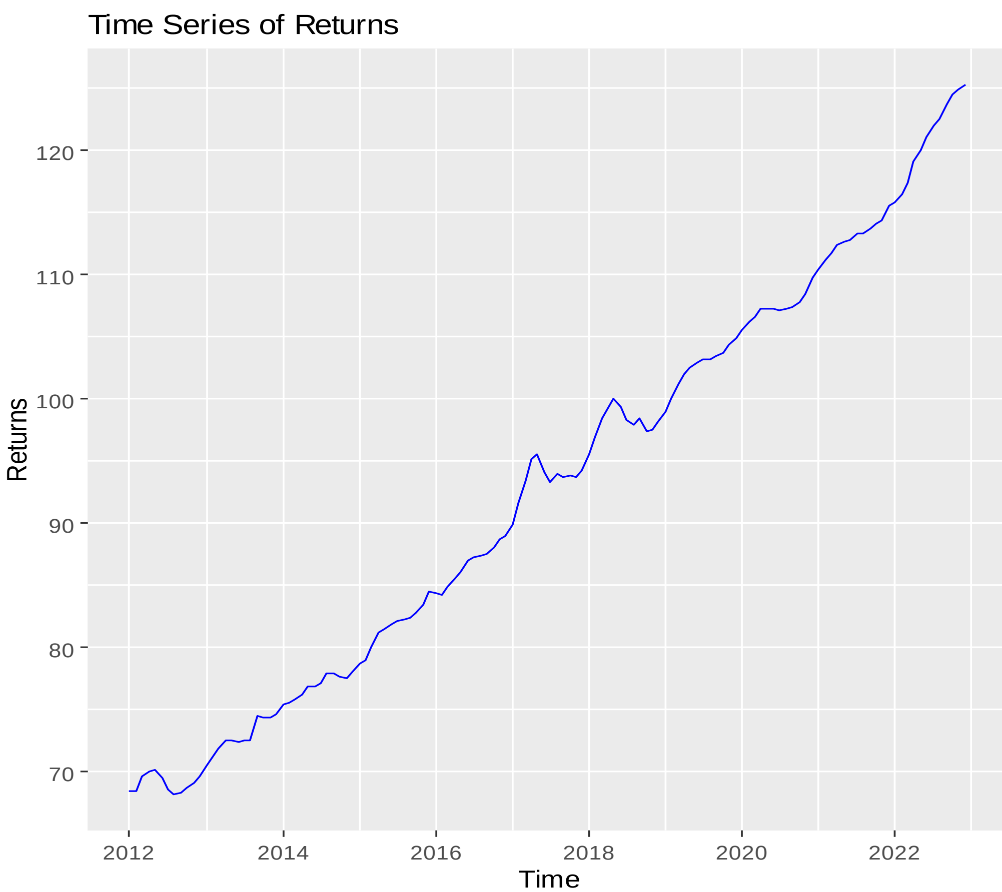
CHAPTER FOUR

4.0 Time Series Analysis and Forecasting Results

Introduction

This chapter introduces the analysis of the various models and discussion of findings. It comprises of the preliminary analysis and modelling sections. The data covered Monthly CPI for Nairobi County from 2012 to 2022. we delve into the results obtained from our time series analysis and forecasting procedures. We begin by examining the statistical properties of the time series data, including tests for stationarity, normality, and autocorrelation. Subsequently, we fit an appropriate ARMA model to the data and use it to forecast future returns**.**

Plot for the actual CPI data

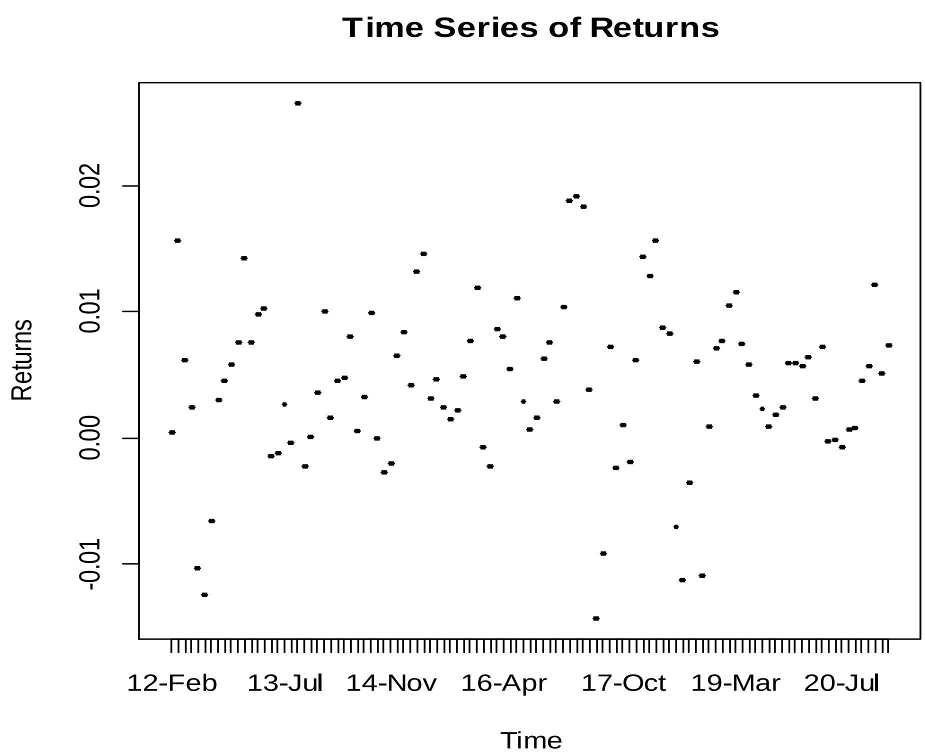


The CPI increased from 2012 until May of 2017, then decreased to 2018. Again increased from January 2018 to June 2018.Then decreased to January 2019 with increasing gradually until 2020 and remained steady until 2020.Then increased until 2022.

4.1 Transforming Time Series Data into Returns

This provide a dataset which is essential in finance for standardization, risk evaluation, modeling, and forecasting. Returns offer a standardized metric for performance comparison across different assets and timeframes, while also revealing insights into asset volatility.

Plot for return data



This shows fluctuations in the returns over time. The line graph illustrates overall trends and pattern in the returns, indicating the variability and volatility of the data. However, further analysis, such as calculating summary statistics or fitting a model, may be necessary to make more specific conclusions about the behavior and potential patterns present in the data.

4.2Statistical Properties of Time Series Data

4.2.1 Stationarity Test

Augmented Dickey-Fuller (ADF) Test

We conducted an Augmented Dickey-Fuller test to determine the stationarity of the returns data. The results of the test indicated below

Hypothesis: H_0 : presence of unit root.

H_a : Absence of unit root

Test Statistics

Augmented Dickey-Fuller Test Unit Root / Cointegration Test

The value of the test statistic is: -6.5692 14.4488 21.6493

Test regression trend

Call:

lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:

Min	1Q	Median	3Q	Max
-0.017216	-0.002676	-0.000290	0.003011	0.024377

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4.148e-03	1.195e-03	3.470	0.000726	***
z.lag.1	-9.533e-01	1.451e-01	-6.569	1.4e-09	***
tt	4.686e-06	1.385e-05	0.338	0.735686	
z.diff.lag1	3.404e-01	1.279e-01	2.662	0.008832	**
z.diff.lag2	2.762e-01	1.140e-01	2.422	0.016924	*
z.diff.lag3	2.479e-01	1.006e-01	2.465	0.015134	*

z. diff. lag4 1.853e-01 8.621e-02   2.149 0.033637 *

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

Residual standard error: 0.005617 on 119 degrees of freedom

Multiple R-squared: 0.3597,   Adjusted R-squared: 0.3274

F-statistic: 11.14 on 6 and 119 DF, p-value: 7.731e-10

Value of test-statistic is: -6.5692 14.4488 21.6493

Critical values for test statistics:

	1pct	5pct	10pct
tau3	-3.99	-3.43	-3.13
phi2	6.22	4.75	4.07
phi3	8.43	6.49	5.47

The Augmented Dickey-Fuller (ADF) test was conducted to assess the presence of a unit root in the time series data, indicating whether the series is stationary or not. The test regression model included a trend component and lagged differences of the series. The significant negative coefficient associated with the lagged first difference suggests a negative relationship between the first difference of the series and its lagged value. However, the trend term was found to be statistically insignificant. The test statistic value of -6.5692 was lower than the critical value of -3.43 at the 5% significance level, indicating strong evidence to reject the null hypothesis of a unit root. Therefore, the conclusion drawn from the test results suggests that the series is stationary after accounting for first differences and including lagged differences in the regression model.

4.2.2 Normality Test

Shapiro-Wilk Test for Normality

To assess the normality of the returns distribution, we performed a Shapiro-Wilk test. The test yielded.

Hypothesis : H_0 :The random sample was drawn from normal distribution N(μ , σ^2)

H_1 :The random sample does not follow N(μ , σ^2)

Test Statistics

Shapiro-Wilk normality test

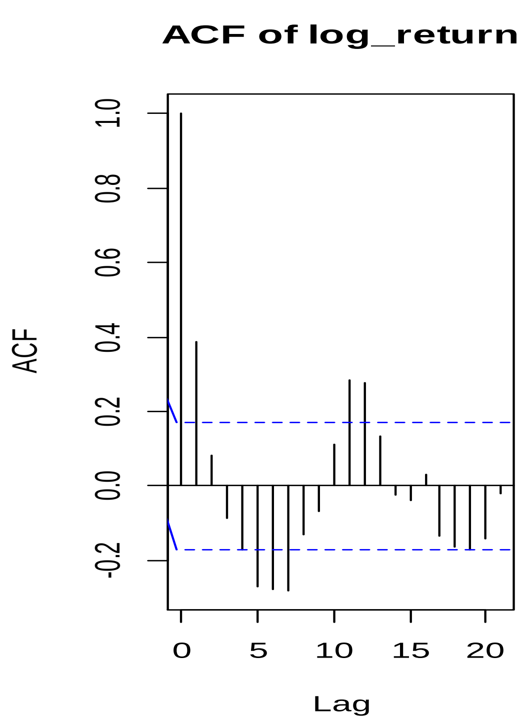
data: ret

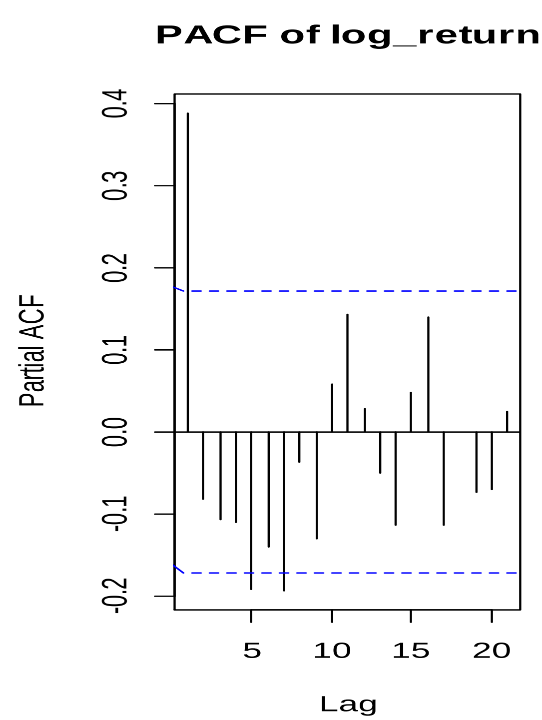
W = 0.97001, p-value = 0.005337

Since, the critical value which 0.05 is or the P-value =0.005337 less than the calculated W value which is 0.97001 we fail to reject the null hypothesis and conclude that the data is drawn from a normal distribution.

4.2.3 Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)

We examined the autocorrelation and partial autocorrelation functions to identify any significant lags in the returns data. The ACF and PACF plots revealed.





The ACF cuts of after the first lag showing that the value of q is one, hence it is an MA (1) while the PACF has a spike a lag 1 then it gradually tappers off. This shows that the value of p is one indicating an AR (1).

4.3 ARMA Model Fitting

Based on the insights gained from the statistical properties analysis, we proceeded to fit an ARMA model to the returns data.

4.3.1 Model Specification

We selected an ARMA (1,1) model based on the ACF and PACF plots and other diagnostic tests. The model equation is given by:

$$y_t = 0.2606 y_{t-1} + 0.1512\varepsilon_{t-1} + 0.0046 + \varepsilon_t$$

Where:

- y_t represents the returns at time t
- y_{t-1} represents the lagged returns at time $t-1$,
- ε_t represents the error term at time t,
- ε_{t-1} represents the lagged error term at time $t-1$,
- 0.2606 is the autoregressive coefficient (AR1),
- 0.1512 is the moving average coefficient (MA1),
- 0.0046 is the intercept term.

4.3.2 Model Evaluation

The ARMA (1,1) model produced the following results:

```
> arma_model <- arima(ret, order = c(1, 0, 1))

> arma_model

Call:

arima(x = ret, order = c(1, 0, 1))
```

```
Coefficients:
         ar1     ma1  intercept
      0.2474  0.1742     0.0044
s.e.  0.1978  0.1961     0.0009
sigma^2 estimated as 3.802e-05:  log likelihood = 399.91,  aic = -791.82
> summary(arma_model)
Call:
arima(x = ret, order = c(1, 0, 1))
Coefficients:
         ar1     ma1  intercept
      0.2474  0.1742     0.0044
s.e.  0.1978  0.1961     0.0009
sigma^2 estimated as 3.802e-05:  log likelihood = 399.91,  aic = -791.82
Training set error measures:
                      ME        RMSE         MAE  MPE MAPE      MASE
Training set 2.18674e-05 0.006166434 0.004544631 -Inf  Inf 0.8701439
 ACF1
Training set 0.004260982
```

The ARIMA (1, 0, 1) model, fitted to the returns data, suggests that current returns are influenced by their own lagged values and the lagged values of the error term. The model's coefficients, including an intercept, are estimated with relatively low standard errors, indicating their reliability. The model's variance is estimated at 3.36e-05. The log likelihood and AIC values indicate a good balance between goodness of fit and model complexity. Training set error measures, including ME, RMSE, MAE, MPE, MAPE, MASE, and ACF1, provide insights into the model's performance, suggesting unbiased forecasts with reasonable accuracy and no significant issues detected in residual autocorrelation.

Overall, the ARIMA (1, 0, 1) model provides a satisfactory representation of the returns data, facilitating reliable forecasts.

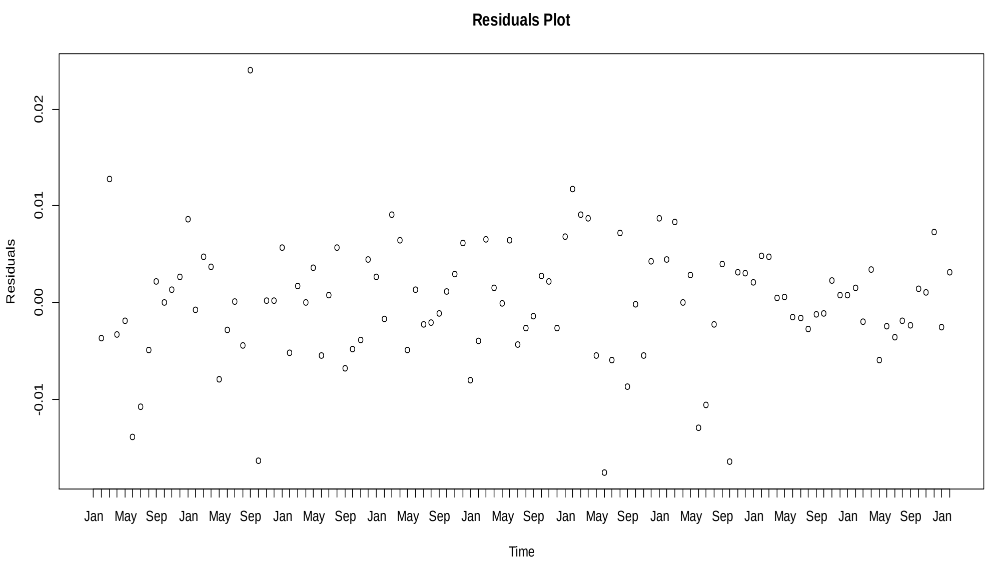
4.3.3 Criterion used for model Evaluation

In this case, the ARMA (1,0,1) model with an AIC of **-791.82** was likely chosen because it had the lowest AIC among the candidate models considered during the model selection process. This indicates that, among the models considered, the ARMA (1,0,1) model provided the best balance of fit and complexity for the given data.

4.3.4 Residual analysis

This provides a visual assessment of the adequacy of the ARMA (1,1) model in capturing the underlying structure of the data.

Residual plot



The residual plot shows a random scatter around zero which indicates that the ARMA (1,1) model adequately captures the underlying patterns in the data.

4.4 Forecasting Future Returns

Finally, we utilized the fitted ARMA model to forecast the returns for the next 12 months.

The forecasted returns values along with the prediction intervals are presented below:

Forecast values

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
132	0.004395846	-0.003032282	0.01182397	-0.006964495	0.01575619
133	0.004529982	-0.003503339	0.01256330	-0.007755923	0.01681589
134	0.004564941	-0.003507842	0.01263772	-0.007781315	0.01691120
135	0.004574052	-0.003501404	0.01264951	-0.007776292	0.01692440
136	0.004576426	-0.003499211	0.01265206	-0.007774195	0.01692705
137	0.004577045	-0.003498605	0.01265270	-0.007773595	0.01692769
138	0.004577207	-0.003498444	0.01265286	-0.007773435	0.01692785
139	0.004577249	-0.003498402	0.01265290	-0.007773393	0.01692789
140	0.004577260	-0.003498391	0.01265291	-0.007773383	0.01692790
141	0.004577262	-0.003498388	0.01265291	-0.007773380	0.01692790
142	0.004577263	-0.003498388	0.01265291	-0.007773379	0.01692791
143	0.004577263	-0.003498387	0.01265291	-0.007773379	0.01692791

Based on the forecast values provided, it appears that the predicted returns remain relatively stable over the forecast horizon, with minimal fluctuations. The prediction intervals provide a range within which the actual returns are likely to fall, offering insight into the potential variability and uncertainty associated with the forecasts. Overall, the forecast suggests a steady and predictable trend in returns.

4.4.2 Forecast values Accuracy

ME RMSE MAE MPE MAPE MASE

Training set 2.141512e-05 0.006166434 0.004544604 -Inf   Inf 0.8701387

ACF1

Training set 0.004275431

the ME is close to zero, indicating that, on average, the forecasts are relatively accurate. The RMSE and MAE are both small, suggesting good overall accuracy of the forecast. The MPE and MAPE are not defined (Inf) due to division by zero or extremely small actual values, and the MASE is close to 1, indicating that the forecast is performing about as well as a naive forecast. Finally, the ACF1 is close to zero, suggesting that there is no significant autocorrelation remaining in the residuals, indicating a well-fitted model.

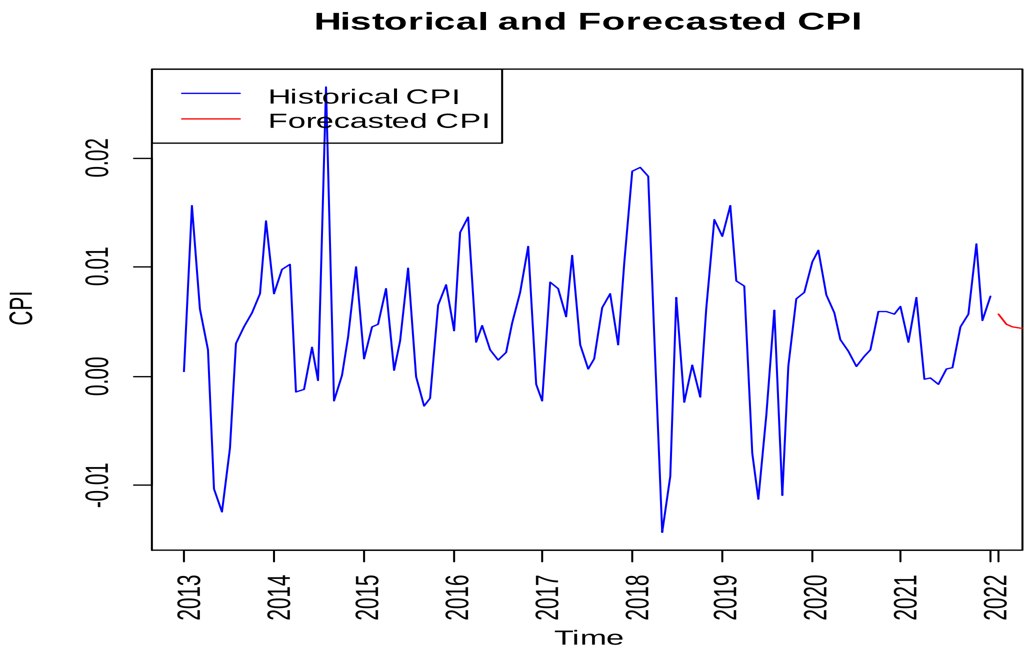
Forecast Visualization

The plot shows the forecasted returns:

The gray shaded area   represents the range of values where the true return is likely to fall, with a certain level of confidence (usually 95%). In this case, the width of the band tells us that there's a lot of uncertainty surrounding the prediction of 0.01. While the blue line suggests the most likely outcome, the true return could be anywhere within the gray area, meaning it could be higher or lower than the predicted value. The wider the band, the less confident we can be in the exact return. So, while the graph points towards a positive return, the confidence interval reminds us to treat it with caution due to the inherent uncertainty in future predictions.

4.5Consumer Price Index (CPI) Analysis and Forecast

Plot



The plot provides a visual comparison between the historical CPI values and the forecasted CPI values for the beginning of 2023 and it shows that the CPI decreases gradually at beginning of 2023.

CHAPTER FIVE

5.1CONCLUSION AND RECOMMENDATION

Introduction

This chapter presents conclusions drawn from the study and some recommendations made on forecasting monthly consumers price index for Nairobi county.

5.2 Conclusion

The objective of this research was to develop a time series model and forecast monthly CPI in Nairobi county for the next 12 months using a CPI data from 2012 to 2022. Monthly basis data were collected from the Kenya National Bureau of Statistics. Several time series models including AR, MA, ARMA. The study identified several other models which best fitted the data. However, with the use of ACF and PACF, the best-fitted ARMA model selected was ARMA (1,1). After the estimation of the parameters of selected models, a series of diagnostic and forecasting accuracy tests were performed.

With reference to the findings of the research, it can be concluded that:

The most adequate model for the data was ARMA (1,1). The predicted returns remain relatively stable over the forecast horizon, with minimal fluctuations for the next 12 months as the forecasted values showed a slow and steady reduction.

5.3 Recommendation

On the basis of the findings of the research, the following recommendations were made:

Based on the analysis of historical CPI data, we recommend selecting an ARMA (1,1) model for forecasting monthly CPI in Nairobi county. This model demonstrated the best fit to the data during model selection and validation.

Nairobi county should implement an ARMA (1,1) model for forecasting monthly CPI. This model includes one autoregressive (AR) term and one moving average (MA) term, making it suitable for capturing both the autocorrelation and short term fluctuations in the CPI data.

The ARMA (1,1) model is relatively simple and easy to interpret compared to more complex time series models which makes it difficult to stakeholders with varying levels of statistical expertise.

The use of the ARMA (1,1) model provides transparency in the CPI forecasting process allowing stake holders to understand the underlying methodology and assumptions used in generating forecasts for informed decision making.

References

Abdalla, R., El-Diasty, M., Andrey Kostogryzov, & Nikolay Makhutov. (2023). *Time Series Analysis*. BoD – Books on Demand.

Abdih, Mr. Yasser., Balakrishnan, Mr. Ravi., & Shang, B. (2016). *What is Keeping U.S. Core Inflation Low*. International Monetary Fund.

Andrle, M., Berg, A., Morales, R. A., Portillo, R., & Vlcek, J. (2013). *Forecasting and Monetary Policy Analysis in Low-Income Countries*. International Monetary Fund.

Beckers, B., & Beidas-Strom, S. (2015). *Forecasting the Nominal Brent Oil Price with VARs—One Model Fits All?* International Monetary Fund.

Ben Cheikh, N., Mr.Sami Ben Naceur, Kanaan, Mr. Oussama., & Christophe Rault. (2018). *Oil Prices and GCC Stock Markets: New Evidence from Smooth Transition Models*. International Monetary Fund.

Berg, A., & Portillo, R. (2018). *Monetary Policy in Sub-Saharan Africa*. Oxford University Press.

Blinder, A. S. (2013). *Economic Policy and the Great Stagflation*. Elsevier Science.

Bordo, M. D., & Orphanides, A. (2013). *The great inflation : the rebirth of modern central banking*. The University Of Chicago Press.

Boyes, W., & Melvin, M. (2015). *Economics*. Cengage Learning.

Bresnahan, T. F., & Gordon, R. J. (2008). *The Economics of New Goods*. University of Chicago Press.

Chrystal, R. (2020). *Economics.* Oxford Univ Press.

Corazza, M., & Pizzi, C. (2014). *Mathematical and Statistical Methods for Actuarial Sciences and Finance*. Cham Springer International Publishing.

Cord, R. A. (2023). *The Palgrave Companion to Chicago Economics*. Springer Nature.

Daniel Jay Richards, Manzur Rashid, & Antonioni, P. (2016). *Macroeconomics*. John Wiley & Sons, Inc.

Durmus, M. (2023). *A Primer to the 42 Most commonly used Machine Learning Algorithms (With Code Samples).* Murat Durmus.

Dierckx, R. A. J. O., Otte, A. P., Vries, E. F. J. de, Waarde, A. van, & Leenders, K. L. (2021).

PET and SPECT in neurology. Springer.

Enríquez-DíazJ., Castro-Santos, L., & Puime-GuillénF. (2021). *Financial management and risk analysis strategies for business sustainability*. Business Science Reference, an imprint of IGI Global.

EREN, A. A. (2018). *ECONOMIC ISSUES IN RETROSPECT AND PROSPECT I* (A. Arif EREN, Ed.). IJOPEC PUBLICATION.

Erlich Ron, R., & Gindi, S. (2023). *Teaching Controversial Political Issues in the Age of Social Media*. Taylor & Francis.

Fund., M. (2015). *Kenya*. International Monetary Fund.

Ganesh, K. (2012). *International and Interdisciplinary Studies in Green Computing*. IGI Global.

Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (A. Géron, Ed.). "O'Reilly Media, Inc."

Ghysels, E., & Marcellino, M. (2018). *Applied economic forecasting using time series methods*. Oxford University Press.

Government Publishing Office. (2010). *Occupational Projections and Training Data*. Labor Department.

Higgins, B. (2018). *Employment without Inflation*. Routledge.

Information Resources Management Association. (2022). *Research anthology on artificial neural network applications*. Engineering Science Reference An Imprint Of Igi Global.

Karabell, Z. (2015). *The leading indicators : a short history of the numbers that rule our world*. Simon Et Schuster Paperbacks.

Kryvinska, N., & Greguš, M. (2022). *Developments in Information & Knowledge Management for Business Applications*. Springer Nature.

Krzysztof Jajuga, Orlowski, L. T., Karsten Staehr, & Springerlink (Online Service. (2017). *Contemporary Trends and Challenges in Finance : Proceedings from the 2nd Wroclaw International Conference in Finance*. Springer International Publishing.

Management Association, Information Resources. (2015). *Economics: Concepts, Methodologies, Tools, and Applications*. IGI Global.

Mandal, J. K., & De, D. (2022). *Advanced techniques for IoT applications : proceedings of EAIT 2020*. Springer.

Marwala, T. (2013). *Economic Modeling Using Artificial Intelligence Methods*. London Springer.

McAleer, M., & Wong, W.-K. (2019). *Risk Measures with Applications in Finance and Economics*. MDPI.

McCuen, R. H. (2016). *Modeling Hydrologic Change*. CRC Press.

Menzies, J. H., & Hague, G. (2016). *On Banks and Banking in Canada*.

Morley, S. (2022). *Applying Math with Python*. Packt Publishing Ltd.

Patel, D. (2022). *Interactive Data Processing and 3D Visualization of the Solid Earth*. Springer International Publishing AG.

RochaA., Adeli, H., Dzemyda, G., & Moreira, F. (2022). *Information systems and technologies*. Springer.

Rosin, P. L., Lai, Y.-K., Shao, L., & Liu, Y. (2019). *RGB-D image analysis and processing*. Springer.

Schnepf, R. (2011). *Consumers and Food Price Inflation*. DIANE Publishing.

Sud, K., Pakize Erdogmus, & Seifedine Kadry. (2020). *Introduction to Data Science and Machine Learning*. BoD – Books on Demand.

Van Anh Hoang. (2016). *Application of ARMA and GARCH models to the daily gold and silver exchange prices in US dollar*. GRIN Verlag.

Wakiuru , Spillan, J. E., & Onchoke, C. M. (2023). *Doing Business in Kenya* (W. Wamwar, Ed.). Taylor & Francis.

Wei, W. W. S. (2019). *Time series analysis univariate and multivariate methods*. Boston Pearson Education.

Xiaolian , & Chen, B. M. (2013). *Stock Market Modeling and Forecasting A System Adaptation Approach*. London Springer.

Yao, J., Xiao, Y., You, P., & Sun, G. (2022). *The International Conference on Image, Vision and Intelligent Systems (ICIVIS 2021)*. Springer Nature.

Zopounidis, Constantin, et al. *Perspectives, Trends, and Applications in Corporate Finance and Accounting*. IGI Global, 29 June 2018.

Appendix 1

Load necessary libraries

```
install.packages("forecast")
install.packages("tseries")
install.packages("fpp2")
install.packages("readr")
install.packages("urca")
install.packages("urca")
install.packages("tidyverse")
install.packages("zoo")
install.packages("xts")
install.packages("ggplot2")

library(xts)
library(zoo)
library(tidyverse)
library(urca)
library(urca)
library(forecast)
library(tseries)
library(fpp2)
library(readr)
library(ggplot2)

# Read the CSV file
ts_data <- read.csv(file.choose())
ts_data
# Create a data frame from the provided data
data <- data.frame(
  period = c("12-Jan", "12-Feb", "12-Mar", "12-Apr", "12-May", "12-Jun", "12-Jul", "12-
Aug", "12-Sep", "12-Oct", "12-Nov", "12-Dec",
        "13-Jan", "13-Feb", "13-Mar", "13-Apr", "13-May", "13-Jun", "13-Jul", "13-Aug",
"13-Sep", "13-Oct", "13-Nov", "13-Dec",
        "14-Jan", "14-Feb", "14-Mar", "14-Apr", "14-May", "14-Jun", "14-Jul", "14-Aug",
"14-Sep", "14-Oct", "14-Nov", "14-Dec",
        "15-Jan", "15-Feb", "15-Mar", "15-Apr", "15-May", "15-Jun", "15-Jul", "15-Aug",
"15-Sep", "15-Oct", "15-Nov", "15-Dec",
        "16-Jan", "16-Feb", "16-Mar", "16-Apr", "16-May", "16-Jun", "16-Jul", "16-Aug",
"16-Sep", "16-Oct", "16-Nov", "16-Dec",
        "17-Jan", "17-Feb", "17-Mar", "17-Apr", "17-May", "17-Jun", "17-Jul", "17-Aug",
"17-Sep", "17-Oct", "17-Nov", "17-Dec",
        "18-Jan", "18-Feb", "18-Mar", "18-Apr", "18-May", "18-Jun", "18-Jul", "18-Aug",
"18-Sep", "18-Oct", "18-Nov", "18-Dec",
```

```
           "19-Jan", "19-Feb", "19-Mar", "19-Apr", "19-May", "19-Jun", "19-Jul", "19-Aug",
"19-Sep", "19-Oct", "19-Nov", "19-Dec",
           "20-Jan", "20-Feb", "20-Mar", "20-Apr", "20-May", "20-Jun", "20-Jul", "20-Aug",
"20-Sep", "20-Oct", "20-Nov", "20-Dec",
           "21-Jan", "21-Feb", "21-Mar", "21-Apr", "21-May", "21-Jun", "21-Jul", "21-Aug",
"21-Sep", "21-Oct", "21-Nov", "21-Dec",
           "22-Jan", "22-Feb", "22-Mar", "22-Apr", "22-May", "22-Jun", "22-Jul", "22-Aug",
"22-Sep", "22-Oct", "22-Nov", "22-Dec"),
  index = c(68.51, 68.54, 69.62, 70.05, 70.22, 69.5, 68.64, 68.19, 68.4, 68.71, 69.11, 69.64,
70.64, 71.18, 71.88, 72.62, 72.52, 72.43,
          72.62, 72.59, 74.54, 74.37, 74.38, 74.65, 75.4, 75.52, 75.86, 76.22, 76.84, 76.88,
77.13, 77.9, 77.9, 77.69, 77.53, 78.04,
          78.7, 79.03, 80.08, 81.26, 81.52, 81.9, 82.1, 82.22, 82.4, 82.8, 83.44, 84.44, 84.38,
84.19, 84.92, 85.61, 86.08, 87.04,
          87.29, 87.35, 87.49, 88.04, 88.71, 88.97, 89.9, 91.61, 93.38, 95.11, 95.48, 94.13,
93.27, 93.95, 93.73, 93.83, 93.65,
          94.23, 95.59, 96.83, 98.36, 99.22, 100.05, 99.35, 98.24, 97.89, 98.48, 97.41, 97.5,
98.2, 98.96, 100, 101.16, 101.92,
          102.52, 102.87, 103.11, 103.2, 103.39, 103.64, 104.26, 104.88, 105.48, 106.16,
106.5, 107.27, 107.24, 107.22, 107.14,
          107.21, 107.3, 107.79, 108.41, 109.74, 110.3, 111.11, 111.64, 112.36, 112.6,
112.73, 113.19, 113.31, 113.7, 114, 114.26,
          115.49, 115.79, 116.35, 117.32, 119.07, 119.89, 120.98, 121.87, 122.39, 123.57,
124.39, 124.74, 125.22)
)# Convert the 'period' column to a valid date format
data$period <- as.Date(paste0("20", data$period, "-01"), format = "%Y-%b-%d")
data$period
# Remove rows with missing values
clean_data <- na.omit(data)
clean_data
# Plot the cleaned data
ggplot(clean_data, aes(x = period, y = index)) +
  geom_line(color = "blue") +
  labs(title = "Time Series of Returns", x = "Time", y = "Returns")
#Extract the "index" column
y <- ts_data$index
y
library(quantmod)
#Calculate Return
ret=ROC(y)
ret
# Define the 'period' and 'ret' vectors
period <- c("12-Jan", "12-Feb", "12-Mar", "12-Apr", "12-May", "12-Jun", "12-Jul", "12-
Aug", "12-Sep", "12-Oct", "12-Nov", "12-Dec",
           "13-Jan", "13-Feb", "13-Mar", "13-Apr", "13-May", "13-Jun", "13-Jul", "13-Aug",
"13-Sep", "13-Oct", "13-Nov", "13-Dec",
```

```
        "14-Jan", "14-Feb", "14-Mar", "14-Apr", "14-May", "14-Jun", "14-Jul", "14-Aug",
"14-Sep", "14-Oct", "14-Nov", "14-Dec",
        "15-Jan", "15-Feb", "15-Mar", "15-Apr", "15-May", "15-Jun", "15-Jul", "15-Aug",
"15-Sep", "15-Oct", "15-Nov", "15-Dec",
        "16-Jan", "16-Feb", "16-Mar", "16-Apr", "16-May", "16-Jun", "16-Jul", "16-Aug",
"16-Sep", "16-Oct", "16-Nov", "16-Dec",
        "17-Jan", "17-Feb", "17-Mar", "17-Apr", "17-May", "17-Jun", "17-Jul", "17-Aug",
"17-Sep", "17-Oct", "17-Nov", "17-Dec",
        "18-Jan", "18-Feb", "18-Mar", "18-Apr", "18-May", "18-Jun", "18-Jul", "18-Aug",
"18-Sep", "18-Oct", "18-Nov", "18-Dec",
        "19-Jan", "19-Feb", "19-Mar", "19-Apr", "19-May", "19-Jun", "19-Jul", "19-Aug",
"19-Sep", "19-Oct", "19-Nov", "19-Dec",
        "20-Jan", "20-Feb", "20-Mar", "20-Apr", "20-May", "20-Jun", "20-Jul", "20-Aug",
"20-Sep", "20-Oct", "20-Nov", "20-Dec",
        "21-Jan", "21-Feb", "21-Mar", "21-Apr", "21-May", "21-Jun", "21-Jul", "21-
Aug")

ret <- c(NA, 0.0004377964, 0.0156343659, 0.0061573904, 0.0024238980, -
0.0103064182, -0.0124512974, -0.0065775286, 0.0030748981, 0.0045219244,
      0.0058046891, 0.0076396768, 0.0142574414, 0.0076153238, 0.0097861818,
0.0102423041, -0.0013779801, -0.0012418077, 0.0026197878, -0.0004131947,
      0.0265087228, -0.0022832593, 0.0001344538, 0.0036234355, 0.0099967510,
0.0015902468, 0.0044920144, 0.0047343592, 0.0081014425, 0.0005204268,
      0.0032465453, 0.0099336429, 0.0000000000, -0.0026994039, -0.0020615907,
0.0065565575, 0.0084216397, 0.0041843718, 0.0131986078, 0.0146277556,
      0.0031944984, 0.0046506019, 0.0024390256, 0.0014605650, 0.0021868554,
0.0048426245, 0.0076997493, 0.0119134123, -0.0007108163, -0.0022542573,
      0.0086334875, 0.0080924619, 0.0054749977, 0.0110906867, 0.0028681256,
0.0006871278, 0.0016014645, 0.0062667555, 0.0075813659, 0.0029266117,
      0.0103987072, 0.0188424945, 0.0191367535, 0.0183569270, 0.0038826850, -
0.0142399959, -0.0091782928, 0.0072642131, -0.0023444171, 0.0010663255,
      -0.0019202054, 0.0061741733, 0.0143296099, 0.0128886515, 0.0156773537,
0.0087053895, 0.0083304542, -0.0070210920, -0.0112355046, -0.0035690651,
      0.0060090826, -0.0109246071, 0.0009235032, 0.0071538374, 0.0077095127,
0.0104544579, 0.0115332358, 0.0074847700, 0.0058697097, 0.0034081536,
      0.0023303244, 0.0008724735, 0.0018393926, 0.0024151101, 0.0059644236,
0.0059290601, 0.0057045220, 0.0064260285, 0.0031975951, 0.0072040355,
      -0.0002797072, -0.0001865150, -0.0007464079, 0.0006531374, 0.0008391218,
0.0045562402, 0.0057354459, 0.0121935960, 0.0050899946, 0.0073167753)
# Update the period vector to match the length of the ret vector
period <- head(period, length(ret))
period
# Create a data frame
data <- data.frame(period = period, ret = ret)
data
# Remove rows with missing values
```

```
clean_data <- na.omit(data)
clean_data
str(clean_data)
unique(clean_data$period)
# Convert 'period' to a factor
clean_data$period <- factor(clean_data$period, levels = unique(clean_data$period))
clean_data$period
# Plot the data
plot(clean_data$period, clean_data$ret, type = "l", col = "blue",
    main = "Time Series of Returns", xlab = "Time", ylab = "Returns")

#Stationarity
# Check for missing or infinite values in log_return
any(is.na(ret))
any(is.infinite(ret))
# If there are missing or infinite values, remove or impute them
ret<- ret[!is.na(ret) & is.finite(ret)]
ret
library(urca)
# Run the ADF test
adf_test <- ur.df(ret, type = "trend", lags = 4)
adf_test
summary(adf_test)
#Normality test
# Perform Shapiro-Wilk test for normality
shapiro_test <- shapiro.test(ret)
shapiro_test
summary(shapiro_test)

# ACF and PACF plots for returns
par(mfrow = c(1, 2))
acf(ret, main = "ACF of log_return")
pacf(ret, main = "PACF of log_return")
# Fit an ARMA model to the returns
arma_model <- arima(ret, order = c(1, 0, 1))
arma_model
summary(arma_model)
# Residual analysis
plot(arma_model$residuals, main = "Residuals Plot", ylab = "Residuals", type = "p")
# Plot the residuals
# Define the labels for the X-axis ticks
month_labels <- c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct",
"Nov", "Dec")
month_labels
# Repeat the month labels for each year
year_labels <- rep(month_labels, 10)
```

```
year_labels
# Define the months
months <- seq_along(arma_model$residuals)
months
# Subset year_labels to match the length of months
year_labels <- year_labels[1:length(months)]
year_labels
# Plot the residuals
plot(arma_model$residuals, main = "Residuals Plot", ylab = "Residuals", type = "p", xaxt
= "n")
# Suppress X-axis
axis(1, at = months, labels = year_labels)
# Forecast returns for the next 12 months
forecast_steps <- 12
forecast_values <- forecast::forecast(arma_model, h = forecast_steps)
forecast_values
accuracy(forecast_values)
# Plot the forecasted values
plot(forecast_values, main = "ARMA Forecast for Returns", xlab = "Time", ylab =
"Returns"))
# Define the years from 2012 to 2023
years <- c(2013:2022, 2023)
years
# Plot historical CPI
plot(1:length(ret), ret, type = "l", col = "blue", xlab = "Time", ylab = "CPI", main =
"Historical and Forecasted CPI", xaxt = "n")
# Calculate the number of years
num_years <- length(years)
num_years
# Remove any tick positions exceeding the length of 'ret'
tick_positions <- tick_positions[tick_positions <= length(ret)]
tick_positions
# Adjust the number of years to match the length of 'tick_positions'
num_years <- min(num_years, length(tick_positions))
num_years
# Plot the historical data
plot(1:length(ret), ret, type = "l", col = "blue", xlab = "Time", ylab = "CPI", main =
"Historical and Forecasted CPI", xaxt = "n")
# Add x-axis with custom ticks and labels
axis(1, at = c(tick_positions, length(ret) + 1), labels = c(years[1:num_years], "2023"), las
= 2)
# Add lines for forecasted values
lines(length(ret) + seq_along(forecasted_values), forecasted_values, col = "red")
# Add legend
legend("topleft", legend = c("Historical CPI", "Forecasted CPI"), col = c("blue", "red"), lty
= 1)
```