ABSTRACTCOVID-19 has now taken a frightening form as the days pass, it is becoming more and more widespread and now it has become an epidemic. The case rate, which was earlier in the hundreds, changed to thousands and then progressing to million. This research focus on to forecast the future of covid-19 confirmed instances (cases) using Time series ARIMA. The research tool will assist you in discovering answers to some questions like.What are the prospects for new confirmed cases on Covid-19 in the future?, What effect does this have on healthcare delivery in the country?, Has Covid-19 confirmed cases had a link to high covid-19 deaths in the country?. This dataset has been collected precisely from DHIMS-Covid-19 Dataset from 14th March 2020 to 8th July 2021. The ARIMA model is used to generate ARIMA results, AC residual plots, PAC results and corresponding prediction plots on the dataset. By applying ARIMA model to forecast confirmed cases, the coefficients generated by the regression model are estimated, and the actual covid-19 confirmed cases and expected confirmed cases are compared and analyzed. It is found that the predicted case has a consistent decrease after 19-OCT-21 to 9-JAN-22. The study also found that, the country would experience an additional case of **66,382** with in the forecasted period adding to the existing cases of **96,971** as at July 8th 2021. This will help the government and doctors prepare for the forthcoming of the virus. Based on the period predictions, The Center for Disease Control should adopt the ARIMA (6, 1, 8) model in their disease control department (DCD) when planning activities to combat the spread of covid-19. In terms of regions and districts, vaccination should be prioritized in covid-19-prone areas. In order to adequately prepare for the country's continued overwhelming covid-19 cases, the government/authorities should rely on anticipated data (forecasted figures) in their planning operations.

**Table of Contents**

[DECLARATION ii](#_Toc85890678)

[DEDICATION iii](#_Toc85890679)

[ACKNOWLEDGEMENTS iv](#_Toc85890680)

[ABSTRACT v](#_Toc85890681)

TABLE OF CONTENT……………………………..………………………….…………………………………..…………vi & vii

LIST OF TABLES………………………………..…………………………………………………………………………….…….viii

LIST OF FIGURES….…………..………….…………..…………………………………………………………………………....ix

LIST OF ABBREVIATIONS.……….……………..…………………………………………………………………………….…. x

[**CHAPTER ONE: INTRODUCTION** 1](#_Toc85890682)

[1.1 Background 1](#_Toc85890683)

[1.2 Statement of the problem 2](#_Toc85890684)

[1.3 Objectives 4](#_Toc85890685)

[1.3.1 Main objectives: 4](#_Toc85890686)

[1.3.2 Specific objectives: 4](#_Toc85890687)

[1.4. Research Questions. 4](#_Toc85890688)

[1.5 Hypothesis of the study 5](#_Toc85890689)

[1.6. Significant of the study, 5](#_Toc85890690)

[1.7. Study area and scope. 6](#_Toc85890691)

[1.8. Limitation of the study. 6](#_Toc85890692)

[1.9 Outline/Organization of the study. 7](#_Toc85890693)

[**CHAPTER TWO:** **LITERATURE REVIEW** 8](#_Toc85890694)

[2.0 Introduction: 8](#_Toc85890696)

[2.1 Topics related to the research problem. 8](#_Toc85890697)

[2.2 Work of other researchers 9](#_Toc85890698)

[2.2 Theoretical literature 12](#_Toc85890699)

[2.3 Empirical literature 13](#_Toc85890700)

[2.4 Conclusion 14](#_Toc85890701)

[**CHAPTER THREE:** **METHODOLOGY** 15](#_Toc85890702)

[3.0 introduction 15](#_Toc85890704)

[3.1 Research Design 15](#_Toc85890705)

[Population of Study 15](#_Toc85890706)

[Sample Size 16](#_Toc85890707)

[3.2 Data Collection 16](#_Toc85890708)

[3.2.1 Reliability and Validity of Data 16](#_Toc85890709)

[3.3 Data Analysis 16](#_Toc85890710)

[3.3.1 Time Series Analysis 17](#_Toc85890711)

[**CHAPTER FOUR:** **DATA ANALYSIS AND RESULTS** 32](#_Toc85890712)

[4.0 Introduction 32](#_Toc85890714)

[ARIMA and Box-Jenkins model selections 33](#_Toc85890715)

[Estimation of ARIMA model 44](#_Toc85890728)

[Diagnostic and Forecasting our ARIMA model (6, 1, 8) 45](#_Toc85890730)

[Portmanteau test for white noise 47](#_Toc85890733)

[Forecasting 48](#_Toc85890735)

[forecasted values 49](#_Toc85890737)

[correlation results 61](#_Toc85890738)

[**CHAPTER FIVE:** **CONCLUSION AND RECOMMENDATIONS** 62](#_Toc85890739)

[5.1 Introduction 62](#_Toc85890741)

[Conclusion 62](#_Toc85890742)

[Recommendations 63](#_Toc85890743)

[References 64](#_Toc85890744)

**LIST OF TABLES**

[**Table 4.1**: A DICKEY-FULLER (ADF) TEST 36](#_Toc85890718)

[**Table 4.2**: PHILLIPS-PERRON TEST 38](#_Toc85890720)

[**Table 4.3**: A dicky-fuller test in 1st difference 39](#_Toc85890722)

[**Table 4.4**: Philips-Perron test in 1st difference 40](#_Toc85890723)

**Table 4.5:** Model selection criteria…………………………………………….…………44

[**Table 4.6** summarize error 46](#_Toc85890731)

## Table 4.7 correlation results …………………………………………………………..61

**LIST OF FIGURES**

[Figure 4.1 Trend in Monthly/Daily Covid-19 cases between 2020 and 2021 42](#_Toc85890726)

**Figure 4.2** autocorrelation among the COVID-19 confirmed cases….………………....35

[**Figure 4.3** ACF of 1st difference of covid-19 confirmed cases 42](#_Toc85890726)

[**Figure 4.4** PACF of 1st difference of covid-19 confirmed cases 43](#_Toc85890727)

[**Figure 4.5** Residual plot 46](#_Toc85890732)

[**Figure 4.6** inverse roots of ARMA 47](#_Toc85890734)

[**Figure 4.7** confirmed and predicted covid-19 cases 49](#_Toc85890736)

**LIST OF ABBREVIATIONS**

**ARIMA** Auto-Regressive Integrated Moving Average.

**SARIMA** Seasonal Auto Regressive Integrated Moving Average.

**COVID-19** Coronavirus disease 2019 (COVID-19).

**WHO** World Health Organization.

**MERS** Middle East Respiratory Syndrome Severe Acute.

**SARS** Respiratory Syndrome.

**HCW, s** Health Care Workers.

**AR** Autoregressive

**GHS** Ghana Health Service

**MA** Moving Average

**AC**, Autocorrelation

**PACF**, Partial Autocorrelation Function

**ADF**,A Dickey-Fuller

**PP**, Phillips-Perron

**PPE,** Personal Protective Equipment

**SEIR,** Susceptible, Exposed, Infectious and Recoverd

**MAPE,** Mean Absolute Percentage Error

**Ro**, Reproduction number

**SVR**, Support Vector Regression

**RF**, Random Forest

**RMSLE**, Root Mean Squared Log Error

## CHAPTER ONE

**INTRODUCTION**

## 1.1 Background

The Coronavirus disease (COVID-19) is a contagious illness that spread from person to person. Personal touch and minute droplets in an infected person's breath can transmit the virus and cause severe acute respiratory syndrome. For unknown reasons, world health organisation (WHO) originally alerted humankind about the pneumonia in January 2020, and it was learnt globally that this disease is transmitted from person to person. The outbreak of covid-19 began in China, in the city of Wuhan. Ghana confirmed its first instances on March 14, 2020. According to Ghana Health Service (GHS) status report, 796 individuals had died, 1,674 active cases, 85 new cases, and 96,971 verified cases had been reported across the nation from March 14, 2020 to July 08, 2021, and these figures were steadily rising. The Healthcare professionals were in an unusual position. They were in charge of keeping people well, yet the close contact exposes both patients and carers to the illness. The confirmation of the country's first fatality cases has put health personnel in a higher risk, which jeopardized the country's healthcare service system. Despite the fact that the COVID-19 pandemic had been in the nation for a year and some months, many researchers had done a lot of work on it, and it is still being researched on.

This investigation will consist of confirmed instances (cases) based on a machine learning regression model as well as Time series ARIMA. Both models are employed in order to forecast and determine the relationship between the variables. More crucially, a critical component of this pandemic investigation is educating residents about the virus's spread. The function of information researchers and information mining analysts is to coordinate important information, which is more likely to aid in understanding the illness and its quality in the country, allowing for better decisions and solid actions. This will encourage individuals to take more aggressive actions to create frameworks, medications, antibodies, and control of comparable pandemics with better possibilities.

## 1.2 Statement of the problem

Numerous research on COVID-19 have been conducted, and it appears to play a role in the health facilities to seek medical care, mortality prediction, and other activities that might help in the spread of the new COVID-19. For instance, (Sylvia Angeletti, 2020) performed Auto Regressive Integrated Moving Average (ARIMA) model prediction on the Johns Hopkins epidemiological data to predict the epidemiological trend of the prevalence and incidence of COVID-19, where the correlogram reporting the auto correlogram function (ACF) and partial auto correlogram function (PACF) showing that both prevalence and incidence of COVID-19 are not influenced by the seasonality. (Farhan Mohammad Khan and Rajiv Gupta, 2020) have applied the univariate time series model to predict the number of COVID-19 infected cases that can be expected in upcoming days in India, where the daily predicted number of COVID-19 cases in India using the ARIMA model and NAR neural network is approximately 1500 cases per day for the next 50 days. This prediction is made with existing conditions. (Vikas Chaurasia and Saurabh Pal’s, 2020). In his study, two auto integrated (AI) models, ARIMA and regression models, were used to decompose and predict changes in the spread of COVID-19 infection. They investigated these information and predicted that the number of deaths would reduce and cases would shoot up.

When reviewing the literature on the subject, it is apparent that little research has been done which focus on Perceived preparedness to respond to the COVID-19 pandemic: a study with healthcare workers (HCW’s) in Ghana, however HCWs reported low perceived preparedness to respond to COVID-19. Training, clear protocols, personal protective equipment (PPE) availability, isolation wards, and communication all help to improve preparation. The government and other stakeholders must work to strengthen HCWs' readiness to respond to the current outbreak and plan for future pandemics, because the regulatory regimes in the countries where this research were conducted differ from those in Ghana, it is impossible to conclude if the findings are relevant in Ghana. In Ghana, studies focused on readiness, immunization, protocol adherence, and social distance.

Hence there exists a knowledge gap when it comes to the covid-19 confirmed case predictions and the effect of covid-19 on health care delivery.. Given that Ghana is the fastest growing Africa country to be in order to treat covid-19, it is necessary to know the future instances (cases) in the country. As a result, the plan is to forecast future occurrences in the country.

The goal of the study is to give us an estimate of how many confirmed instances we could see in the future, and it is more likely to help us comprehend the illness and its quality, which will allow us to make better decisions and engage in more solid actions. The findings of the investigation can provide insight into understanding the patterns of the occurrence and provide an idea of the epidemiological stage of these cases. This will encourage individuals to take more aggressive actions to create frameworks, medications, antibodies, and control of comparable pandemics with better possibilities.

## 1.3 Objectives

Introduction**:**

Our study's purpose is to forecast the future of confirmed instances (cases) using a machine learning regression model and Time series ARIMA model, as well as to examine the relationship between some variables in the contest of covid-19 in the country.

## 1.3.1 Main objectives:

1. To forecast the future of covid-19 confirmed instances (cases) using Time series ARIMA

## 1.3.2 Specific objectives:

1. Extracting different statistical measures from these graphs.
2. Create a hypothesis to evaluate the significance or dependability of the regression and correlation coefficients

## 1.4. Research Questions.

Introduction:

Providers are using every tool available to them in order to provide care in the safest and most effective manner possible. Ghanaian officials have devised a strategy to ensure that the remainder of the health-care system is not jeopardized. However, it is protected and operational during the pandemic, which means that people may continue to receive important services and feel secure visiting hospitals and clinics. The crisis is actually worldwide, and science is at the vanguard of the virus's struggle. This includes medical personnel attempting to treat the sick while putting their own health at risk, as well as public health officials tracking the virus. Researchers are now introducing diagnostics, treatments, and immunizations, as well as fiercely campaigning for measures such as social isolation to curb the spread of the disease.

All of this will be especially important in a country where resources are few, scientific infrastructure is still in its early stages, and health-care services are underfunded. The research tool will assist you in discovering answers to the following questions.

1. What are the prospects for new confirmed cases on Covid-19 in the future??
2. What effect does this have on healthcare delivery in the country?
3. Has Covid-19 confirmed cases had a link to high covid deaths in the country?

## 1.5 Hypothesis of the study

The study's null hypothesis, marked by Ho: state that there is no association between covid-19 confirmed cases and covid-19 death in the country.

And the alternative hypothesis, marked by Hi, asserts that there is a link between covid-19 confirmed cases and death in the country.

## 1.6. Significant of the study,

a) Researchers and information mining analysts.

The study will aid in the coordination of important information, and it is more likely to aid in the understanding of the illness and its quality in the country, allowing for better decisions and solid actions. This will encourage individuals to take more aggressive efforts.

b) Health Practitioners

This study should be useful to health practitioners, Non-Governmental Organizations (NGOs), investors, and other parties that want to examine specific health regulatory in the country based on the performance of the country in dealing with covid-19 and other health parameters.

c) Academicians and Researchers

This study is intended to serve as a foundation for future research and as a reference point for academics and researchers alike, as it will provide insight into the rapid spread of the novel coronavirus and its resulting conditions.

## 1.7. Study area and scope.

The study is primarily concerned with forecasting new cases in the country in terms of covid-19, expectations of new cases from July 8th, 2021 to 31st March, 2022. Structured interviews will not be scheduled by us. We shall be influenced most broadly by the interpretative viewpoint, and most narrowly by Vikas Chaurasia1 and Saurabh Pal’s (29 August 2020) ARIMA and Regression approach. The interpretative viewpoint emphasizes analyzing the meanings and viewpoints of covid-19 case prediction, as well as how these meanings are mediated. The study will also look into the distribution of covid-19 throughout the country. In addition, the research report will integrates my grasp of the relevant theory and past research with the findings of my empirical investigation. Only secondary data from the District Health Information Management System (DHIMS) were used in this research.

## 1.8. Limitation of the study.

The semester's scheduling restrictions need less time than would be optimal for a quantitative investigation working as a National Service Personnel and a student at the same time. Submissions of my chapters and the comb binding of my work on working days will have an impact on me because I will be at work post on those days.

Secondly, one clear disadvantage of this study is that the amount of data to be utilized may not be up to date, since new cases do not remain constant. Issued an update on covid-19 for the 2020–2021 year, and the quality the validity of this study is solely dependent on the correctness, dependability, and quality of the secondary data source. The outcome may be influenced by approximation and relative measure in relation to the data source.

## 1.9 Outline/Organization of the study.

This is a high-level overview/summary of the study's structure. The study paper is divided into five (5) sections. The "first chapter outlines the study's background, problem statement, aims/objectives, research questions, hypotheses, and significance.", The study's scope and limitations. The second chapter includes a review of the literature, which includes theoretical and empirical investigations on the ARIMA and Regression Analysis on New Cases Forecasts. The third chapter is dedicated to research methods. The data analysis and empirical findings are presented in the fourth chapter. The last chapter summarizes the findings and makes recommendations.

## CHAPTER TWO

## LITERATURE REVIEW

## 2.0 Introduction:

This chapter provides an overview of prior research on the subject. The part includes a theoretical, and empirical review. In this section, there is a review of the work of several authors concerning concept definitions and various researches done to uncover the academic work.

## 2.1 Topics related to the research problem.

The Pandey Etal SEIR and Regression Model Based Covid-19 Outbreak Prediction in India. Application of the ARIMA model on the COVID-19 epidemic dataset (Domenico Benvenuto, Marta Giovanetti, Lazzaro Vassallo, Silvia Angeletti, Massimo Ciccozzi). Descriptive analysis of the data was performed, and to evaluate the incidence of new confirmed cases of COVID-19 and to prevent eventual bias, the difference between the cases confirmed on that day and the cases confirmed on the previous day were calculated. COVID‑19 Pandemic: ARIMA and Regression Model‑Based Worldwide Death Cases Predictions (Vikas Chaurasia, Saurabh Pal). COVID-19 virus outbreak forecasting of registered and recovered cases after sixty day lockdown in Italy: A data driven model approach (Nalini Chintalapudi, Gopi Battineni, Francesco Amenta). SEIR and Regression Model based COVID-19 outbreak predictions in India (Gaurav Pandey, Poonam Chaudhary, Rajan Gupta, Saibal Pal).

## 2.2 Work of other researchers

Recent studies on COVID-19 include only exploratory analysis of the available limited data. Effective and well-tested vaccine against COVID-19 had not been invented and hence a key part in managing this pandemic is to decrease the epidemic peak, also known as flattening the epidemic curve. The role of data scientists and data mining researchers is to integrate the related data and technology to better understand the virus and its characteristics, which can help in taking right decisions and concrete plan of actions. It will lead to a bigger picture in taking aggressive measures in developing infrastructure, facilities, vaccines and restraining similar epidemics in future. Time series data provided by John Hopkins University; USA has been used for the empirical result analysis. The time period of data is from 30/01/2020 to 30/03/2020. The data includes confirmed cases, death cases and recovered cases of all countries. The Pandey Etal SEIR and Regression Model Based Covid-19 Outbreak Prediction in India transmission is person to person via respiratory droplets among close contact with the average number of people infected by a patient being 1.5 - 3.5 but the virus is not considered airborne. There exist a large number of evidences where machine learning algorithms have proven to give efficient predictions in healthcare. Nsoesie etal has provided a systematic review of approaches used to forecast the dynamics of influenza pandemic. They have reviewed research papers based on deterministic mass action models, regression models, prediction rules, Bayesian network, SEIR model, ARIMA forecasting model etc.

These models can be used to evaluate disease from within the host model i.e. influence interaction within the cells of the host to meta population model i.e. how its spread in geographically separated populations. The most important part of this model is to calculate the R0 value. The value of R0 tells about the contagiousness of disease. It is the fundamental goal of epidemiologists studying a new case. In simple terms R0 determines an average of what number of people can be affected by a single infected person over a course of time. If the value of R0 < 1, this signifies the spread is expected to stop. If the value of R0 = 1, this signifies the spread is stable or endemic. If the value of R0 >, 1 this signifies the spread increasing in the absence of intervention. Equation (1) calculates the percentage of the population needed to be vaccinated to stabilize the spread of disease. 2.1 In this study, two AI models, ARIMA and regression models were used to decompose and predict changes in the spread of covid-19 infection. We have investigated this information and predicted that the number of deaths will be reduced compared to the overall situation. The decline shown in the ARIMA model graph indicates that the future mortality rate will decrease (based on the current situation). The training dataset verified by the mean absolute percentage error (MAPE=99.09%) indicates the accuracy of the model. The regression model also indicated an increase in the initial number of deaths, but over time, it predicted fewer deaths than actual deaths from 2nd May 2020. Based on the above result and discussion, through ARIMA and regression models, we can conclude that there is possibility of reducing deaths worldwide and should be reduced. Over time, there must be new opportunities to deal with this pandemic. Many researchers, scientists, doctors, nurses, medical support staff, and government agencies are all playing their roles. However, we ourselves have a responsibility to follow the guidelines provided by these agencies. If we do not maintain social estrangement, gather in public places, and do not keep the neighborhood clean, how can we overcome the covid-19 pandemic?

Additionally, Nalini Chintalapudi, Gopi Battineni, Francesco Amenta (13 April 2020) also conducted a related study on the topic (COVID-19 virus outbreak forecasting of registered and recovered cases after sixty-day lockdown in Italy: A data driven model approach), they indicated that, Predictions were done with 93.75% accuracy for registered case models and 84.4% accuracy for recovered case models. The forecasting of infected patients could be reaching the value of 182,757, and recovered cases could be registered value of 81,635 at end of May. In their Conclusions: Their study highlights the importance of country lockdown and self-isolation to control the disease transmissibility among Italian population through data driven model analysis. Their findings suggest that nearly 35% decrement of registered cases and 66% growth of recovered cases will be possible. It is also evident in their findings that from 60- day forecasting of infected cases might rise in between the range of 105,732 and 182,757, and recovered cases could increase in between the range of 16,742 and 81,635 with Confidence Interval of 80% and 95%. The regressive distribution of patient cases while two plots had observed to estimate the fitting accuracy. Their model validation was assessed by prediction errors. Based on the ARIMA model accuracy evolution of COVID-19 Italian epidemic data on mentioned time period, they considered mean absolute prediction error (MAPE) parameter. The accuracy (Acc) is defined in their equation; “Acc 100-MAPE\*100. The models of ARIMA (1, 2, 0) registered, and ARIMA (3, 2, 0) recovered cases are validated with an accuracy of 93.75%, 84.4% respectively.

## 2.2 Theoretical literature

To prevent ambiguity and idiosyncratic interpretations of key terminologies utilized in this study, I've defined the terms used in this study below.

**ARIMA** Auto-Regressive Integrated Moving Average.

**SARIMA** Seasonal Auto Regressive Integrated Moving Average.

**COVID-19** Coronavirus disease 2019 (COVID-19).

**WHO** World Health Organization.

**MERS** Middle East Respiratory Syndrome Severe Acute.

**SARS** Respiratory Syndrome.

**HCW, s** Health Care Workers.

**AR** Autoregressive

**GHS** Ghana Health Service

**MA** Moving Average

**AC**, Autocorrelation

**PACF**, Partial Autocorrelation Function

**ADF**,A Dickey-Fuller

**PP**, Phillips-Perron

**PPE,** Personal Protective Equipment

**SEIR,** Susceptible, Exposed, Infectious and Recoverd

**MAPE,** Mean Absolute Percentage Error

**Ro**, Reproduction number

**SVR**, Support Vector Regression

**RF**, Random Forest

**RMSLE**, Root Mean Squared Log Error

## 2.3 Empirical literature

The empirical literature talks about the results of others related to our topics;the following are description of some of the researches presented.

Some ARIMA models with different ARIMA boundaries were selected, which includes ARIMA (0, 2, 1) for the lowest MAPE (4.7520) for Italy, similarly for Spain and France selected separately with ARIMA (1, 2, 0) and ARIMA (0, 2, 1) and the lowest MAPE (5.58486) and (5.6335) respectively. This test shows that the ARIMA model is appropriate to understand the effect of COVID-19. The aftereffects of the examination can reveal insight into understanding the patterns of the episode and give a thought of the epidemiological phase of these locales. In Mahdim etal’s work he stated that for the purpose of time series analysis, different models such as ARIMA, Cubist regression, RF, Ridge regression, SVR and stacking-ensemble method were assessed. His created models can produce exact forecasting, with errors of 0.87–3.51%, 1.02–5.63%, and 0.95–6.90% in 1, 3, and 6 days respectively. His positioning of models, from the best to the most noticeably worst with respect to precision, in all situations is SVR, stacking–gathering learning, ARIMA, Cubist regression, Ridge regression, and RF models. Taking into account the data accumulated from the Johns Hopkins University depository in the period from January 30, 2020 to March 30, 2020, the SEIR model and the regression model were used. RMSLE evaluated the introduction of the model, and the data of the SEIR model were 1.52 and 1.75, respectively. The RMSLE tightening rate between the SEIR model and the regression model is 2.01. In addition, the estimation of R0 as the diffusion of pollution was analyzed to 2.02. It is foreseeable that in the next 14 days, the number of cases may rise to 5000–6000. Chakraborty and Ghosh collected the data as of April 4, 2020, and showed a pandemic flare-up in excess of 1,116,643 affirmed diseases and in excess of 59,170 revealed deaths around the world. The primary objective of their paper is in two fold (1) producing present moment (constant) estimates of things to come off COVID-19 cases for various nations; (2) chance evaluation of the novel COVID-19 for some significantly influenced nations. To take care of the primary issue, they introduced a half breed approach dependent on ARIMA model and wavelet-based forecasting model that can create present moment (10 days ahead) conjectures of the quantity of day by day affirmed cases for Canada, France, India, South Korea, and the UK. They applied an ideal relapse tree calculation to discover basic causal factors that altogether influence the case casualty rates for various nations. Chintalapudi also studied, from mid-February to the end of March, COVID-19 data of deleted cases registered and restored on-site by the Italian Ministry of Health where his appointment of the accidental ARIMA vision group using R real model was completed. The accuracy of the enrollment case model reached 93.75%, and the accuracy of the recovery case model reached 84.4%, he discovered that at the end of May, the forecasting of infected patients could reach the value of 182,757, and recovered cases could be a registered value of 81,635. The above findings indicate that it is possible to reduce enrollment of cases by approximately 35% and improve recovery of cases by approximately 66%.

## 2.4 Conclusion

According to the literature, studies on Covid-19 have been undertaken all over the world, but none have been undertaken in Ghana. By filling that vacuum, our study aims to contribute to the literature.

## CHAPTER THREE

## METHODOLOGY

## 3.0 introduction

The approach utilized to perform the study is discussed in this chapter. Discussions on research design, population, sampling, data collection, data processing, and relativity testing are among the topics covered.

## 3.1 Research Design

The process of gathering data and assessing situations in such a way that blends extreme relevance with the objective of the study to be done is known as research design (Kothari, 1990). It is the theoretical framework within which research, data collecting, and analysis are carried out. This research is an explanatory study that aims to explain relationships. The desk study matches the concept since it used a computer to retrieve data from the District Health Information System 2. (DHIMS2). The investigation was quantitative since it used numerical values.

## Population of Study

A population generally has a lot of characters to examine, an inquiry is usually confined to one or more samples taken from it. All the regions in the country were suitable for this study and for the foundation of prediction, based on the goal. All regions in the country are prone to covid-19 as listed on the Ghana Health service website, the country’s data (all regions) was selected for the study as the total population.

## Sample Size

The study used all regions covid-19 data in the country. Based on convenience, all the sixteen (16) regions from March 2020 to July 2021 were chosen for this study. This is because data for this study period was accessible.

## 3.2 Data Collection

Secondary data from the District Health Information System 2 (DHIMS2) was used in this investigation. The period under consideration was 2020-2021. The information comes from the District Health Information System 2 (DHIMS2) and includes a set of covid-19 statements (data) from the country obtained from the District Health Information System 2 (DHIMS2).

## 3.2.1 Reliability and Validity of Data

In the case of validity, the “data was gathered from the District Health Information System 2 (DHIMS2). The website is a recognized and an accepted place to collect accurate data. In ensuring reliability, several statistical tests were conducted on the data to ensure that the results were not misleading.

## 3.3 Data Analysis

Data was analyzed using Stata 13.1, Microsoft excel was used to build the data, prediction was run in Stata software. The data set had information in two dimensions; both the time (2020-2021) and in different regions, which correspond to panel data, which is often used in a situation where data includes both time series and cross-sectional elements (different regions). Panel data regression has some advantages. Firstly, panel information can handle more complex data because they combine both cross-section and time series data. This leads to increased rates of freedom and increased test power. In addition, the effect of some of the variables left on the regression can be mitigated from panel regression (Brooks 2014). Gujarati (2004) also pointed out that paneled information can bring out better and measure consequences which could not be done with clean cross-section or time series data because panel information provides more information, greater variation, less variable collinearity, greater efficiency, and better dynamics.

However, panel data has some evaluation and termination problems. Since this information includes both cross-sectional and time dimensions, it is necessary to look at the problem of cross-sectional data (e.g., Heteroscedasticity) and time series data (e.g., autocorrelation) (Gujaratis, 2004) thereby making it quite complicated.

## 3.3.1 Time Series Analysis

Time series analysis comprises methods or processes that breakdown a series into components and explainable portions that allows trends to be identified, estimates and forecasts to be made. Basically time series analysis attempts to understand the underlying context of the data points through the use of a model to forecast future values based on known past values. Such time series models include MA, AR, ARIMA, GARCH, TARCH, EGARCH, FIGARCH, CGARCH, etc but the main focus of this study is based on MA, AR, and ARIMA models.

**3.3.2 Lag**

Lag is the time periods between two observations. For example, lag 1 is between Yt and Yt-1. Lag 2 is between Yt and Yt-2. Time series can also be lagged forward, Yt and Yt+1. the observation at the current time, Yt , depends on the value of the previous observation, Yt-1.

**3.3.3 Differencing**

Differencing simply means subtracting the value of an earlier observation from the value of a later observation. Calculating differences among pairs of observations at some lag to make a non-stationary series stationary. There are possible shifts in both the mean and the dispersion over time for this series. The mean may be edging upwards, and the variability may be increasing. If the mean is changing, the trend is removed by differencing once or twice. If the variability is changing, the process may be made stationary by logarithmic transformation. Differencing the scores is the easiest way to make a non-stationary mean stationary (flat). The number of times you have to difference the scores to make the process stationary determines the value of d. If d=0, the model is already stationary and has no trend. When the series is differenced once, d=1 and linear trend is removed. When the difference is then differenced, d=2 and both linear and quadratic trend are removed. For non stationary series, d values of 1 or 2 are usually adequate to make the mean stationary.

**3.3.4 Stationary and Non-stationary**

Series Stationary, series vary around a constant mean level, neither decreasing nor increasing systematically over time, with constant variance. Non-stationary series have systematic trends, such as linear, quadratic, and so on. A non-stationary series that can be made stationary by differencing is called ―non-stationary in the homogenous sense. Stationarity is used as a tool in time series analysis, where the raw data are often transformed to become stationary. For example, economic data are often seasonal or dependent on a non-stationary price level. Using non-stationary time series produces unreliable and spurious results and leads to poor understanding and forecasting. The solution to the problem is to transform the time series data so that it becomes stationary. If the non-stationary process is a random walk with or without a drift, it is transformed to stationary process by differencing. Differencing the scores is the easiest way to make a non-stationary mean stationary (flat). The number of times you have to difference the scores to make the process stationary determines the value of d. If d = 0, the model is already stationary and has no trend. When the series is differenced once, d=1 and linear trend is removed. When the difference is then differenced, d =2 and both linear and quadratic trend are removed. For non-stationary series, d values of 1 or 2 are usually adequate to make the mean stationary. If the time series data analysed exhibits a deterministic trend, the spurious results can be avoided by detrending. Sometimes the non-stationary series may combine a stochastic and deterministic trend at the same time and to avoid obtaining misleading results both differencing and detrending should be applied, as differencing will remove the trend in the variance and detrending will remove the deterministic trend. A non-stationary process with a deterministic trend becomes stationary after removing the trend, or detrending. For example, Yt = α + βt + εt is transformed into a stationary process by subtracting the trend βt: Yt - βt = α +εt . No observation is lost when detrending is used to transform a non-stationary process to a stationary one. Non-stationary data, as a rule, are unpredictable and cannot be modelled or forecasted. The results obtained by using non-stationary time series may be spurious in that they may indicate a relationship between two variables where one does not exist. In order to receive consistent, reliable results, the non-stationary data was transformed into stationary data. In contrast to the non-stationary process that has a variable variance and a mean that does not remain near, or returns to a long-run mean over time, the stationary process reverts around a constant long-term mean and has a constant variance independent of time.

**3.4 Components of Time Series**

A vital step in choosing our appropriate modeling and forecasting procedure is to consider the type of data patterns exhibited from the time series graphs of the time plots. The sources of variation in terms of patterns in time series data are mostly classified into four main components:

1. Horizontal – when data values fluctuate around constant value
2. Trend – when there is long term increase or decrease in the data
3. Seasonal – when a series is influenced by seasonal factors and recurs on a regular periodic basis.
4. Cyclic – when the data exhibit rises and falls that are not of a fixed period.

**3.4.1 The Trend (d)**

The trend is simply the underlying long term behavior or pattern of the data or series. The Australian Bureau of Statistics (ABS, 2008) defined trend as the long term ‘movement in a time series without calendar related and irregular effects, and is a reflection of the underlying level. It is the result of influences such as population growth, price inflation and general economic changes. A model with two trend terms (dC2) has to be differenced 32 twice to make it stationary. The first difference removes linear trend, the second difference removes quadratic trend, and so on.

**3.4.2 Seasonal Variation (S)**

A seasonal effect is a systematic and calendar related effect. Some examples include the sharp escalation in most Retail series which occurs around December in response to the Christmas period, or an increase in water consumption in summer due to warmer weather. Other seasonal effects include trading day effects (the number of working or trading days in a given month differs from year to year which will impact upon the level of activity in that month) and moving holidays (the timing of holidays such as Easter varies, so the effects of the holiday will be experienced in different periods each year). Seasonal adjustment is the process of estimating and then removing from a time series influences that are systematic and calendar related. Observed data needs to be seasonally adjusted as seasonal effects can conceal both the true underlying movement in the series, as well as certain non-seasonal characteristics which may be of interest to analysts. Seasonality in a time series can be identified by regularly spaced peaks and troughs which have a consistent direction. Other techniques that can be used in time series analysis to detect seasonality include:

1. A seasonal subseries plot is a specialised technique for showing seasonality.

2. Multiple box plots can be used as an alternative to the seasonal subseries plot to detect seasonality.

3. The autocorrelation plot can help identify seasonality.

**3.4.3 Common Assumptions in Time Series**

Techniques and common assumption in many time series techniques is that the data are stationary. A stationary process has the property that the mean, variance and autocorrelation structure do not change over time.

If the time series is not stationary, we can often transform it to stationary with one of the following techniques:

1. We can difference the data. That is, given the series Zt , the differenced data will contain one less point than the original data. Although you can difference the data more than once, one difference is usually sufficient.

2. If the data contain a trend, we can fit some type of curve to the data and then model the residuals from that fit. Since the purpose of the fit is to simply remove long term trend, a simple fit, such as a straight line, is typically used.

3. For non-constant variance, taking the logarithm or square root of the series may stabilize the variance. For negative data, you can add a suitable constant to make the entire data positive before applying the transformation. This constant can then be subtracted from the model to obtain predicted (i.e., the fitted) values and forecasts for future points.

**3.4.5 Autocorrelation Function (ACF)**

Autocorrelation refers to the correlation of a time series with its own past and future values. Autocorrelation is also sometimes called ―lagged correlation” or “serial correlation‖, which refers to the correlation between members of a series of numbers arranged in time. The pattern of autocorrelations in a time series at numerous lags; the correlation at lag 1, then the correlation at lag 2, and so on. Correlations among 35 sequential scores at different lags.

**3.5.6 Partial Autocorrelation Function (PACF)**

Partial autocorrelation function measures the degree of association between Yt and Yt+k when the effect of other time lags on Y are held constant. The partial autocorrelation function PACF denoted by the set of partial autocorrelations at various lags k are defined by ( k=1, 2, 3…). Specifically, partial autocorrelations are useful in identifying the order of an autoregressive model. The partial autocorrelation of an AR (p) process is zero at lag p+1 and greater. The approximate 95% confidence interval for the partial autocorrelations is at +2/N .

Partial autocorrelation plots are formed by:

* Vertical axes: partial autocorrelation at coefficient at lag, k,
* Horizontal axes: time lag k (k = 0, 1, 2...)

In addition, 95% confidence interval bands are typically included on the plot.

**3.6 Common Approaches to Univariate Time Series**

There are a number of approaches to modeling time series. Few of the most common approaches are below:

**3.6.1 Decomposition**

One approach is to decompose the time series into a trend, seasonal, and residual component. In other words decomposition refers to separating a time series into trend, cyclical, and irregular effects. Decomposition may be linked to de-trending and deseasonalising data so as to leave only irregular effects, which are the main focus of time series analysis. Triple exponential smoothing is an example of this approach. Another example, called seasonal loess, is based on locally weighted least squares and is discussed by Cleveland (1993).

**3.6.2 Autoregressive (AR) Models**

An autoregressive model is simply a linear regression of the current value of the series against one or more prior values of the series. The value of p is called the order of the AR 39 model. AR models can be analyzed with one of various methods, including standard linear least squares techniques. They also have a straightforward interpretation. A common approach for modeling univariate time series is the autoregressive (AR) model:

In this general case, the ACF damps down and the PACF cuts off after p lags. An AR (p) model is stationary if the roots of φ(L) = 0 all lie outside the unit circle. A necessary condition for stationary is that rk = 0 as k = ∞.

**3.6.3 Moving Average (MA)**

Models Moving average model is conceptually a linear regression of the current value of the series against the white noise or random shocks of one or more prior values of the series. The random shocks at each point are assumed to come from the same distribution, typically a normal distribution, with location at zero and constant scale. The distinction in this model is that these random shocks are propagated to future values of the time series. Fitting the MA estimates is more complicated than with AR models because the error terms are not observable. This means that iterative non-linear fitting procedures need to be used in place of linear least squares. MA models also have a less obvious interpretation than AR models. Moving Average (MA) is another common approach for modeling univariate time series models is the moving average (MA) model. An MA (q) is said to be invertible if φ(L) can be inverted, in other words if it can be expressed as an AR. An MA (q) is invertible if the roots of φ(L)=0 all lie outside the unit circle. A finite AR is always invertible. The random shocks at each point are assumed to come from the same distribution, typically a normal distribution, with location at zero and constant scale. The distinction in this model is that these random shocks are propagated to future values of the time series. Sometimes the ACF and PACF will suggest that a MA model would be a better model choice and sometimes both AR and MA terms should be used in the same model. It is also important to note, however, that the error terms after the model is fit should be independent and follow the standard assumptions for a univariate process. Box and Jenkins popularized an approach that combines the moving average and the autoregressive approaches (Box, Jenkins, and Reinsel, 1994). This resulted in autoregressive moving average model (ARMA). The Box-Jenkins model assumes that the time series is stationary. Box and Jenkins recommend differencing non-stationary series one or more times to achieve stationarity. Doing so produces an ARIMA model, with the "I" standing for Integrated. This is described in detail below since it is the main method used in the analysis of data in this research.

**3.6.4 Autoregressive Moving Average (ARMA)**

Models Autoregressive and Moving Average processes can be combined to obtain a very flexible class of univariate processes (proposed by Box and Jenkins), known as ARMA processes. The time series Xt is an ARMA (p, q) process, if it is stationary and An ARMA process is stationary if the roots of φ(L) all lie outside the unit circle and invertible if the roots of θ(L) all lie outside the unit circle.

The acronym for an auto-regressive integrated moving average model. The three terms to be estimated in the model are auto-regressive (p), integrated (trend—d), and moving average (q).The ARIMA (auto-regressive, integrated, moving average) model of a time series is defined by three terms (p, d, q). Identification of a time series is the process of finding integer, usually very small (e.g., 0, 1, or 2), values of p, d, and q that model the patterns in the data. When the value is 0, the element is not needed in the model. The middle element, d, is investigated before p and q. The goal is to determine if the process is stationary and, if not, to make it stationary before determining the values of p and q. Recall that a stationary process has a constant mean and variance over the time period of the study.

**3.7 Box-Jenkins ARIMA Process**

In statistics and econometrics, and in particular in time series analysis, an autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. The Box–Jenkins methodology, named after the statisticians George Box and Gwilym Jenkins, applies ARIMA models to find the best fit of a time series to past values of this time series, in order to make forecasts. They are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the ―integrated‖ part of the model) can be applied to remove the non-stationarity. The model is generally referred to as an ARIMA (p, d, q) model where p, d, and q are non-negative integers that refer to the order of the autoregressive, integrated, and moving average parts of the model respectively.

**3.7.1 Box-Jenkins Modeling Approach**

The Box-Jenkins model uses iterative three-stage modeling approach which is:

1. Model identification and model selection: making sure that the variables are stationary, identifying seasonality in the dependent series (seasonally differencing it if necessary), and using plots of the autocorrelation and partial autocorrelation functions of the dependent time series to decide which (if any) autoregressive or moving average component should be used in the model.

2. Parameter estimation using computation algorithms to arrive at coefficients which best fit the selected ARIMA model. The most common methods use maximum likelihood estimation or non-linear least-squares estimation.

3. Diagnostic and forecasting, Model checking by testing whether the estimated model confirms 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 mean and variance of residuals over time and performing a ADF and PP test or plotting autocorrelation and partial autocorrelation 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.

**3.8 Box-Jenkins Model Identification**

**3.8.1 Stationarity and Seasonality**

The first step in developing a Box–Jenkins model is to determine if the time series is stationary and if there is any significant seasonality that needs to be modeled.

**3.8.2 Detecting Stationarity**

Stationarity can be assessed from a run sequence plot. The run sequence plot should show constant location and scale. It can also be detected from an autocorrelation plot. Specifically, non-stationarity is often indicated by an autocorrelation plot with very slow decay. Finally, unit root tests provide a more formal approach to determining the degree of differencing such as A Dicky Fuller and Phillips-Perron Unit Root Tests are carried out employing the unit root testing procedures of Hamilton (1994). The Dicky Fuller test for the null hypothesis of a level stationary against an alternative of unit root together with the Philips-Peron test for the null hypothesis of a unit root against the alternative of a stationary series. The decision rule is that for the ADF test if the p-value of its test statistic is greater than the critical value, say 0.05, then reject the null hypothesis of having a level stationary series and therefore conclude the alternate hypothesis that it has a unit root. The Philips-Peron Test, on the other hand, test for the null hypothesis of unit root against an alternative hypothesis of stationarity by rejecting the null hypothesis if its p-value is less than the critical value chosen.

**3.8.3 Differencing to achieve Stationarity**

Box and Jenkins recommend the differencing approach to achieve stationarity. However, fitting a curve and subtracting the fitted values from the original data can also be used in the context of Box-Jenkins models.

**3.8.4 Identifying p and q**

Once stationarity and seasonality have been addressed, the next step is to identify the order (i.e. the p and q) of the autoregressive and moving average terms. These are determined by examining the values of the autocorrelations and the partial autocorrelations with their corresponding plots as explained below.

**3.8.5 Autocorrelation and Partial Autocorrelation Plots**

The primary tools for doing this are the autocorrelation plot and the partial autocorrelation plot. The sample autocorrelation plot and the sample partial autocorrelation plot are compared to the theoretical behaviour of these plots when the order is known.

**3.8.6 Best Model Identification and Selection Criteria**

**H0:** The data are independently distributed (i.e. the correlations in the population from which the sample is taken are 0, so that any observed correlations in the data result from randomness of the sampling process).

**Ha:** The data are not independently distributed.

The choice of a plausible model depends on its p-value for the modified Box-Jenkins if is well above .05, indicating ―non-significance. In other words, the smaller the p-value, the better the model.

**3.9 Some Applications of ARIMA Model in Real Life Situations**

ARIMA modeling techniques have been applied in many fields of research. For example, Aidoo (2010) applied ARIMA model on the monthly inflationary rates in Ghana. He indicated that Ghana faces a macroeconomic problem of inflation for a long period of time. The problem in somehow slows the economic growth in this country. Using monthly inflation data from July 1991 to December 2009, we find that ARIMA (1,1,1)(0,0,1)12 can represent the data behavior of inflation rate in Ghana well. Based on the selected model, we forecast seven (7) months inflation rates of Ghana outside the sample period (i.e. from January 2010 to July 2010). The observed inflation rate from January to April which was published by Ghana Statistical Service Department fall within the 95% confidence interval obtained from the designed model. The forecasted results show a decreasing pattern and a turning point of Ghana inflation in the month of July. Again, Cui (2011) researched on the topic: ―ARIMA Models for Bank Failures: Prediction and Comparison.‖ They said that the number of bank failures has increased dramatically over the last twenty-two years. A common notion in economics is that some banks can become ―too big to fail.‖ Is this still a true statement? What is the relationship, 48 if any, between bank sizes and bank failures? In this thesis, the proposed modeling techniques are applied to real bank failure data from the FDIC. In particular, quarterly data from 1989:Q1 to 2010:Q4 are used in the data analysis, which includes three major parts: 1) pairwise bank failure rate comparisons using the conditional test (Przyborowski & Wilenski, 1940), 2) development of the empirical recurrence rate (Ho, 2008) and the empirical recurrence rates ratio time series; and 3) the Autoregressive Integrated Moving Average (ARIMA) model selection, validation, and forecasting for the bank failures classified by the total assets.

**3.9.1 ARIMA Modeling in Disease Surveillance: Advantages and Disadvantages.**

According to Reis and Mandl (2003), one of the primary benefits of ARIMA models is their ability to correct for local trends in the data – what has happened on the previous day is incorporated into the forecast of what will happen today. This works well, for example, during a particularly severe flu season, where prolonged periods of high visit rates are adjusted to by the ARIMA model, thus preventing the alarm from being triggered every day throughout the flu season. However, if the ARIMA model ―adjusts‖ to an actual outbreak instead of detecting it, a slowly spreading outbreak or attack might be missed because of this correction. This correction is most likely to affect detection of outbreaks occurring over several days, rather than those that occur suddenly. It is therefore also important to rely on the non-ARIMA or non-classical ARIMA models for outbreak detection. ARMA models in Reis and Mandl (2003) and Reis, Pagano, and Mandl (2003) require large historic records of patient visits in order to begin surveillance. This is a substantial disadvantage. As can be seen from Moore et al. (2002), in some cases long historical data are not available and not necessary. Also, combining both historical and recent trends is quite realistic. Another disadvantage of ARMA is that the corresponding detector is not sensitive to the slow growth. According to Rizzo et al. (2005), outbreaks that evolve over a matter of days, for example, can often be detected with ARMA models that generate single-day predictions based on historical data. More gradually developing outbreaks are generally easier to detect by using such techniques as CuSum (Hawkins & Olwell, 1998).

**3.9.2 Regression Model**

Linear regression is a prescient measurable methodology for displaying connection between a dependent variable with a given arrangement of autonomous factors. It is a direct way to deal with displaying the connection between a dependent variable and at least one independent variable. At the point when we have just a single independent variable, it is as called straightforward linear regression. For more than one independent factor, the procedure is called multiple linear regressions. This investigation has utilized linear regression and multiple regressions for expectation of COVID-19 cases [12]. The linear regression description includes a linear condition that adds a specific information literacy particular arrangement x, whose response is the predictable return y of the data particular arrangement (y). The linear condition gives each information value or part a scale factor, called the coefficient, which is represented by the Greek word beta β. including an additional coefficient in the same way provides additional degrees of freedom for the line and is repeatedly called the intercept or offset coefficient. In a straightforward regression issue, the type of the model would be:

y = β0 + β1x. Where β0 is the intercept, β1 is the coefficient, x is the independent variable, and y is the dependent variable. In higher estimates, when we have multiple information x, the line is called a plane or hyperplane. Described in this way are the kinds of conditions and specific characteristics for the coefficients (β0 and β1). The general condition for a multiple linear regression with n independent factors is: **y = β0 + β1 x1 + β1x2 + ⋯ + βnXn + ε**, Where β0, β1, β2… βn are the coefficients, x1, x2,…. Xn–x are the variables, y is the y-variable, and ϵ is the random error “noise”.

**3.9.3 Correlation Coefficients**

The statistical measure correlation coefficient is the strength of the relationship between the relative motions of two variables. The range is defined as −1 to+1. Incorrect correlation measurement occurs when the values are greater than+1 and less than −1. The correlation measurement at −1 is completely negative, the correlation measurement at+1 is positive, and the value at 0.0 is the nonlinear relationship between the two variables

## CHAPTER FOUR

## DATA ANALYSIS AND RESULTS

## 4.0 Introduction

This chapter provides an overview of the various models and their examination, as well as a discussion of the findings. The AR, MA, ARMA and ARIMA model shall be used in the modelling. The data covered covid-19 data collected on a daily basis from 14th March 2020 to 8th July 2021. The variables are described in detail. The regression findings, as well as the other tests performed on the data, are reported.

**4.1 Data Analysis**

Confirmed cases, total cases, death cases, and recovered cases are all included in the statistics. However, for the sake of analysis and prediction of COVID-19 cases, this paper focuses solely on the confirmed cases data. For the purpose of calculating and forecasting the number of COVID-19 instances. It is recommended that a lengthy time series data is required for univariate time series forecasting. Meyler etal (1988) recommended that at least 50 observations should be used for such a univariate time series forecasting. This could be problematic if few observations are used. However, when using a long time series data, it could be possible that the series contains a structural break which may necessitate only examining a subsection of the entire data series or alternatively using intervention analysis or dummy variables. This is because there may be some conflict between the 50 needed for sufficient degrees of freedom for statistical robustness and having a shorter data sample to avoid structural breaks. The series should be plotted against time to assess whether any structural breaks, outliers or data errors occurred. This step may also reveal whether there is significant seasonal pattern in the times series or not. A dimension of the preliminary analysis for examining non-stationarity of the data is by considering the time series plot of our total confirmed cases from 14 March, 2020 to 8 July, 2021 as shown in Figure 4.1.

## ARIMA and Box-Jenkins model selections

We focus on identification of the ARIMA model, ARIMA is one of the most widely used approaches to time series forecasting.

* ARIMA( p,d,q): Autoregressive Integrated Moving Average

Box and Jenkins (1970) introduced three steps methods to select appropriate models for estimating and forecasting univariate models.

“Let my past predict my future”

Box and Jenkins three stages are

* Identification
* Estimation
* Diagnostic and forecasting

Under **identification**, we focus on

* (a) Stationarity
* (b) ARIMA(p,d,q), that determine p,d,q

We look at the properties of the variable we're interested in. **Is it stationary**? If the mean and covariance of a series do not change with time, it is said to be covariance stationary. A stationary series has no trend, has constant amplitude variations around its mean, and wiggles continuously. We'll take a difference of our COVID-19 data to remove the trend because it's non-stationary.

**Stationarity:** Graph, correlogram and formal test (ADF, PP) were done on the data.

* If stationary, Then we use ARMA model (p,q)
* If Non-stationary: Then we use ARIMA model (p,d,q)

The following codes were entered into Stata to describe our COVID-19 data as time series data.

examining non-stationarity of the data is by considering the time series plot of total cases from 14 March, 2020 to 8 July, 2021 as shown in Figure 4.1.



Figure 4.1**: Trend in Monthly/Daily Covid-19 cases between 2020 and 2021**

Figure 4.1 revealed that COVID cases from 2020 to 2021 have been largely seasonal and a positive trend. This means that there is no stationarity in the data (non-stationary) since it follows a trend (dependent on time).

We therefore compliment the above Graph with a correlogram,

The diagram below has produce an autocorrelation among the COVID-19 confirmed cases, which we can observe that the decay is really slow (follow a slow decay pattern) and it is a sign that our variable is not stationary.

****

## Figure 4.2: autocorrelation among the covid-19 confirmed cases.

We then again perform a formal test (ADF and PPerron tests) to confirm the above result (non-stationarity)

**For the Trend**

Ho: Trend > 0.05(the trend does not belong to the series)

H1: Trend < 0.05(the trend does belong to the series)

**For the Constant**

Ho: Constant > 0.05(the constant does not belong to the series)

H1: Constant < 0.05 (the constant belongs to the series)

**For the confirmed cases**

Ho: p> 0.05(total cases has a unit root, therefore not stationary)

H1: p< 0.05 (total cases has no unit root, therefore stationary)

## A DICKEY-FULLER (ADF) TEST

**Number of observations = 481**

Test Statistic 1% Critical Value 5% Critical Value 10% Critical Value

Z(t) 0.274 -3.981 -3.421 -3.130

MacKinnon approximate p-value for **Z(t) = 0.9962**

D.confiremedcases Coef. Std. Err. t P>t [95% Conf. Interval]

L1. .0006391 .0023351 0.27 0.784 -.0039493 .0052276

trend -.2762372 .5393895 -0.51 0.609 -1.336105 .7836303

cons 235.6307 28.42612 8.29 0.000 179.7751 291.4863

## Table 4.1: Dicky-fuller test result

In table 4.1, Our trend has a p-value of **0.609**, which is more than **0.05, (0.609>0.05)**, therefore we fail to reject the null hypothesis and say that including the trend is insignificant, it is different than 0, and conclude that the trend does not belong to the series, which indicate that the trend is not significant.

For our constant, the p-value **0.000,** is less than **0.05,** (**0.000>0.05),** we therefore reject the null hypothesis and conclude that the constant belongs to the series. Which indicate it is significant in the series.

For the **dickey-fuller (adf) test** on confirmed cases, the result above indicates that the p-value **0.9962** is greater than **0.05** (**0.9962>0.05**), therefore we can’t reject the null hypothesis, we then conclude that our confirmed cases have a unit root which makes it non-stationary.

To be sure of our Data not being stationary, we performed a second formal test (PPerron test) to confirm the non-Stationarity.

**For the Total cases**

Ho: p> 0.05(total cases has a unit root, therefore not stationary)

H1: p< 0.05 (total cases has no unit root, therefore stationary)

## PHILLIPS-PERRON TEST

**Number of observations = 481**

**Newey-West lags = 5**

Test Statistic 1% Critical Value 5% Critical Value 10% Critical Value

Z(rho) -0.937 -28.862 -21.485 -18.092

Z(t) -0.485 -3.981 -3.421 -3.130

MacKinnon approximate p-value for **Z(t) = 0.9840**

confirmedcases Coef. Std. Err. t P>t [95% Conf. Interval]

L1. 1.000639 .0023351 428.51 0.000 .9960507 1.005228

Trend -.2762372 .5393895 -0.51 0.609 -1.336105 .7836303

Cons 235.6307 28.42612 8.29 0.000 179.7751 291.4863

## Table 4.2: Philips-perron test result

We could see in table 4.2 that the **pperron test** also confirms that our data is not stationary. For the **pperron test** on confirmed cases, the result above indicate that the p-value **0.9840** is greater than **0.05** (**0.9962>0.05**), therefore we can’t reject the null hypothesis, we then conclude that our confirmed cases has a unit root which makes it Non-stationary. Whiles trend is not significant, the constant is significant per our result above.

## Differencing

Since our data is not stationary, we take differences of the data by testing the formal tests (ADF and PPerron) in 1st differences to check if it would be stationary by 1st differences

Ho: p > 0.05(confirmed cases has a unit root, therefore not stationary)

H1: p < 0.05 (confirmed cases has no unit root, therefore stationary)

**DICKEY-FULLER TEST IN 1ST DIFFERENCE**

**Number of observations = 480**

Test Statistic 1% Critical Value 5% Critical Value 10% Critical Value

Z(t) -15.858 -3.442 -2.871 -2.570

MacKinnon approximate p-value for **Z(t) = 0.0000**

D.confirmedcases Coef. Std. Err. t P>t [95% Conf. Interval]

LD. -.6890689 .0434512 -15.86 0.000 -.7744479 -.60369

Cons 139.3675 15.53004 8.97 0.000 108.8519 169.8831

**Table 4.3:** **Dickey-fuller test in 1st difference**

**PHILLIPS-PERRON TEST IN 1ST DIFFERENCE**

**Number of observation = 480**

**Newey-West lags = 5**

Test Statistic 1% Critical Value 5% Critical Value 10% Critical Value

Z(rho) -468.893 -20.484 -14.000 -11.200

Z(t) -17.578 -3.442 -2.871 -2.570

MacKinnon approximate p-value for **Z(t) = 0.0000**

D. confirmedcases Coef. Std. Err. t P>|t| [95% Conf. Interval]

LD. .3109311 .0434512 7.16 0.000 .2255521 .39631

cons 139.3675 15.53004 8.97 0.000 108.8519 169.8831

Table 4.4: **Phillips-perron test in 1st difference.**

our two results shows that our data is stationary at 1st difference, where the p-values (**0.0000<0.05**) and (**0.0000<0.05**) respectively for both tests are significant, we therefore reject the null hypothesis and conclude that, confirmed cases has no unit root, therefore it is stationary.

The constants are also significant, where we record a p-value of **0.000** for both tests less than **0.05, (0.000<0.05).**

Therefore, satisfying the stationarity at 1st difference, we will now be working with an AR(I)MA model.

## Determination of “P” and “Q”

Determining the values of: p , q the below equation is applicable.

## Equation 1

Where Yt is the total covid-19 cases in Ghana, c is the constant, also is the autoregressive component with a “**p**” order and is themoving average component with a “**q**” order and finally the error term

By identification of the model (p,q): check the corrologram,

* Autocorrelation function (ACF)- we determine “q”
* Partial Autocorrelation (PACF)- we determine “P”
* ACF and PACF may suggest diverse possible models

We also consider parsimony: Adding more variables will increase the model fit (R²) at cost of decreasing degrees of freedom.(we don’t want to add needless or insignificant variables).

Below is the Autocorrelation function graph on our differenced data,



## Figure 4.3: ACF of 1st difference of covid-19 confirmed cases

To determine the order of our MA component of “q” we will use all the lags exceeding the confidence band, and with our result (figure 4.3), we have 21 lags exceeding the 95% confidence bands, because of parsimony, we will use just 2 lags out the 21 lags for our MA process, which are lag 1 and lag 8.

Below is the Partial Autocorrelation function graph on our differenced data,

## 

## Figure 4.4: PACF of 1st difference of covid-19 confirmed cases

To determine the order of our AR component of “p” we will use all the lags exceeding the 95% confidence bands, and with our result, we have 7 lags exceeding the 95% confidence bands, because of parsimony, we will use just 2 out the 7 lags for our AR process, which are lag 6 and lag 5 since they are closer to the 95% confidence bands

Therefore, we determine 4 possible ways of estimating our ARIMA model, which are AR=6, 5 and MA=1, 8

ARIMA (6,1,1), ARIMA (5,1,1), ARIMA (6,1,8), ARIMA (5,1,8)

## Estimation of ARIMA model

Estimating all our four (4) possible candidate models, the objective is to find a stationary and parsimonious model that fits the data well. Which model is better?

**Model selection criteria**

* Significance of the ARIMA components
* Compare Akaike and Bayesian criterions (**Smaller is better**)
* Maximum likelihood (**Bigger is better**)
* SigmaSQ: estimate of the error variance (**Smaller is better**)

Model selection criteria

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Model A | **Model B** | Model C | Model D | **Best Model** |
| Criteria | ARIMA (6,1,1) | **ARIMA (6,1,8)** | ARIMA (5,1,1) | ARIMA (5,1,8) |  |
| C, AR and MA | 4/8 | **13/14** | 5/7 | 7/14 | B |
| SigmaSQ | 224.6632 | **216.9659** | 224.8141 | 221.1908 | B |
| Log Likelihood | -3287.7278 | **-3272.803** | -3288.018 | -3280.398 | B |
| Akaike | 6593.456 | **6575.606** | 6592.035 | 6590.797 | B |
| Bayeseian | 6631.039 | **6638.244** | 6625.442 | 6653.435 | C |
| **Best Model** |  |  |  |  | **B** |

**Table 4.5: Model selection criteria**

Based on the above commands performed on our 1st differenced data for model estimation, the above table is the result, and it shows the significant levels of all the 4 models.

The above result in table 4.5 shows that **Model B** met almost all the selection criteria, In Model B, out of the fourteen (14) coefficients of the constant (C), Autoregressive (AR) and the Moving Average (MA), thirteen out of fourteen coefficients were significant (**13/14**). The variance of the error (SigmaSQ= **-3272.803**) was the lowest/smallest in the above result as compared to the rest of the models, the biggest log likelihood in the above result is under **Model B**, the smallest Akaike criterion value is also found in **Model B** whiles the smallest Bayeseian criterion value is found in Model C. therefore the best Model for our forecasting is **Model B**.

## Diagnostic and Forecasting our ARIMA model (6, 1, 8)

We have our “potential” candidate model: ARIMA (6, 1, 8), we therefore have to check if it will be stable for our univariate process.

Requirements for a stable univariate process:

* Residuals of the model are white noise: Portmanteau Test

Null Hypothesis: residuals are white noise.

* Check if the estimated ARMA process is (covariance) stationary, which implies the AR roots should lie inside the unit circle.
* Check if the estimated ARMA process is invertible: all MA roots should lie inside the unit circle.

If the above conditions are satisfied with our selected model, we can then forecast with that model. If not satisfied, we have to repeat the selection and the estimation method (Try another method).

We therefore first predict our error to see if it wiggles around the mean and use the mean to check if our model is white noise.

**Variable Obs Mean Std. Dev. Min Max**

**Error 481 .4601788 217.9228 -806.9655 1074.984**

## Table 4.6: Summarize error

## Figure 4.5: Residual plot

We could see our values wiggle around the mean (**.4601788),**

We again confirm the above result with a formal test (portmanteau test)

Null hypothesis: residuals are white noise if p>0.05

Alternative hypothesis: residuals are not white noise if p<0.05

## Portmanteau test for white noise

**Portmanteau (Q) statistic = 30.4622**

**Prob > chi2(40) = 0.8619**

Our p-value 0.8619 is greater than 0.05 (**0.8619>0.05**), we therefore fail to reject the null hypothesis and conclude that our residuals are white noise.

We check to see if our ARMA process is (covariance) stationary and ARMA process is invertible

Figure 4.6: Inverse roots of ARMA

## 

The result above shows that our estimated ARMA process is (covariance) stationary, which implies the AR roots lie inside the unit circle therefore our AR parameters satisfy stability condition and our estimated ARMA process is invertible: At least one eigenvalue is at least 1.0. Therefore, not all our MA parameters do satisfy invertibility condition. Which indicate all MA roots do not lie inside the unit circle.

## FORECASTING

Forecasting plays an important role in decision making process. It is a planning tool which helps decision makers to foresee the future uncertainty based on the behavior of past and current observations. Forecasting as described by Box and Jenkins (1976), provide the basis for economic and business planning, inventory and production control and optimization of industrial processes. Forecasting is the process of predicting some unknown quantities. From previous studies, most research work has found that the selected model is not necessary the model that provides best forecasting. In this sense, further forecasting accuracy test such as Mean Error (ME), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) must be performed on the model. Table 4.10 summarizes the forecasted values of COVID-19 cases over the period of 9th July 2021 to March 31st 2022 with 95% confidence level using the ARIMA (6 1 8) model which has lower p-value of 0.000 (thus, less than alpha value of 0.05) indicating that it is the best model according to Box-Jenkins model selection criteria.

We forecasted for the next 266 days. The below result shows both original cases from our raw data and the forecasted cases from our differenced data.

## 

## Figure 4.7 Confirmed and predicted covid-19 cases

The blue thick line indicates our confirmed COVID-19 cases (original data), and the red thin line indicates our predicted cases (differenced data) from July 9TH ,2021 to 31ST March 2022.

## FORECASTED VALUES

|  |  |  |
| --- | --- | --- |
| days | Forecasted confirmed cases | Difference in forecasted values |
| 9-Jul-21 | 97119.99 | 148.99 |
| 10-Jul-21 | 97285.72 | 165.73 |
| 11-Jul-21 | 97512.66 | 226.94 |
| 12-Jul-21 | 97745.3 | 232.64 |
| 13-Jul-21 | 97975.23 | 229.93 |
| 14-Jul-21 | 98183.81 | 208.58 |
| 15-Jul-21 | 98424.13 | 240.32 |
| 16-Jul-21 | 98679.63 | 255.5 |
| 17-Jul-21 | 98953.63 | 274 |
| 18-Jul-21 | 99206.98 | 253.35 |
| 19-Jul-21 | 99472.05 | 265.07 |
| 20-Jul-21 | 99745.66 | 273.61 |
| 21-Jul-21 | 100045.8 | 300.14 |
| 22-Jul-21 | 100340 | 294.2 |
| 23-Jul-21 | 100637.1 | 297.1 |
| 24-Jul-21 | 100934.2 | 297.1 |
| 25-Jul-21 | 101251.3 | 317.1 |
| 26-Jul-21 | 101575.5 | 324.2 |
| 27-Jul-21 | 101902.6 | 327.1 |
| 28-Jul-21 | 102228.3 | 325.7 |
| 29-Jul-21 | 102561.7 | 333.4 |
| 30-Jul-21 | 102907.3 | 345.6 |
| 31-Jul-21 | 103257.3 | 350 |
| 1-Aug-21 | 103610.5 | 353.2 |
| 2-Aug-21 | 103962.2 | 351.7 |
| 3-Aug-21 | 104325.2 | 363 |
| 4-Aug-21 | 104691.3 | 366.1 |
| 5-Aug-21 | 105066.1 | 374.8 |
| 6-Aug-21 | 105436.3 | 370.2 |
| 7-Aug-21 | 105814.9 | 378.6 |
| 8-Aug-21 | 106193.4 | 378.5 |
| 9-Aug-21 | 106582.8 | 389.4 |
| 10-Aug-21 | 106968.3 | 385.5 |
| 11-Aug-21 | 107360.2 | 391.9 |
| 12-Aug-21 | 107749.4 | 389.2 |
| 13-Aug-21 | 108147.6 | 398.2 |
| 14-Aug-21 | 108544 | 396.4 |
| 15-Aug-21 | 108945.4 | 401.4 |
| 16-Aug-21 | 109343.6 | 398.2 |
| 17-Aug-21 | 109746.9 | 403.3 |
| 18-Aug-21 | 110149.9 | 403 |
| 19-Aug-21 | 110556 | 406.1 |
| 20-Aug-21 | 110960.4 | 404.4 |
| 21-Aug-21 | 111365.9 | 405.5 |
| 22-Aug-21 | 111771.8 | 405.9 |
| 23-Aug-21 | 112178.4 | 406.6 |
| 24-Aug-21 | 112585.2 | 406.8 |
| 25-Aug-21 | 112990.1 | 404.9 |
| 26-Aug-21 | 113395.8 | 405.7 |
| 27-Aug-21 | 113799.4 | 403.6 |
| 28-Aug-21 | 114204.4 | 405 |
| 29-Aug-21 | 114605.8 | 401.4 |
| 30-Aug-21 | 115008 | 402.2 |
| 31-Aug-21 | 115406 | 398 |
| 1-Sep-21 | 115805.5 | 399.5 |
| 2-Sep-21 | 116200.4 | 394.9 |
| 3-Sep-21 | 116595.7 | 395.3 |
| 4-Sep-21 | 116986 | 390.3 |
| 5-Sep-21 | 117376.6 | 390.6 |
| 6-Sep-21 | 117762.3 | 385.7 |
| 7-Sep-21 | 118147.6 | 385.3 |
| 8-Sep-21 | 118527.8 | 380.2 |
| 9-Sep-21 | 118906.9 | 379.1 |
| 10-Sep-21 | 119281.2 | 374.3 |
| 11-Sep-21 | 119653.7 | 372.5 |
| 12-Sep-21 | 120021.6 | 367.9 |
| 13-Sep-21 | 120386.8 | 365.2 |
| 14-Sep-21 | 120747.7 | 360.9 |
| 15-Sep-21 | 121105.3 | 357.6 |
| 16-Sep-21 | 121458.8 | 353.5 |
| 17-Sep-21 | 121808.3 | 349.5 |
| 18-Sep-21 | 122153.9 | 345.6 |
| 19-Sep-21 | 122494.8 | 340.9 |
| 20-Sep-21 | 122832.3 | 337.5 |
| 21-Sep-21 | 123164.5 | 332.2 |
| 22-Sep-21 | 123493.3 | 328.8 |
| 23-Sep-21 | 123816.5 | 323.2 |
| 24-Sep-21 | 124136.4 | 319.9 |
| 25-Sep-21 | 124450.4 | 314 |
| 26-Sep-21 | 124761.1 | 310.7 |
| 27-Sep-21 | 125065.8 | 304.7 |
| 28-Sep-21 | 125367 | 301.2 |
| 29-Sep-21 | 125662.2 | 295.2 |
| 30-Sep-21 | 125953.9 | 291.7 |
| 1-Oct-21 | 126239.6 | 285.7 |
| 2-Oct-21 | 126521.6 | 282 |
| 3-Oct-21 | 126797.7 | 276.1 |
| 4-Oct-21 | 127069.9 | 272.2 |
| 5-Oct-21 | 127336.5 | 266.6 |
| 6-Oct-21 | 127599 | 262.5 |
| 7-Oct-21 | 127856.1 | 257.1 |
| 8-Oct-21 | 128108.8 | 252.7 |
| 9-Oct-21 | 128356.5 | 247.7 |
| 10-Oct-21 | 128599.6 | 243.1 |
| 11-Oct-21 | 128837.9 | 238.3 |
| 12-Oct-21 | 129071.5 | 233.6 |
| 13-Oct-21 | 129300.7 | 229.2 |
| 14-Oct-21 | 129525 | 224.3 |
| 15-Oct-21 | 129745.2 | 220.2 |
| 16-Oct-21 | 129960.5 | 215.3 |
| 17-Oct-21 | 130171.8 | 211.3 |
| 18-Oct-21 | 130378.3 | 206.5 |
| 19-Oct-21 | 130581.1 | 202.8 |
| 20-Oct-21 | 130779.1 | 198 |
| 21-Oct-21 | 130973.6 | 194.5 |
| 22-Oct-21 | 131163.5 | 189.9 |
| 23-Oct-21 | 131350 | 186.5 |
| 24-Oct-21 | 131532.1 | 182.1 |
| 25-Oct-21 | 131710.9 | 178.8 |
| 26-Oct-21 | 131885.6 | 174.7 |
| 27-Oct-21 | 132057.2 | 171.6 |
| 28-Oct-21 | 132224.8 | 167.6 |
| 29-Oct-21 | 132389.5 | 164.7 |
| 30-Oct-21 | 132550.5 | 161 |
| 31-Oct-21 | 132708.7 | 158.2 |
| 1-Nov-21 | 132863.6 | 154.9 |
| 2-Nov-21 | 133015.7 | 152.1 |
| 3-Nov-21 | 133164.9 | 149.2 |
| 4-Nov-21 | 133311.5 | 146.6 |
| 5-Nov-21 | 133455.4 | 143.9 |
| 6-Nov-21 | 133596.8 | 141.4 |
| 7-Nov-21 | 133736 | 139.2 |
| 8-Nov-21 | 133872.8 | 136.8 |
| 9-Nov-21 | 134007.7 | 134.9 |
| 10-Nov-21 | 134140.3 | 132.6 |
| 11-Nov-21 | 134271.4 | 131.1 |
| 12-Nov-21 | 134400.4 | 129 |
| 13-Nov-21 | 134528.2 | 127.8 |
| 14-Nov-21 | 134654.1 | 125.9 |
| 15-Nov-21 | 134779 | 124.9 |
| 16-Nov-21 | 134902.3 | 123.3 |
| 17-Nov-21 | 135024.8 | 122.5 |
| 18-Nov-21 | 135146 | 121.2 |
| 19-Nov-21 | 135266.7 | 120.7 |
| 20-Nov-21 | 135386.4 | 119.7 |
| 21-Nov-21 | 135505.6 | 119.2 |
| 22-Nov-21 | 135624.2 | 118.6 |
| 23-Nov-21 | 135742.6 | 118.4 |
| 24-Nov-21 | 135860.5 | 117.9 |
| 25-Nov-21 | 135978.5 | 118 |
| 26-Nov-21 | 136096.3 | 117.8 |
| 27-Nov-21 | 136214.3 | 118 |
| 28-Nov-21 | 136332.5 | 118.2 |
| 29-Nov-21 | 136451 | 118.5 |
| 30-Nov-21 | 136570 | 119 |
| 1-Dec-21 | 136689.4 | 119.4 |
| 2-Dec-21 | 136809.6 | 120.2 |
| 3-Dec-21 | 136930.4 | 120.8 |
| 4-Dec-21 | 137052.2 | 121.8 |
| 5-Dec-21 | 137174.9 | 122.7 |
| 6-Dec-21 | 137298.7 | 123.8 |
| 7-Dec-21 | 137423.5 | 124.8 |
| 8-Dec-21 | 137549.8 | 126.3 |
| 9-Dec-21 | 137677.2 | 127.4 |
| 10-Dec-21 | 137806.2 | 129 |
| 11-Dec-21 | 137936.5 | 130.3 |
| 12-Dec-21 | 138068.6 | 132.1 |
| 13-Dec-21 | 138202.2 | 133.6 |
| 14-Dec-21 | 138337.7 | 135.5 |
| 15-Dec-21 | 138474.8 | 137.1 |
| 16-Dec-21 | 138614 | 139.2 |
| 17-Dec-21 | 138755.1 | 141.1 |
| 18-Dec-21 | 138898.2 | 143.1 |
| 19-Dec-21 | 139043.4 | 145.2 |
| 20-Dec-21 | 139190.7 | 147.3 |
| 21-Dec-21 | 139340.2 | 149.5 |
| 22-Dec-21 | 139492 | 151.8 |
| 23-Dec-21 | 139646.1 | 154.1 |
| 24-Dec-21 | 139802.5 | 156.4 |
| 25-Dec-21 | 139961.3 | 158.8 |
| 26-Dec-21 | 140122.5 | 161.2 |
| 27-Dec-21 | 140286.2 | 163.7 |
| 28-Dec-21 | 140452.3 | 166.1 |
| 29-Dec-21 | 140621 | 168.7 |
| 30-Dec-21 | 140792.2 | 171.2 |
| 31-Dec-21 | 140966.1 | 173.9 |
| 1-Jan-22 | 141142.4 | 176.3 |
| 2-Jan-22 | 141321.5 | 179.1 |
| 3-Jan-22 | 141503 | 181.5 |
| 4-Jan-22 | 141687.3 | 184.3 |
| 5-Jan-22 | 141874.2 | 186.9 |
| 6-Jan-22 | 142063.7 | 189.5 |
| 7-Jan-22 | 142255.8 | 192.1 |
| 8-Jan-22 | 142450.6 | 194.8 |
| 9-Jan-22 | 142648 | 197.4 |
| 10-Jan-22 | 142848.1 | 200.1 |
| 11-Jan-22 | 143050.8 | 202.7 |
| 12-Jan-22 | 143256.1 | 205.3 |
| 13-Jan-22 | 143463.9 | 207.8 |
| 14-Jan-22 | 143674.3 | 210.4 |
| 15-Jan-22 | 143887.3 | 213 |
| 16-Jan-22 | 144102.8 | 215.5 |
| 17-Jan-22 | 144320.7 | 217.9 |
| 18-Jan-22 | 144541.1 | 220.4 |
| 19-Jan-22 | 144764 | 222.9 |
| 20-Jan-22 | 144989.3 | 225.3 |
| 21-Jan-22 | 145216.9 | 227.6 |
| 22-Jan-22 | 145446.8 | 229.9 |
| 23-Jan-22 | 145679 | 232.2 |
| 24-Jan-22 | 145913.5 | 234.5 |
| 25-Jan-22 | 146150.1 | 236.6 |
| 26-Jan-22 | 146388.9 | 238.8 |
| 27-Jan-22 | 146629.8 | 240.9 |
| 28-Jan-22 | 146872.7 | 242.9 |
| 29-Jan-22 | 147117.6 | 244.9 |
| 30-Jan-22 | 147364.5 | 246.9 |
| 31-Jan-22 | 147613.3 | 248.8 |
| 1-Feb-22 | 147863.8 | 250.5 |
| 2-Feb-22 | 148116.2 | 252.4 |
| 3-Feb-22 | 148370.2 | 254 |
| 4-Feb-22 | 148625.9 | 255.7 |
| 5-Feb-22 | 148883.2 | 257.3 |
| 6-Feb-22 | 149142 | 258.8 |
| 7-Feb-22 | 149402.3 | 260.3 |
| 8-Feb-22 | 149664 | 261.7 |
| 9-Feb-22 | 149927 | 263 |
| 10-Feb-22 | 150191.3 | 264.3 |
| 11-Feb-22 | 150456.7 | 265.4 |
| 12-Feb-22 | 150723.3 | 266.6 |
| 13-Feb-22 | 150991 | 267.7 |
| 14-Feb-22 | 151259.7 | 268.7 |
| 15-Feb-22 | 151529.3 | 269.6 |
| 16-Feb-22 | 151799.7 | 270.4 |
| 17-Feb-22 | 152071 | 271.3 |
| 18-Feb-22 | 152342.9 | 271.9 |
| 19-Feb-22 | 152615.6 | 272.7 |
| 20-Feb-22 | 152888.8 | 273.2 |
| 21-Feb-22 | 153162.5 | 273.7 |
| 22-Feb-22 | 153436.7 | 274.2 |
| 23-Feb-22 | 153711.3 | 274.6 |
| 24-Feb-22 | 153986.2 | 274.9 |
| 25-Feb-22 | 154261.3 | 275.1 |
| 26-Feb-22 | 154536.7 | 275.4 |
| 27-Feb-22 | 154812.1 | 275.4 |
| 28-Feb-22 | 155087.6 | 275.5 |
| 1-Mar-22 | 155363.1 | 275.5 |
| 2-Mar-22 | 155638.5 | 275.4 |
| 3-Mar-22 | 155913.8 | 275.3 |
| 4-Mar-22 | 156188.9 | 275.1 |
| 5-Mar-22 | 156463.8 | 274.9 |
| 6-Mar-22 | 156738.3 | 274.5 |
| 7-Mar-22 | 157012.5 | 274.2 |
| 8-Mar-22 | 157286.2 | 273.7 |
| 9-Mar-22 | 157559.4 | 273.2 |
| 10-Mar-22 | 157832.1 | 272.7 |
| 11-Mar-22 | 158104.2 | 272.1 |
| 12-Mar-22 | 158375.7 | 271.5 |
| 13-Mar-22 | 158646.5 | 270.8 |
| 14-Mar-22 | 158916.5 | 270 |
| 15-Mar-22 | 159185.7 | 269.2 |
| 16-Mar-22 | 159454.1 | 268.4 |
| 17-Mar-22 | 159721.6 | 267.5 |
| 18-Mar-22 | 159988.2 | 266.6 |
| 19-Mar-22 | 160253.8 | 265.6 |
| 20-Mar-22 | 160518.5 | 264.7 |
| 21-Mar-22 | 160782 | 263.5 |
| 22-Mar-22 | 161044.5 | 262.5 |
| 23-Mar-22 | 161305.9 | 261.4 |
| 24-Mar-22 | 161566.1 | 260.2 |
| 25-Mar-22 | 161825.2 | 259.1 |
| 26-Mar-22 | 162083 | 257.8 |
| 27-Mar-22 | 162339.6 | 256.6 |
| 28-Mar-22 | 162594.9 | 255.3 |
| 29-Mar-22 | 162849 | 254.1 |
| 30-Mar-22 | 163101.7 | 252.7 |
| 31-Mar-22 | 163353.1 | 251.4 |
|  |  | **66,382.1** |

By applying ARIMA (6 1 8) model to forecast confirmed cases. It is found that the predicted case has a consistent decrease after 19-OCT-21 to 9-JAN-22. Which in general indicate a rise and fall of cases in the country.

The study also found that, the country would experience an additional case of 66,382 with in the forecasted period adding to the existing cases of 96,971 as at July 8th 2021. This will help the government and doctors prepare for the forthcoming of the virus. Based on the period predictions, these methods can also be used to forecast the mortality rate for a long period.

Per the above results, we also analyzed that health facilities/ hospitals will be overwhelmed with the higher rise in confirmed cases and it is going to affect healthcare delivery across the nation, since the country lacks isolation centers and that would lead to more of community spread within the country.

## CORRELATION RESULTS

**Command**

(observations=482)

Totaldeath confirmedcases recoveredcases

totaldeath 1.0000

confirmedcases 0.9830 1.0000

recoveredcases 0.9824 0.9981 1.0000

**Table 4.7 Correlation result**

the above result indicates that there is a strong positive relationship between all the three variables, which implies, as our death toll rises up the country turns to experience more confirmed cases and more recoveries likewise suggest that confirmed cases and recovered cases are positively related.

## CHAPTER FIVE

## CONCLUSION AND RECOMMENDATIONS

## 5.1 Introduction

This chapter discusses the study's findings as well as some recommendations for policymakers in the country. .

## Conclusion

The objective of this research was to develop a time series model and forecast COVID-19 cases for the next 266 days (09 July, 2021 to 31 March 2022). Data from March 14, 2020 to July 08, 2021 were collated from DHIMS. The preliminary analysis revealed that our data was non-stationary with a positive increasing trend from 14th march 2020 to 8th July 2021, with the Country’s cases going upward. Several time series models including AR, MA, and Autoregressive Integrated Moving Average (ARIMA) were used in modelling the data in the Stata 13.1. The study identified 4 ‘candidate ‘models which best fitted the data ARIMA (5,1,1), ARIMA (5,1,8), ARIMA (6,1,1) and ARIMA (6,1,8). However, with the use of the Modified Box-Jenkins (Significance of the ARIMA components Compare, Akaike and Bayesian criterions (**Smaller is better**), Maximum likelihood (**Bigger is better**), SigmaSQ: estimate of the error variance (**Smaller is better**)) model selection statistic criteria of the lowest p-value, our best-fitted ARIMA models tend to be ARIMA (6, 1, 8). ARIMA (6, 1, 8) was selected and evaluated. After the estimation of the parameters of selected model, a series of diagnostic and forecasting accuracy tests were performed. Having satisfied the Box-Jenkins model selection criteria, ARIMA (6, 1, 8) model was adjudged to be the best and most plausible model for forecasting our COVID-19 total cases in the Country. With reference to the findings of the research, it can be concluded that:

1. The most adequate model for the data was ARIMA (6, 1, 8)

2. In the next 266 days, there will be an increase in COVID-19 cases in the country, as anticipated values fell in the vicinity of 97,120 cases on July 9, 2021, and 163,353 cases in March, 2022.

## Recommendations

The following recommendations were provided based on the findings of the study:

1. The Center for Disease Control should adopt the ARIMA (6, 1, 8) model in their disease control department (DCD) when planning activities to combat the spread of covid-19.
2. In terms of regions and districts, vaccination should be prioritized in covid-19-prone areas.

3. In order to adequately prepare for the country's continued overwhelming covid-19 cases, the government/authorities should rely on anticipated data (forecasted figures) in their planning operations.

4. In order to deliver quality health care to the citizens, the government must continue to assist health facilities in terms of personnel protection equipment (PPEs) and logistics.

5. The government must look into the causes of high cases in high-case regions and launch aggressive health and environmental initiatives in the areas where they occur.

## REFERENCES

Vikas Chaurasia (august 2020) ARIMA and Regression Model‑Based Worldwide Death Cases Predictions, pp 32-34.

JAPHETH ARTHUR (may 2013) application of statistical models to outpatient department (opd) attendance data in saltpond municipal hospital of the central region of Ghana, Pp 8-11.

Nalini Chintalapudi a , Gopi Battineni a, \*, Francesco Amenta a (April 2020) COVID-19 virus outbreak forecasting of registered and recovered cases after sixty day lockdown in Italy: A data driven model approach, Pp 1-7.

Gaurav Pandeya , Poonam Chaudharya , Rajan Guptab , Saibal Pal SEIR and Regression Model based COVID-19 outbreak predictions in India.

World Health Organization, 2020. Coronavirus disease 2019 (COVID-19): situation report, 67.

https://www.ghs.gov.gh/covid19/dashboardm.php