**Module Title:** Predictive Analysis for Decision Making

**Module Code:** 7FNCE044W

**Course:** MSc Fintech and Business Analytics, Semester 2, 2023/2024

**Empirical report for forecasting Indian GDP using ARIMA Model**

# Abstract

The Gross Domestic Product (GDP) is the total worth of all goods and services produced within a country's borders in a given year. This Empirical Report has constructed an appropriate Autoregressive-Integrated Moving-Average (ARIMA) model for the Indian GDP data using the Box-Jenkins technique. Data on India's annual GDP from 1972 to 2022 was sourced from the World Bank. We conclude that ARIMA (1, 2, 1) is the suitable statistical model for the GDP of India. Lastly, we projected India's GDP for the ensuing ten years using the fitted ARIMA model.

**Keywords**: Box-Jenkins approach, India, Forecasting, Goodness-of-fit measures, Gross domestic product, Residuals analysis

Table of Contents

[Abstract 2](#_Toc165932273)

[Introduction 4](#_Toc165932274)

[Theory 4](#_Toc165932275)

[Autoregressive (AR) model 4](#_Toc165932276)

[Autoregressive moving average (ARMA) model 5](#_Toc165932277)

[ARIMA Models 5](#_Toc165932278)

[Box-Jenkins Approach 5](#_Toc165932279)

[The Data 7](#_Toc165932280)

[A Proposed Indian GDP Model 8](#_Toc165932281)

[Forecasting Results 14](#_Toc165932282)

[Interpretation and conclusion 15](#_Toc165932283)

[References 17](#_Toc165932284)

# Introduction

The market worth of all commodities and services generated by the economy during the time under consideration, including paid-in construction expenses, private inventories, government purchases, personal consumption, and the international trade balance (exports minus imports), is represented by the GDP. There are three approaches to measuring GDP: the expenditure approach, the production approach, and the income approach.

Of the macroeconomic variables, GDP has emerged as the one that concerns people the most. Data on GDP is considered to be a crucial indicator for evaluating both the macroeconomic situation overall and the state of national economic development (Ning et al, 2010). It is frequently regarded as the most accurate indicator of the state of the economy. Furthermore, it serves as a crucial foundation for the government's economic growth plans and initiatives.

For this reason, a precise GDP forecast is essential to gaining a clear understanding of an economy's future trajectory.

In order to estimate and forecast the Indian GDP, we shall employ time series analytic statistical techniques in this study. ARIMA models are among these methods that are most frequently used. Numerous research, including Bhuiyan et al. (2008), Ning et al. (2010), Dritsaki (2015), Wabomba et al. (2016), Maity and Chatteriee (2012), and Uwimana et al. (2018), have employed these models to investigate the GDP in various nations.

The structure of the Emperical Report is as follows. The statistical foundational theory for univariate time series analysis is provided in Section 2. Further in Section 3 we have discussed the sources of the data, the sample size and frequency of the data, definitions of the variables and decribed the main feature of data with the graphs. In Section 4, we have presented our suggested Box-Jenkins model for the Indian GDP. Lastly, the study's summary and closing thoughts are provided in section 5.

# Theory

For a sufficiently enough amount of data on the relevant variables, the time series analysis can produce very accurate short-term forecasts (see Granger and Newbold, 1986). The ARIMA models are popular and adaptable in univariate time series analysis. Three processes are combined to form the ARIMA model: the Moving-Average (MA) process, the Differencing process, and the Autoregressive (AR) process. These procedures, which are widely employed in numerous applications, are referred to as primary univariate time series models in statistical literature.

## Autoregressive (AR) model

AR (p), an autoregressive model of order p, can be written as follows:

**𝑋𝑡 = 𝑐 + 𝛼1𝑋𝑡−1 + 𝛼2𝑋𝑡−2 + ⋯ + 𝛼𝑝𝑋𝑡−𝑝 + 𝜀𝑡 ; 𝑡 = 1,2, … 𝑇, (1)**

where 𝜀𝑡 is the equation's error term; 𝜀𝑡 is a white noise process, which is a series of independently distributed (iid) random variables with E(𝜀𝑡) = 0 and 𝑣𝑎𝑟(𝜀𝑡) = 𝜎2; that is, 𝜀𝑡 ~𝑖𝑖𝑑 𝑁(0, 𝜎2). This model is a long-term memory model since all previous values can have cumulative effects on this level 𝑋𝑡 and so on.

* 1. Moving-average (MA) model

A moving-average process of order q, or MA (q), is said to be a time series {𝑋𝑡} if:

**𝑋𝑡 = 𝜀𝑡 − 𝜃1𝜀𝑡−1 − 𝜃2𝜀𝑡−2 − ⋯ − 𝜃𝑞𝜀𝑡−𝑞. (2)**

The explanatory variables in this model are represented in terms of previous errors.  
Consequently, only q faults will affect 𝑋𝑡; higher order errors will not affect 𝑋𝑡, indicating that this is a short memory model.

## Autoregressive moving average (ARMA) model

A time series {𝑋𝑡} is said to follow an autoregressive moving-average process of order p and q, or ARMA (p, q). if:

**𝑋𝑡 = 𝑐 + 𝛼1𝑋𝑡−1 + ⋯ + 𝛼𝑝𝑋𝑡−𝑝 + 𝜀𝑡 − 𝜃1𝜀𝑡−1 − ⋯ − 𝜃𝑞𝜀𝑡−𝑞. (3)**

This model may combine elements of the previous MA and AR models.

## ARIMA Models

The ARMA models may be further extended to non-stationary series to generate ARIMA models by allowing the data series to be differencing. The moving-average order (q), the degree of differencing (d), and the autoregressive order (p) make up the three parameters of the general non-seasonal model, ARIMA (p, d, q). For example, if 𝑋𝑡 is a non-stationary series, the ARIMA(p, 1, q) model would be as follows after taking a first-difference of 𝑋𝑡 to make ∆𝑋𝑡 stationary:

**∆𝑋𝑡 = 𝑐 + 𝛼1∆𝑋𝑡−1 + ⋯ + 𝛼𝑝∆𝑋𝑡−𝑝 + 𝜀𝑡 − 𝜃1𝜀𝑡−1 − ⋯ − 𝜃𝑞𝜀𝑡−𝑞 (4)**

in which ∆𝑋𝑡 = 𝑋𝑡 − 𝑋𝑡−1. However, the model turns into an ARIMA (0, 1, 0) random walk model if p = 0 and q = 0 in equation (4).

## Box-Jenkins Approach

The statisticians George Box and Gwilym Jenkins are honoured by the Box-Jenkins (1970) technique of time series analysis, which use ARIMA models to identify which time series model best matches the historical data from the time series. For additional information regarding the Box-Jenkins time series analysis, refer to Chatfield (2016), Frain (1992), Kirchgässner et al. (2013), Young (1977), and Young (1977). The four iterative phases of modelling using this method are depicted in Figure 1.

A diagram of a model

Description automatically generated

*Figure 1: The Box-Jenkins iterative approach's phases*

The Box-Jenkins iterative approach's four steps of modelling are as follows:

* Model identification: By examining the plots of the series' Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF), one may confirm that the variables are stationary, identify seasonality in the series, and decide which moving average or autoregressive component to use in the model.
* Model estimation: using computational techniques to determine which coefficients are most appropriate for the selected ARIMA model. The most popular techniques make use of non-linear least-squares estimation or maximum likelihood estimation (MLE).
* Model checking: by assessing if the calculated model complies with the specifications of a stationary univariate process. Misspecification can be found by plotting the residuals' ACF and PACF; in particular, the residuals should be independent of one another and have a consistent mean and variance over time. If the estimation proves to be inadequate, we must return to step one and attempt to build a better model. The estimated model needs to be contrasted with other ARIMA models in order to determine which one best fits the data. The Bayesian Information Criterion (BIC) and Akaike's Information Criterion (AIC), which are defined as follows, are the two widely used model selection criteria.

**𝐴𝐼𝐶 = 2𝑚 − 2 ln(𝐿̂), 𝐵𝐼𝐶 = 𝑙𝑛(𝑛)𝑚 − 2 𝑙𝑛(𝐿̂) (5)**

whereby n denotes the sample size (number of observations), 𝑚 is the number of parameters the model estimates, and 𝐿̂ is the maximum value of the likelihood function for the model. The conventional criteria, Mean Squared Error (MSE), is typically used in combination with AIC and BIC in practical applications.

* Forecasting: When the selected ARIMA model meets the conditions of a stationary univariate process, we may use it for forecasting.

# The Data

The World-Bank provided the annual GDP of India for this study from 1972 to 2022. This indicates that our 51 GDP observations meet the requirement of having more than 50 observations in order to use the Box-Jenkins method for time series forecasting (Chatfield, 2016). We will utilise this data to determine the suitable ARIMA model and then anticipate the Indian GDP for the ensuing 10 years, from 2023 to 2033. R Programming was used to analyse the data collected.

A screenshot of a data

Description automatically generated

*Figure 2: Descriptive Statistics of Indian GDP Data*

The descriptive statistics from Figure 2 Shows:

**Minimum (Min.):** The smallest GDP value in the dataset is approximately $71.46 billion (7.146×10107.146×1010).

**First Quartile (1st Qu.):** The first quartile, representing the 25th percentile, is approximately $225.4 billion (2.254×10112.254×1011), demonstrating that a quarter of the GDP figures in the sample fall short of this threshold.

**Median:** The median GDP value is approximately $415.9 billion (4.159×10114.159×1011), which is the middle value of the dataset when ordered from smallest to largest.

**Mean:** The average GDP value across all observations is approximately $904.3 billion (9.043×10119.043×1011).

**Third Quartile (3rd Qu.):** The third quartile, representing the 75th percentile, is approximately $1.509 trillion (1.509×10121.509×1012), indicating that 75% of the GDP values are below this amount.

**Maximum (Max.):** The largest GDP value in the dataset is approximately $3.417 trillion (3.417×10123.417×1012).

A graph showing the growth of years

Description automatically generated

*Figure 3: Indian GDP over the years since 1972*

GDP data over the time, with the x-axis representing the number of years since the starting year and the y-axis representing the GDP values. The graph shows a steady increase in GDP over the 50-year period, with the value rising significantly towards the end of the period, reaching around $3.0 trillion.

# A Proposed Indian GDP Model

ARIMA model was chosen with the Box-Jenkins approach to forecast the GDP as they have ability to model various patterns like trends and seasonality. They are flexible, handling non-stationary data through differencing, making them suitable for many applications, including finance and economics. ARIMA models are relatively simple to implement and interpret, requiring only historical data without the need for external variables. They provide confidence intervals and error measures, aiding in the assessment of forecast accuracy.

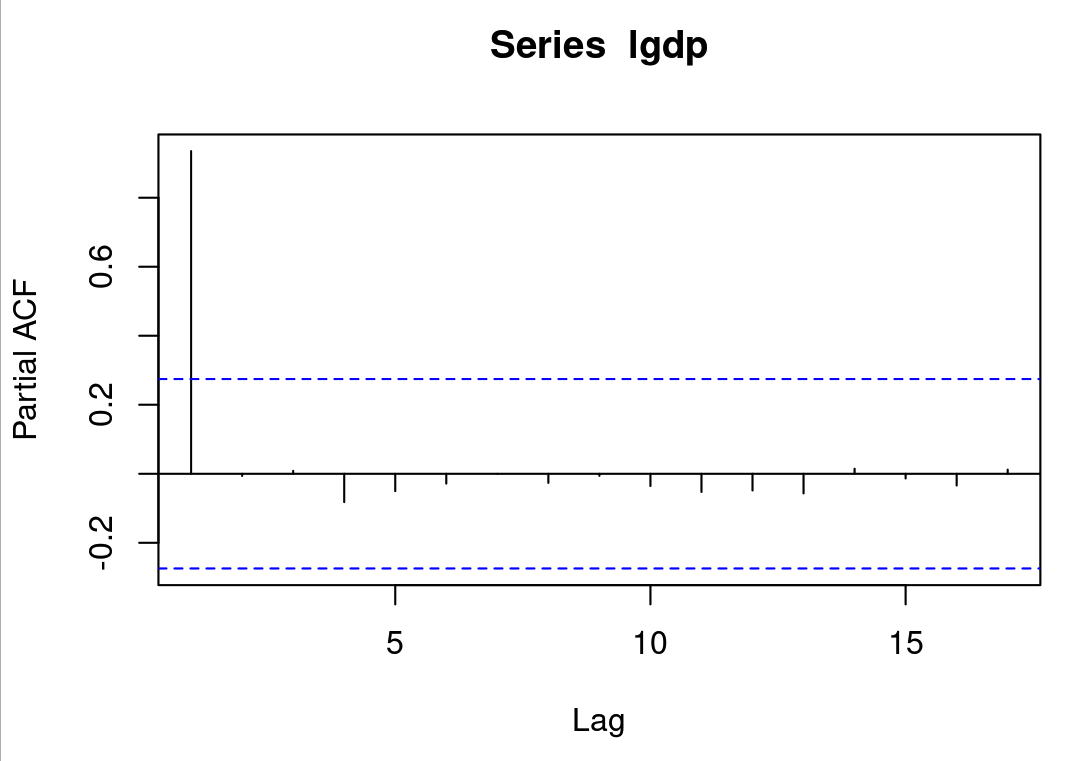
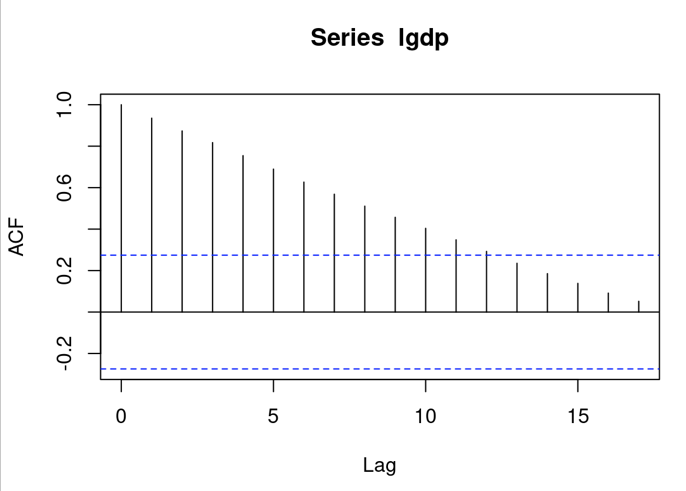
The data was first analysed using time plots of the series, as shown in Figures 2 and 3. The time graphs in Figure 2 may be used to visually examine the fact that the Indian GDP is not a stationary series. Furthermore, the non-stationary behaviour of the series is supported by the ACF and PACF plots of the series, as seen in Figure 3.

It is also contrasted with the 𝜏3 critical values. The null hypothesis of a unit root cannot be rejected at conventional significance levels since -1.7496 is bigger than -3.15 (the crucial value at 10% significance level). This suggests that the series may be non-stationary. As with the development of the ARIMA model, we will use the differencing to arrive at stationary.

A graph showing the growth of a stock market

Description automatically generated

*Figure 4: lgdp Plot.*



*Figure 5: ACF and PACF plots for lgdp.*

A screenshot of a computer code

Description automatically generated

A visual analysis of Figure 4 verifies that the seasonal trend in the Indian GDP series can frequently be achieved by logarithmic processing. As a result, the exponential trend will ultimately transform into a linear trend. The data must be adjusted to obtain stationary before proceeding with additional analysis using the Box-Jenkins method.

In order to achieve stationary in the second instant, by calculating the difference between the natural logarithms of the series values, the series was modified. The transformation is represented by the following equation:

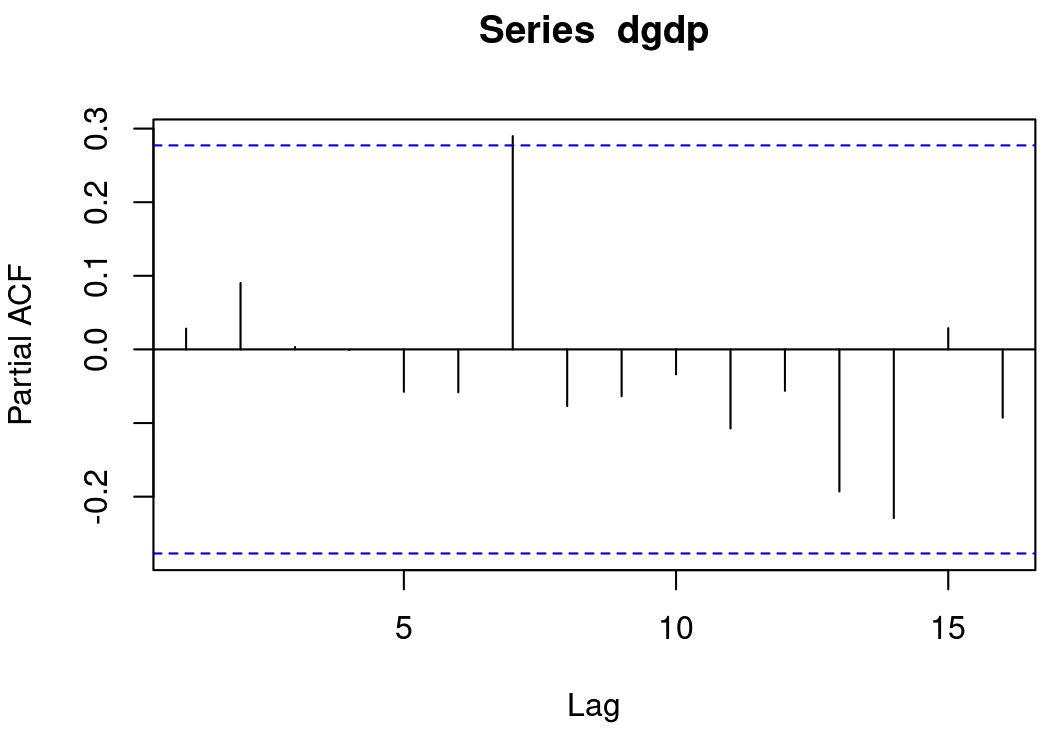
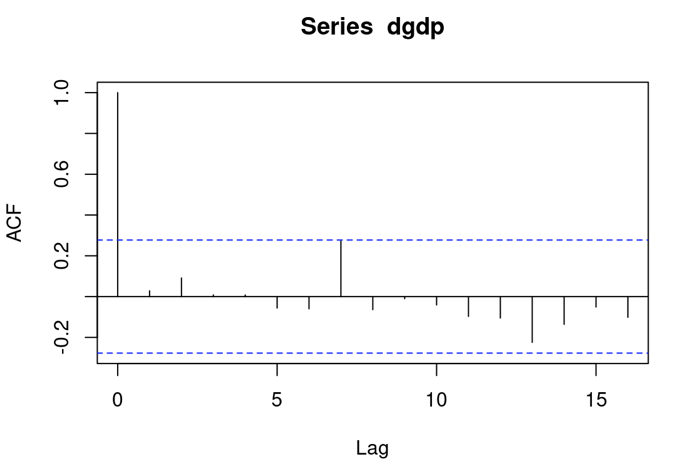
**𝑋𝑡 = ln GDP𝑡 − ln GDP𝑡−1 (6)**

where ln GDP𝑡 = log𝑒 (GDP𝑡).

A graph with lines and numbers

Description automatically generated

*Figure 6: first difference(dgdp) Plot*



*Figure 7: ACF and PACF plots for first difference(dgdp).*

Figure 6 illustrates that 𝑋𝑡 is stationary and lacks a trend following the transformation.In the ARIMA model, d = 1. We may therefore conclude that the model after first difference is stationary by looking at how the ACF and PACF of the differenced GDP data appear, comparing the two plots in Figure 7, and further looking at the findings of the ADF test results below for first difference (dgdp).

A screenshot of a test result

Description automatically generated

ARIMA model is fitted the data of lgdp where we found out that the model which fit the best is ARIMA(0,1,0). The ARIMA(0,1,0) model with drift indicates a simple differencing model with a constant term (drift). Here's a breakdown of the model components and their implications:

**Model Specification: ARIMA(0,1,0) with drift**

* **ARIMA(0,1,0):** The model appears to be a straightforward differencing model without any moving average (MA) or autoregressive (AR) components, based on its specification. The centre '1' denotes that stationarity has been achieved by a single difference of the series.
* **With drift:** The inclusion of a drift component implies that there is a constant linear trend in the differenced series.

**Coefficients:**

* **Drift:** 0.0773
  + This coefficient indicates the average change in the differenced series per time period. A positive drift suggests a general upward trend in the original series.
* **Standard Error (s.e.) of Drift:** 0.0109
  + The standard error measures the variability of the drift estimate. A smaller standard error relative to the drift value suggests that the estimate is precise.

**Model Fit Statistics:**

* **variance (𝜎2):** 0.006067
  + This figure shows the variance of the model's residuals, or mistakes. A lower sigma-squared value suggests that the model fits the data more well.
* **Log Likelihood:** 57.23
  + Higher values indicate a better model fit. The log likelihood gauges the probability of the model given the data.
* **Akaike Information Criterion (AIC):** -110.47
* **Corrected Akaike Information Criterion (AICc): -**110.21
* **Bayesian Information Criterion (BIC**): -106.64
  + AIC, AICc, and BIC are criteria for model selection among a set of models. Lower values generally indicate a better model in terms of explaining the variation in the data with fewer parameters.

Based on the ARIMA(0,1,0) model with drift for the series "lgdp" and the coefficients derieved, the estimated regression equation for this model can be written as follows:

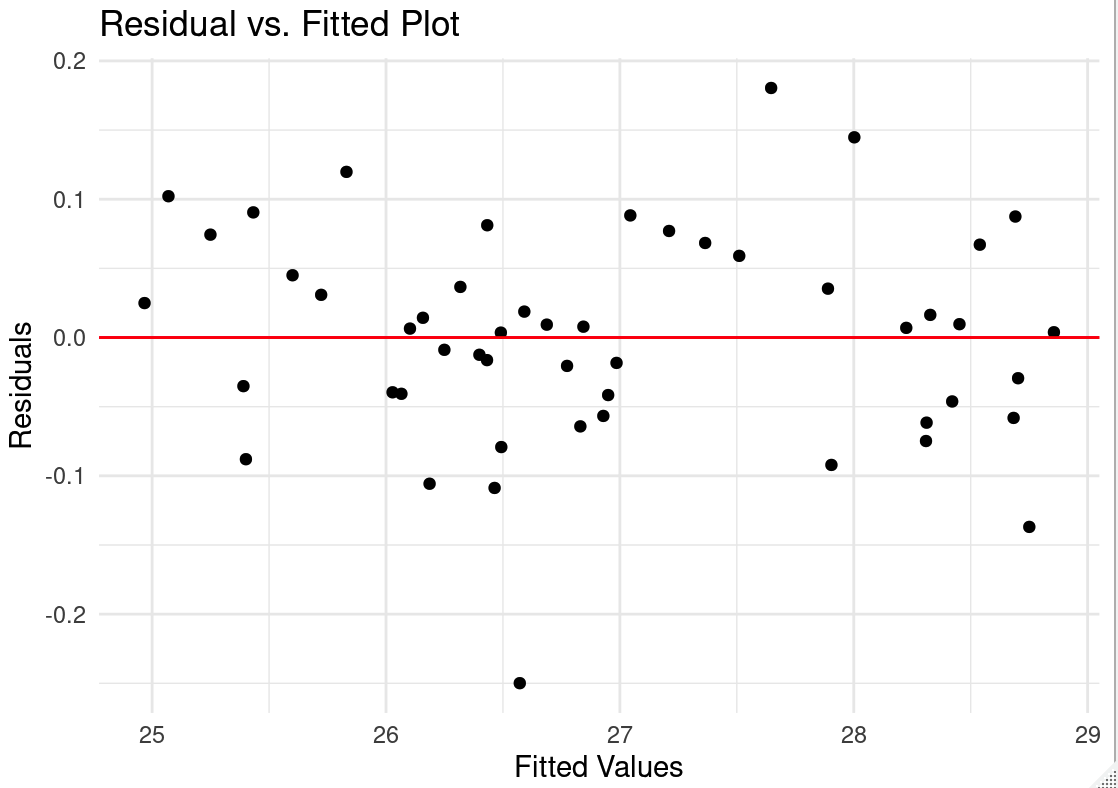
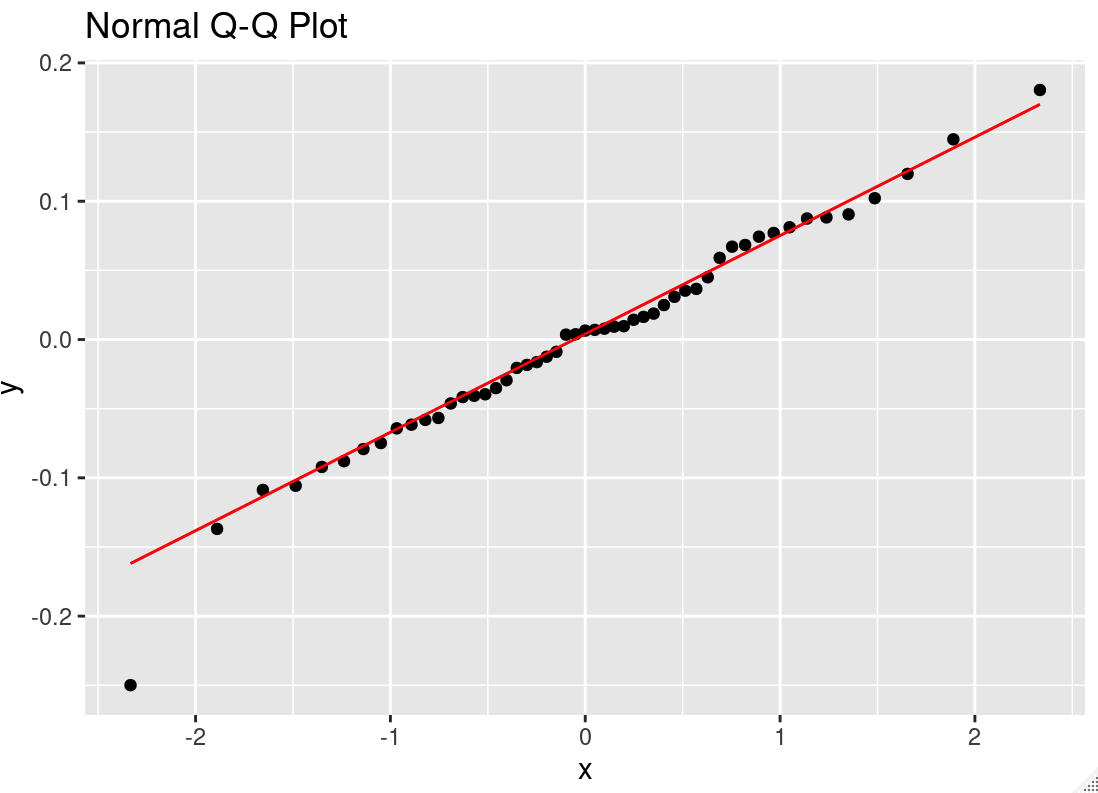
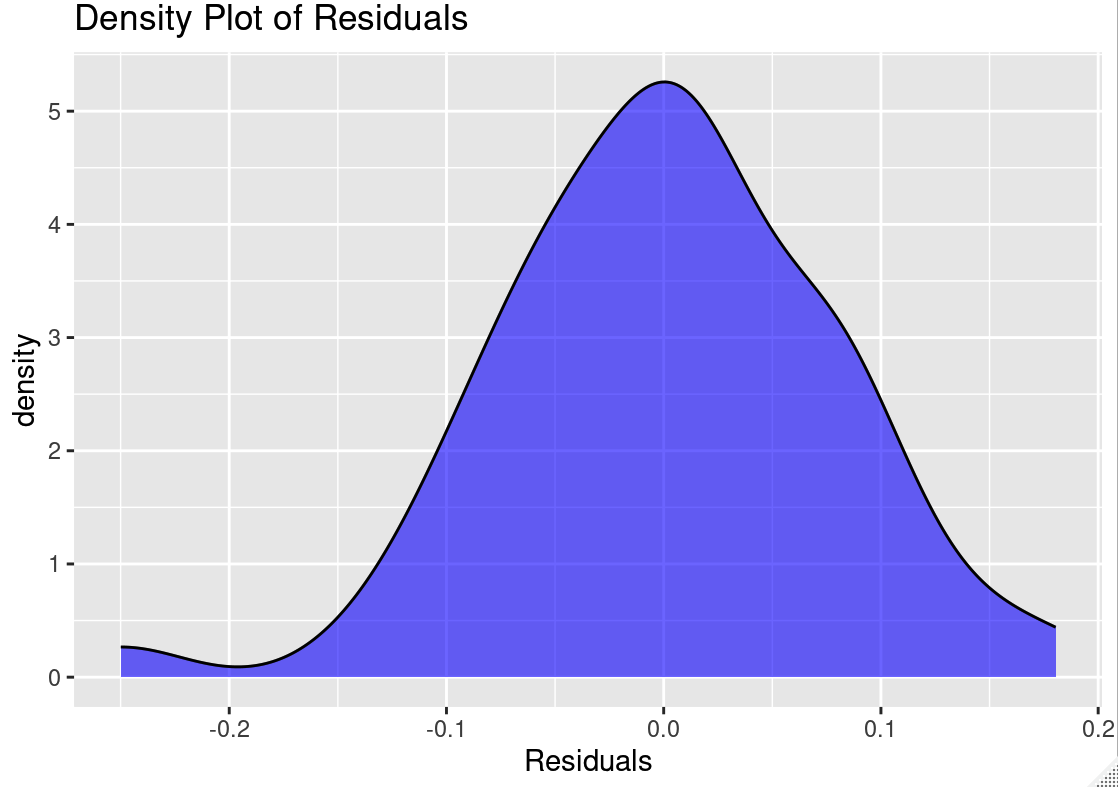
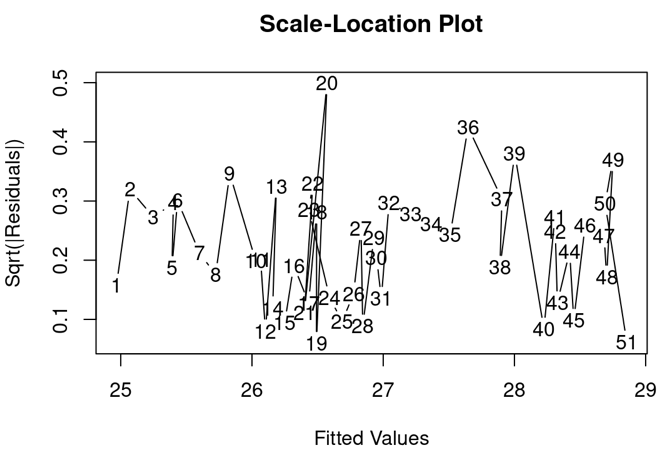
**𝑋𝑡 = 0.0773+ 𝑋𝑡−1 + 𝜀𝑡 (7)**

Here, 𝑋𝑡 represents the forecasted value of the series at time 𝑡, 𝑌𝑡−1is the actual observed value of the series at time 𝑡−1, and 0.0773 is the drift coefficient. This equation indicates that the forecast for the next period is the value of the series in the previous period plus a constant drift term, which represents the average period-to-period change in the series "lgdp".

A screenshot of a computer program

Description automatically generated

**Residual Analysis:**



*Figure 8: Residual Analysis Plots*

The residual analysis of an ARIMA(0,1,0) model, as depicted in the Figure 8, includes four diagnostic plots that are used to evaluate the adequacy of the model. These plots are essential for checking the assumptions of the model, such as the normality of residuals, the absence of patterns in the residuals, and the constant variance of residuals across the range of fitted values.

the residual analysis of the ARIMA(0,1,0) model as shown in the image suggests that the model fits the data reasonably well, with no apparent violations of the assumptions of constant variance, no autocorrelation, and residuals that are approximately normally distributed.

# Forecasting Results

Equation (7) can be used to anticipate GDP values for India for the next 10 years(from 2023 to 2033) because the ARIMA (0, 1, 0) model fits the GDP data. Table 3 presents the GDP prediction values. Figure 9 displays the trend of the actual and forecasted lnGDP values together with their 95% confidence intervals.

The estimated numbers imply that India's GDP will continue to rise. Keep in mind that this result is only an expected value and that the national economy is a dynamic, complex system. To mitigate the risk of severe fluctuations in the economy, we should maintain microeconomic regulation and control for stability and continuity, be cognizant of the risk of adjustment in economic operations, and adjust the corresponding target value based on the circumstances (Wabomba et al., 2016).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Time Point | lnGDP Forecast | GDP Forecast (USD) | 95% Confidence Interval Lower Bound (USD) | 95% Confidence Interval Upper Bound (USD) |
| 52 | 28.93703 | 3.69E+12 | 3.17E+12 | 4.30E+12 |
| 53 | 29.01437 | 3.99E+12 | 3.21E+12 | 4.95E+12 |
| 54 | 29.09171 | 4.31E+12 | 3.31E+12 | 5.61E+12 |
| 55 | 29.16906 | 4.66E+12 | 3.43E+12 | 6.32E+12 |
| 56 | 29.2464 | 5.03E+12 | 3.58E+12 | 7.08E+12 |
| 57 | 29.32375 | 5.43E+12 | 3.74E+12 | 7.90E+12 |
| 58 | 29.40109 | 5.87E+12 | 3.92E+12 | 8.79E+12 |
| 59 | 29.47844 | 6.34E+12 | 4.12E+12 | 9.77E+12 |
| 60 | 29.55578 | 6.85E+12 | 4.34E+12 | 1.08E+13 |
| 61 | 29.63313 | 7.40E+12 | 4.57E+12 | 1.20E+13 |

*Table 1: Indian GDP Forecast*

A graph showing a line

Description automatically generated with medium confidence

*Figure 7: Indian GDP Forecast Plot*

# Interpretation and conclusion

The ARIMA(0,1,0) model with drift effectively captures the underlying trend in the lgdp series. The positive drift coefficient, along with a relatively low standard error, suggests a consistent upward trend in the series after differencing. The low sigma-squared value and high log likelihood further affirm the model's adequacy in fitting the data. The negative values in AIC, AICc, and BIC indicate a good model fit relative to other models with more parameters or less fit.

The report forecasts the Indian GDP for the next ten years, from 2023 to 2033, using the selected ARIMA model. The forecasts suggest a continuing growth trend in India's GDP

The provided forecasts and their confidence intervals are crucial for planning and decision-making processes, especially in economic policy and investment. The increasing trend and widening confidence intervals highlight both the expected economic growth and the associated forecast uncertainty, respectively.

**Limitations:**

This model is useful for forecasting future values of Indian GDP, assuming the trend continues similarly. It is noteworthy to acknowledge that this model does not incorporate any consideration for potential cyclicality or other types of autocorrelation that may exist within the data.

The ARIMA model, while robust for short-term forecasting and handling non-stationary data, does not account for potential structural changes in the economy or external shocks that could significantly alter the growth trajectory.

The model's predictions are based solely on historical GDP data and do not incorporate other potentially influential variables such as political changes, global economic conditions, or technological advancements.

# References

[1] M. N. A. Bhuiyan, K. S. Ahmed and R. Jahan, Study on Modeling and Forecasting of the GDP of Manufacturing Industries in Bangladesh, Chiang Mai University Journal of Social Science and Humanities, 2 (2008), 143- 157.

[2] G. E. P. Box, G. M. Jenkins, Times Series Analysis Forecasting and Control, Holden-Day San Francisco, 1970.

[3] C. Chatfield, The Analysis of Time Series: An Introduction. CRC Press, 2016.

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

[5] J. Frain, Lecture Notes on Univariate Time Series Analysis and Box Jenkins Forecasting, Economic Analysis, Research and Publications, 1992.

[6] C. W. Granger, P. Newbold, Forecasting Economic Time Series, Academic Press, 1986.

[7] C. M. Jarque, A. K. Bera, Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals, Economics Letters, 6 (1980), 255-259. <https://doi.org/10.1016/0165-1765(80)90024-5>

[8] G. Kirchgässner, J. Wolters and U. Hassler, Univariate Stationary Processes, in Introduction to Modern Time Series Analysis, Springer, Berlin, Heidelberg, 2013, 27-93. <https://doi.org/10.1007/978-3-642-33436-8_2>

[9] B. Maity, B. Chatterjee, Forecasting GDP Growth Rates of India: An Empirical Study, International Journal of Economics and Management Sciences, 1 (2012), 52-58.

[10] W. Ning, B. Kuan-jiang and Y. Zhi-fa, Analysis and Forecast of Shaanxi GDP Based on the ARIMA Model, Asian Agricultural Research, 2 (2010), 34-41.

[11] A. Uwimana, B. Xiuchun and Z. Shuguang, Modeling and Forecasting Africa’s GDP with Time Series Models, International Journal of Scientific and Research Publications, 8 (2018), 41-46. https://doi.org/10.29322/ijsrp.8.4.2018.p7608

[12] M. S. Wabomba, M. P. Mutwiri and M. Fredrick, Modeling and Forecasting Kenyan GDP Using Autoregressive Integrated Moving Average (ARIMA) Models, Science Journal of Applied Mathematics and Statistics, 4 (2016), 64-73. <https://doi.org/10.11648/j.sjams.20160402.18>

[13] W. L. Young, The Box-Jenkins Approach to Time Series Analysis and Forecasting: Principles and Applications, RAIRO-Operations ResearchRecherche Opérationnelle, 11 (1977), 129-143. <https://doi.org/10.1051/ro/1977110201291>