

# ***Predicting China's GDP in Future Five Years by Linear Regression and ARIMA Model***

**Songyan Liu**

*Beijing Academy International Division, Beijing, China  
liusy070909@gmail.com*

**Abstract:** The Gross Domestic Product (GDP) is a key metric for assessing a country's economic status and level of development. As a fundamental indicator of national accounts and economic strength, GDP provides a comprehensive measure of economic activity. Therefore, accurately evaluating and forecasting GDP is essential for informed economic policymaking, strategic business investments, and the implementation of social welfare programs. Thus, delving into the GDP has numeral benefits. As the international situation is constantly changing and global economy is continuously developing, this paper particularly forecasts the change in GDP in China for future five years. Under this circumstance, the paper utilizes the Autoregressive Integrated Moving Average (ARIMA) and Linear Regression as the research methods to predict future GDP. Including the historical GDP dataset from National Bureau of Statistics, the paper derives different results from the forecasting, it is found that the predicted value of GDP in 2028 using ARIMA is 1031800 billion CNY, while 1663300 billion CNY using Linear Regression.

**Keywords:** GDP, ARIMA, Linear Regression Model, Time Series.

## **1. Introduction**

It is known that Gross Domestic Product (GDP) has four main components: Consumption (C), Investment (I), Government Spending (G), and Net Exports (NX), represented by the formula  $GDP = C + I + G + NX$ . For the calculation of GDP, Economists use three primary methods to approaches the values: the value-added approach, the production approach, and the income approach. The growth or decline of GDP serves as a critical indicator of economic cycles. A positive GDP growth rate signifies economic expansion, whereas a negative growth rate indicates a recession. GDP is calculated as the total value of goods and services produced within a specific period, multiplied by market prices. This gives rise to two key classifications: nominal GDP, which reflects current price levels, and real GDP, which adjusts for inflation using a base-year price level.

Recent research on forecasting China's GDP has employed diverse methods, each offering valuable insights while facing distinct limitations. One study utilizes nighttime light data from satellites to predict provincial GDP, overcoming challenges associated with mixed-frequency data [1]. By constructing a comprehensive time series, the researchers develop an unbiased predictor for economic growth, though this method may fail to capture economic activities in regions with significant non-light-emitting sectors. Another analysis compares Autoregressive Integrated Moving Average (ARIMA) and linear regression models, demonstrating that ARIMA is better suited for short-term GDP forecasts due to its ability to handle time-series data effectively [2]. However, its

reliance on data up to 2017 limits its applicability to more current economic conditions, and the linear regression model is vulnerable to distortions caused by outliers. A third study applies a Bayesian Vector Autoregression (BCAR) model, focusing on the influence of monetary policy on GDP predictions [3]. While this approach provides valuable insights, the subjectivity inherent in the use of Bayesian priors could introduce bias into the findings. Another study employs mathematical optimization techniques, including the Pontryagin Maximum Principle, to analyze GDP growth from 2016 to 2020, offering scenarios based on different management strategies [4]. However, these advanced methods require specialized knowledge, which may limit their accessibility. Lastly, a fifth study extrapolates past trends to predict China's economic growth over the next two decades, incorporating development theories like structural change and catching-up [5]. While it highlights key drivers such as population growth and education, the assumption that past growth patterns will continue may fail to account for unforeseen disruptions that could impact future growth. Collectively, these studies underscore both the potential and challenges of forecasting China's economic trajectory, highlighting the need for more adaptable and comprehensive models.

The essay focuses on forecasting China's GDP, with two main areas of exploration: comparing the ARIMA and linear regression models and using time series analysis for short-term GDP prediction. The study is structured around four key aspects. First, it examines data from 2000 to 2023, providing a comprehensive historical perspective. Second, it focuses specifically on mainland China to ensure a detailed and region-specific analysis. Third, it evaluates and compares the ARIMA and linear regression models to determine their effectiveness in GDP forecasting. Finally, the study centers on multiple essential economic indicators—the overall GDP dataset, Consumption, Investment, Government Spending and Net Export—which are crucial for understanding economic trends and making accurate predictions.

## 2. Methodology

### 2.1. ARIMA analysis

ARIMA, which is Autoregressive Integrated Moving Average Model, entails mainly three parts: Autoregressive Model (AR), Differential Process (I) and Moving Average Model (MA) [6]. The notion of the model is to utilize the data from the past to predict future trends. For AR, the influence of observations from past events are considered. For Differential Process, the seasonality factors in the times series will be eliminated in order to acquire a precise value. For MA, the past forecast errors are taken into account on the current prediction. In this essay, the historical data of GDP is used to established the time series. Additionally, the stationary data is qualified by Augmented Dickey-Fuller (ADF) test.

$$Y_t = c + \sum_{i=1}^p \phi_i Y_{t-i} + \sum_{j=0}^q \phi_j \epsilon_{t-j} + \epsilon_t \quad (1)$$

where  $Y_t$  is the current value of the time series,  $c$  is a constant term,  $\phi_i$  represents the coefficients of the autoregressive (AR) part,  $\phi_j$  represents the coefficients of the moving average (MA) part, and  $\epsilon_t$  is the white noise error term, satisfying  $\epsilon_t \sim N(0, \sigma^2)$ .  $p$  is the autoregressive order, indicating how many past time points are considered,  $q$  is the moving average order, indicating how many past error terms are considered, and  $d$  is the differencing order, making the time series stationary.

Using the original GDP data shown in Table 1, the essay establishes time series and plots the time series data. According to the plot in Figure 1, the GDP exhibits the exponential growth, indicating that the variance becomes greater overtime. Thus, data is non-stationary and requires transformation

before applying ARIMA. Under this circumstance, the essay utilizes the log transformation to stabilize the variance, making it easier to apply ARIMA method.

The ARIMA model requires the time series to be stationary, which means that the variance, mean do not change overtime. To guarantee these prerequisites, the essay applies first-order differencing to remove the trend and achieve stationarity. For the sake of accuracy, the differenced log-transformed series is then tested again for stationarity.

Table 1: GDP in time series adopted from national bureau of statistics

| Year | GDP (in billions of CNY) | Year | GDP (in billions of CNY) |
|------|--------------------------|------|--------------------------|
| 1990 | 1,866.78                 | 2007 | 26,581.03                |
| 1991 | 2,178.15                 | 2008 | 31,404.54                |
| 1992 | 2,692.35                 | 2009 | 34,090.28                |
| 1993 | 3,533.36                 | 2010 | 40,151.28                |
| 1994 | 4,819.74                 | 2011 | 47,156.43                |
| 1995 | 6,079.37                 | 2012 | 51,947.01                |
| 1996 | 7,117.66                 | 2013 | 56,884.52                |
| 1997 | 7,897.30                 | 2014 | 63,646.27                |
| 1998 | 8,440.20                 | 2015 | 67,670.83                |
| 1999 | 8,967.70                 | 2016 | 74,412.70                |
| 2000 | 9,921.45                 | 2017 | 82,075.43                |
| 2001 | 10,965.52                | 2018 | 90,030.95                |
| 2002 | 12,033.27                | 2019 | 99,086.51                |
| 2003 | 13,582.37                | 2020 | 101,356.72               |
| 2004 | 15,987.83                | 2021 | 114,923.74               |
| 2005 | 18,493.74                | 2022 | 120,583.17               |
| 2006 | 21,631.44                | 2023 | 126,058.20               |

According to Duke University, the definition of first-order difference is that “the first difference of a time series is the series of changes from one period to the next. If  $Y_t$  denotes the value of the time series  $Y$  at period  $t$ , then the first difference of  $Y$  at period  $t$  is equal to  $Y_t - Y_{t-1}$ . In Statgraphics, the first difference of  $Y$  is expressed as  $\text{DIFF}(Y)$ , and in RegressIt it is  $Y\_DIFF1$ . If the first difference of  $Y$  is stationary and also *completely random* (not autocorrelated), then  $Y$  is described by a random walk model: each value is a random step away from the previous value. If the first difference of  $Y$  is stationary but *not* completely random--i.e., if its value at period  $t$  is autocorrelated with its value at earlier periods--then a more sophisticated forecasting model such as exponential smoothing or ARIMA may be appropriate.” [7].

Once the data is stationary, the essay determine the parameters of the ARIMA, which is ARIMA (p, d, q), where P is the Auto Regressive order, obtained Partial Autocorrelation Function (PACF) to identify; d is the Differencing order, the number of times the data was differenced; and q is the Moving Average order, derived by Autocorrelation Function (ACF).

The definition of PACF, according to Penn State University, is that “in general, a partial correlation is a conditional correlation. It is the correlation between two variables under the assumption that people know and take into account the values of some other set of variables” [8]. Thus,

$$\phi_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1} \phi_{k-1,j} \rho_{k-j}}{1 - \sum_{j=1}^{k-1} \phi_{k-1,j} \rho_j} \quad (2)$$

Where  $\rho_k$  is the autocorrelation function (ACF) value at lag  $k$  and  $\phi_{k-1,j}$  is the partial autocorrelation function (PACF) coefficient at lag  $k - 1$ . The definition of ACF, according to Penn State University, is that “the ACF property describes a unique autocorrelation pattern. When  $\phi_1$  is positive, the ACF gradually declines to zero in an exponential manner as the lag  $h$  grows.

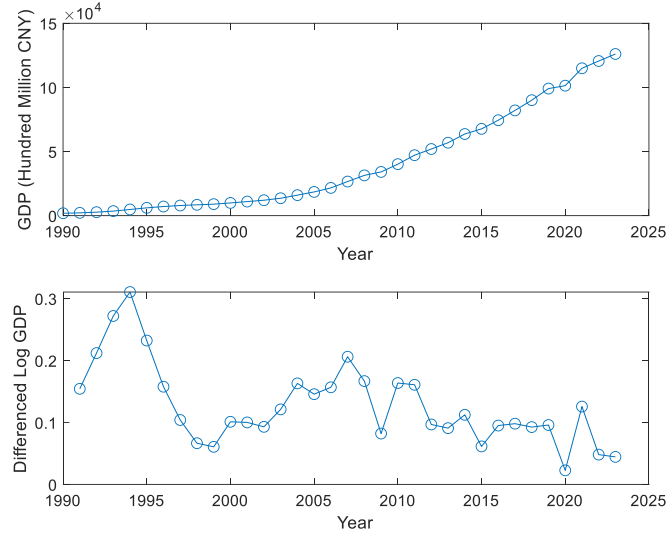


Figure 1: Original GDP and log GDP first difference

Conversely, if  $\phi_1$  is negative, the ACF still shows exponential decay toward zero with increasing lag, but the autocorrelation values oscillate between positive and negative signs [9], leading to

$$\rho_h = \phi_1^h \quad (3)$$

where  $h$  is the time period and  $\rho_h$  is the correlation. After running the PACF and ACF analysis of the data, the essay chooses the parameter of the model. Because ACF of the data has lag of 1, and PACF of the data has lag of 2, which are both significantly higher than the confidence bound, the essay decides to choose the parameter  $p$  as 2 and  $q$  as 1.

After determining the model, the essay utilizes MATLAB to write the code, implementing ARIMA (2,1,1). Then, the essay finds the Gaussian Distribution of the model, see Table 2.

Table 2: ARIMA (2,1,1) model (gaussian distribution)

|          | Value     | Standard Error | T Statistic | P Value     |
|----------|-----------|----------------|-------------|-------------|
| Constant | 0.017833  | 0.0074258      | 2.4015      | 0.016326    |
| AR {1}   | 1.6494    | 0.065631       | 25.131      | 2.2889e-139 |
| AR {2}   | -0.79804  | 0.041574       | -19.196     | 4.0171e-82  |
| MA {1}   | -1        | 0.15924        | -6.2798     | 3.3893e-10  |
| Variance | 0.0015093 | 0.00052225     | 2.89        | 0.0038528   |

Setting this real value set into the original formula, the essay acquires the new formula for the first order differencing data:

$$Y'_t = 0.017833 + \sum_{i=1}^2 (1.6494) Y'_{t-i} + \sum_{j=0}^1 (-1) \epsilon_{t-j} + \epsilon_j, \quad (4)$$

where  $Y_t' = \Delta^d Y_t = (1 - B)^d Y_t$ ,  $B$  is the backshift operator ( $BY_t = Y_{t-1}$ ). In this paper,  $d$  is equal to 1 according to model.

To ensure the ARIMA model's adequacy, the essay examines its residuals—the differences between observed and predicted values. Properly fitted models should produce residuals that resemble white noise. In particular, the essay utilizes the Ljung-Box test to assess ARIMA model. Ljung-Box test testifies whether residuals are uncorrelated up to a specified lag. According to Null Hypothesis, residuals showing no autocorrelation occurs at p-value is less than 0.05. However, in this the model shows that p-value equals to 0.21999, meaning that residuals exhibit no significant autocorrelation. Thus, it is reasonable to use ARIMA (2,1,1) to predict GDP.

## 2.2. Linear regression analysis

Linear regression is a widely used choice for predicting GDP because it's simple, easy to understand, and effective at capturing the relationships between different economic factors. GDP is influenced by several variables like consumption, investment, government spending, and net export, which often move together in a linear way. Using linear regression, economist quantify how each of these factors affects GDP, giving people a clear idea of their individual impact. The beauty of this model is that its results—known as coefficients ( $r$ )—express exactly how much a change in one factor will influence GDP, which is valuable for making informed deductions [10]. Based on these benefits, the essay utilizes the Linear Regression Model to predict GDP in the future 4 years which gives that

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_3 X_3 + \epsilon \quad (5)$$

where  $Y$  is the response variable,  $X_1, X_2, \dots, X_n$  are the explanatory variables,  $\beta_0$  is the intercept (the value of  $Y$  when all  $X$ 's are 0), and  $\beta_1, \beta_2, \dots, \beta_n$  are the coefficients (Table 3).

Table 3: GDP components. adopted from national bureau of statistics

| Year/GDP Component | Investment (in billions of CNY) | Consumption (in billions of CNY) | Government Spending (in billions of CNY) | Net Export (in billions of CNY) |
|--------------------|---------------------------------|----------------------------------|------------------------------------------|---------------------------------|
| 2000               | 3291.7                          | 3917.2                           | 1318.7                                   | 239                             |
| 2001               | 3721.4                          | 4305.5                           | 1637.6                                   | 232.5                           |
| 2002               | 4349.9                          | 4202.7                           | 2205.3                                   | 309.4                           |
| 2003               | 5556.7                          | 4584.2                           | 2460.7                                   | 298.6                           |
| 2004               | 7047.7                          | 5300                             | 2836.1                                   | 407.9                           |
| 2005               | 8877.4                          | 6717.7                           | 3159.5                                   | 1022.3                          |
| 2006               | 10999.8                         | 7641                             | 3483.9                                   | 1665.5                          |
| 2007               | 13732.4                         | 8921                             | 3957                                     | 2342.3                          |
| 2008               | 17249.9                         | 10848.8                          | 4456.2                                   | 2422.7                          |
| 2009               | 22484.6                         | 12534.3                          | 4978.1                                   | 1503.7                          |
| 2010               | 27898.2                         | 15699.8                          | 5643.5                                   | 1505.7                          |
| 2011               | 31192.1                         | 18391.9                          | 6226.4                                   | 1168.8                          |
| 2012               | 36483.5                         | 21030.7                          | 6753.2                                   | 1463.6                          |
| 2013               | 43692.8                         | 23780.5                          | 7228.1                                   | 1455.2                          |
| 2014               | 48926.9                         | 26239.4                          | 7790                                     | NA                              |
| 2015               | 55159                           | 28678.4                          | 8381.4                                   | NA                              |
| 2016               | 60646.6                         | 31580.8                          | 8989.6                                   | NA                              |
| 2017               | 64123.8                         | 34384.7                          | 9581.6                                   | NA                              |
| 2018               | 64567.5                         | 38098.7                          | 10300.9                                  | N/A                             |

Table 3: (continued).

|      |         |         |         |     |
|------|---------|---------|---------|-----|
| 2019 | 64567.5 | 41164.9 | 10703.9 | N/A |
| 2020 | 51890.7 | 39198.1 | 11366.5 | N/A |
| 2021 | 54454.7 | 44082.3 | 12100.1 | N/A |
| 2022 | 57213.8 | 45937.9 | 12987.2 | N/A |
| 2023 | 52091.6 | 48678.9 | 27457.4 | N/A |

There are two different variables. The first is the response variable, i.e., GDP. The second is the explanatory variable. Here, investment includes foreign investment and domestic investment, consumption includes customers spending and government purchase, government spending includes transfers and welfares, and net export is the exports subtract imports. Setting the real-world situation into the Linear Regression Model, the essay assumes the precise formula is:

$$GDP = \beta_0 + \beta_1 Investment + \beta_2 Consumption + \beta_3 Govern. Spending + \beta_4 NetExport \quad (6)$$

Because of the incomplete Net Export data, missing values in Net Export are handled using the method of interpolation padding, which is a technique to predict the unknown values based on existing data points.

Mean Square Error Test is used in this essay to verification the model predicting GDP. Mean Squared Error Test basically is the calculation of the average squared difference between the actual GDP values and the predicted values generated by model. A lower MSE implies that the models' forecasts are closer to the actual values, meaning a better performance.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (7)$$

where  $y_i$  is the actual value of the data and  $\hat{y}_i$  is the estimated value.

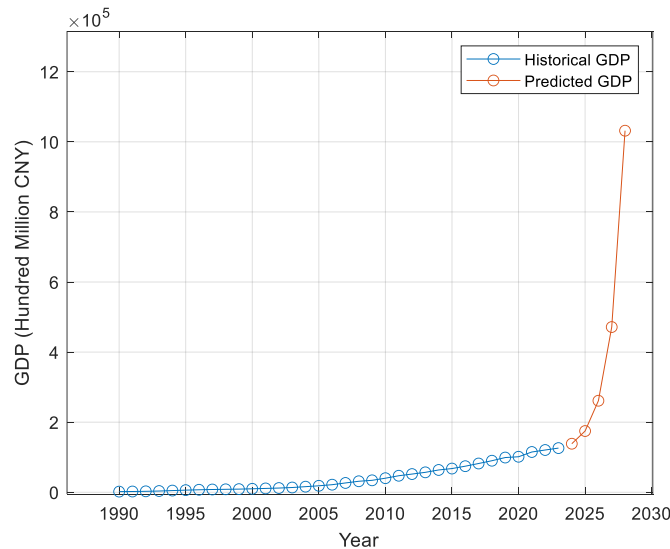


Figure 2: Historical and predicted overall GDP trends using ARIMA (1990-2028)

### 3. Results

#### 3.1. Results for China

After running ARIMA (2,1,1) model, the essay obtains the original prediction of China's overall GDP, as shown in Figure 2. In 2024, the predicted GDP in billions of CNY is  $1.3852e+05$ . In 2025, the predicted GDP in billions of CNY is  $1.7473e+05$ . In 2026, the predicted GDP in billions of CNY is  $2.6125e+05$ . In 2027, the predicted GDP in billions of CNY is  $4.7152e+05$ . In 2028, the predicted GDP in billions of CNY is  $1.0318e+06$ .

After running Linear Regression model, the essay obtains another prediction of China's overall GDP, see Figure 3. In 2024, the predicted GDP in billions of CNY is  $1.3336e+05$ . In 2025, the predicted GDP in billions of CNY is  $1.416e+05$ . In 2026, the predicted GDP in billions of CNY is  $1.4985e+05$ . In 2027, the predicted GDP in billions of CNY is  $1.5809e+05$ . In 2028, the predicted GDP in billions of CNY is  $1.6633e+05$ .

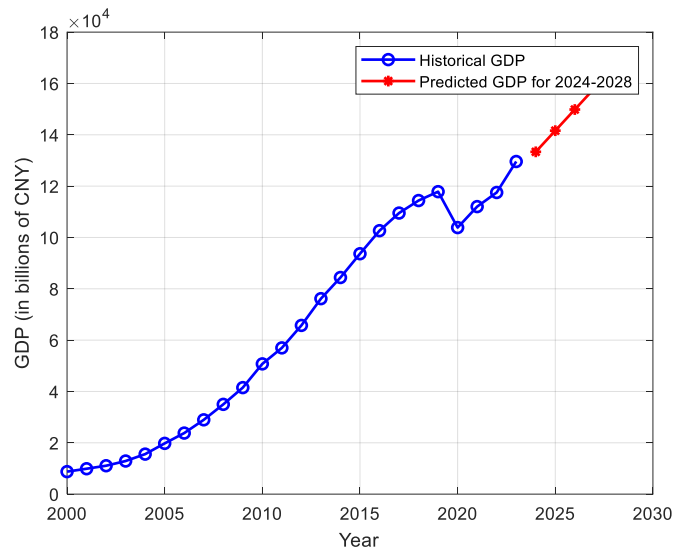


Figure 3: Historical and predicted overall GDP trends using linear regression

#### 3.2. Results for Beijing

Utilizing the same methodology used in predicting overall China's GDP, the essay delves into specific region of China to provide a case study, in which applies ARIMA and Linear Regression in order to derive the regional prediction of Beijing's GDP.

For Beijing's GDP, the steps for ARIMA are similar to those of overall China's GDP. First, searching for enough Beijing's GDP data from National Bureau of Statistics to build a time series. Second, graphing the Beijing's GDP trends and progressing the Augmented Dickey-Fuller Test, using differencing if necessary. Third, if GDP shows an exponential increasing pattern, using log transformation to linearize and difference data. Fourth, determining the parameters of ARIMA model, which is p, d, and q. Then, using MATLAB to establish the model and visualize the graph. Finally, testing the model by residual plot and Ljung-Box test. The result is shown in Figure 4.

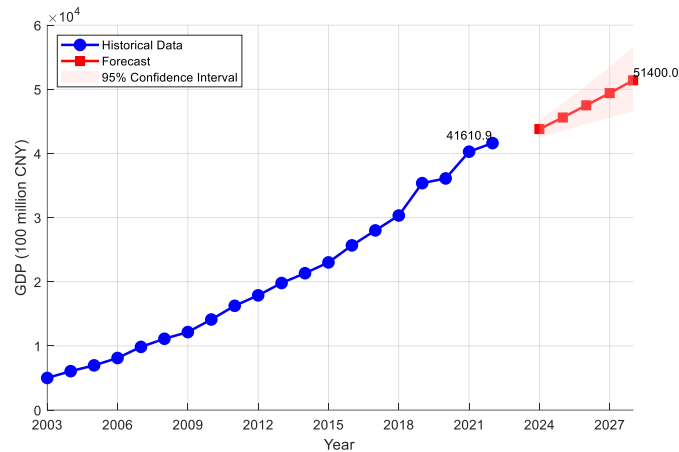


Figure 4: Beijing GDP trend forecast (2003-2028) by ARIMA

In the Linear Regression model, the Beijing's GDP somewhat resembles that of overall China's GDP. Both of the prediction is based on four explanatory variables: investment, consumption, government spending and net export. Then, using MATLAB to implement such dataset. Finally, Mean Square Error test is used to testified the feasibility of the model. The result is shown in Figure 5.

The ARIMA (2,1,1) model demonstrates strong prediction accuracy, making it a reliable choice for forecasting. One of its key strengths is its ability to produce white noise residuals, which is ideal for ensuring clean and unbiased predictions. Additionally, this model effectively captures underlying trends while remaining robust against outliers, enhancing its reliability. In contrast, the Linear Regression model offers moderate prediction accuracy and relies on autoregression for forecasting. However, its residual analysis tends to be weaker, limiting its effectiveness. While it can identify trends, it is more sensitive to outliers, which can distort predictions. For robust and accurate forecasts, ARIMA (2,1,1) is the preferred choice, especially when dealing with GDP calculation.

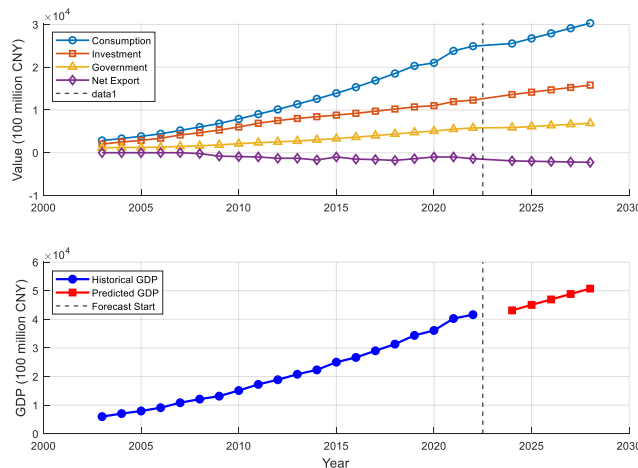


Figure 5: Beijing GDP components and trend forecast (2003-2028) by linear regression

#### 4. Conclusion

From the chart, the essay identifies that GDP prediction between two methodology has significant differences when it comes to 2025 to 2028 years, partially because of some flaws in two methods.

For ARIMA, the essay should add more indicators, such as CPI and PPI or inflation rate, to create more complex model, ARIMAX, to derive a more accurate prediction. Thus, minute inaccuracy exists in the ARIMA (2,1,1), accounting for the discrepancy in the results between two models. For Linear Regression model, using extrapolation might present a huge threat to GDP forecasting, since it is unreliable to set up an assumption under the unknown interval. Considering all drawbacks and applications of the methodologies, the essay presents a holistic comprehension of two models. What makes this research unique is its comparative analysis of the ARIMA and linear regression models to improve GDP prediction accuracy. By combining the strengths of both models, the study aims to enhance forecasting precision and provide a more reliable economic outlook. To test the effectiveness of this approach, the models are validated using data from the National Bureau Statistics and China's official economic records, proving their ability to capture economic trends and improve prediction accuracy.

Additionally, the research incorporates the overall GDP, investment, consumption, government spending and net export, as important inputs, recognizing its role in reflecting economic fluctuations. By integrating these key economic indicators with machine learning techniques, the study strengthens the forecasting model. To make this approach more accessible for future research and practical use, the study also provides MATLAB code for implementing these models. For the discussion of future application, both models offer valuable insights, their effectiveness depends on the nature of the data and the specific forecasting objectives. In practice, combining multiple models—such as using ARIMA for short-term trends and linear regression for long-term structural analysis—can lead to more robust and accurate predictions. As data-driven decision-making continues to evolve, the thoughtful application of these models will remain essential in guiding policies, business strategies, and scientific advancements.

## References

- [1] Holz, C. A. (2008). *China's economic growth 1978–2025: what we know today about China's economic growth tomorrow*. *World Development*, 36(10), 1665-1691.
- [2] Zhang, A. (2024). *Research of Forecasting Chinese Gross Domestic Product Based on Auto Regressive Integrated Moving Average Algorithm and Linear Regression Model*. *Advances in Economics, Management and Political Sciences*, 101, 151-160.
- [3] Higgins, P., Zha, T., & Zhong, W. (2016). *Forecasting China's economic growth and inflation*. *China Economic Review*, 41, 46-61.
- [4] Oliinyk, V., & Kozmenko, S. (2019). *Forecasting and management of gross domestic product*. *Journal of International Studies*, 12(4).
- [5] Gu, Y., Shao, Z., Huang, X., & Cai, B. (2022). *GDP forecasting model for China's provinces using nighttime light remote sensing data*. *Remote Sensing*, 14(15), 3671.
- [6] Chen Xingxing (2020). *Research on influencing factors of domestic tourism consumption based on multiple linear regression analysis*. *Advances in Social Science*, 9(9), 1530-1536.
- [7] Robert Nau, Fuqua School of Business, Duke university. (2020). *Statistical forecasting: notes on regression and time series analysis. Chapter 5, ARIMA models for time series forecasting. Identifying the order of differencing*. [https://people.duke.edu/~rnau/Decision411\\_2007/411arim2.htm](https://people.duke.edu/~rnau/Decision411_2007/411arim2.htm)
- [8] PennState Eberly College of Science STAT 510 Applied Time Series Analysis <https://online.stat.psu.edu/stat510/lesson/2/2.2>
- [9] PennState Eberly College of Science STAT 510 Applied Time Series Analysis <https://online.stat.psu.edu/stat510/lesson/1/1.2>
- [10] Li Wenyue, Li Jungang (2022). *Based on linear regression model, this paper makes an empirical analysis of China's GDP growth*. *Statistics and Application*, 12(29), 1499.