Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/101/20231594

Research of Forecasting Chinese Gross Domestic Product Based on Auto Regressive Integrated Moving Average Algorithm and Linear Regression Model

An Zhang[1,a,*]

[1]*University of California, Davis, Davis, California,95617, United States*
a. aanzhang@ucdavis.edu
**corresponding author*

Abstract: Gross domestic product (GDP) is an important indicator for measuring a country's economic development level. By studying GDP, it can analyze the economic structure of a country or region, determine the development trend of a country. GDP is widely applied in fields such as macroeconomic analysis, policy formulation, and international economic comparison. After Chinese reform and opening up, the economy has developed rapidly, and the GDP has shown a trend of increasing year by year. Therefore, the analysis of Chinese GDP is meaningful. This article analyzes GDP data from 2001 to 2017. Linear regression method is used for prediction first but the goodness of fit for this model to predict GDP is not very high. So then the Auto Regressive Integrated Moving Average (ARIMA) model is used for prediction. The results indicate that the ARIMA algorithm has higher goodness of fit and can better illustrate the situation, which can be used for short-term prediction in China. All this model is realized in the python.

Keywords: GDP, ARIMA, Forecasting, Time Series

1. Introduction

GDP is one of the key issues affecting social stability and social form, which is a very important topic for China in its high-speed development period. Since the reform and opening up, China's economy has achieved rapid development, with GDP increasing from 364.52 billion yuan in 1978 to 74412.7 billion yuan in 2016, an increase of approximately 215 times [1]. The yearbook figures imply that real GDP grew by 24.7% between 1997 and 2000 [2]. However, under the pressure of COVID-19, China is facing a huge problem with restoring Economic development. How to maintain rapid economic development and transform the economic development mode under the pressure of economic downturn is a major issue facing China's economic development [1].

This study applies the ARIMA algorithm and linear regression model to explore China's GDP data, providing a method for understanding China's economic situation and having practical significance for China's adjustment strategies or policies.

The following sections are as follows: Section 2 will include a literature review of previous studies. The third section will introduce the code and detailed research methods. Then, Section 4 will introduce the results and analysis. Finally, the research findings and discussions will be presented in Section 5.

DOI: 10.54254/2754-1169/101/20231594

2. Literature Review

In recent years, China has developed rapidly. China has experienced remarkable economic growth over the past 25 years. According to the World Bank, China's real GDP in1991 was about $0.91 trillion; China's real GDP in1991 was about $0.91 trillion; in 2015, it reached $8.91 trillion (both figures in constant 2010 USD) (World bank 2017). China has become the second largest economy in the world, closely following the United States. From 1991-2015, there has been an eightfold increase in real CDP per capita, rising from about $800 to $6500 (in constant 2010 USD) [3]. Therefore, effectively and quickly predicting the future economic development and GDP trend is more conducive to the country's next economic development. Linear regression model has been widely used in many fields because of its advantages, such as simple and easy to understand, high computational efficiency and strong interpretability. For example, Vehicle miles traveled (VMT) is an essential input for many aspects of transportation engineering, and an accurate estimation of VMT is critical for practicing engineers. Linear regression models are a popular method to estimate VMT as they provide insight into the relationships between VMT and other external factors [4]. In other aspects, Dastan Maulud, Adnan M Abdulazeez et al. discussed the various work of different researchers on linear and polynomial regression and compared their performance, using the best methods to optimize the prediction and accuracy. However, linear regression model has some limitations in classification problems. Because the output range of linear regression model is continuous real numbers, it can not be directly applied to discrete categories in classification problems. And linear regression models are very sensitive to outliers. Even if there is only one outlier, it will have a great impact on the fitting of the model. At the same time, it cannot deal with multicollinearity and nonlinear relations. In order to choose a more accurate approach, we also used the ARIMA model to re-evaluate and compare with the linear regression model to select a better model. As a time series model, the ARIMA model has been successful in many ways. For example, the novel model integrates autoregressive integrated moving average (ARIMA) and adaptive boosting (AdaBoost) machine learning, called ARIMA-AdaBoost. Compared with the previous research, the experiment results have shown that the ARIMA-AdaBoost outperforms the simple AdaBoost and Long Short-Term Memory (LSTM)-AdaBoost for transformer quality predictions [5]. Like linear regression models, ARIMA models can be calculated more accurately and require only endogenous variables without the need for other exogenous variables. However, it requires that the time series data be stable and that it also fails to capture nonlinear relationships [6]. For the ARIMA model, it is necessary to determine whether the data are smooth or not, and if the data are not smooth, the data need to be differenced by different orders to make the data smooth [7]. Prapanna Mondal, Labani Shit, Saptarsi Goswami and others conducted a study on the effectiveness of the autoregressive Composite Moving Average (ARIMA) model on 56 Indian stocks from different industries. They chose the ARIMA model and used Akaike information criterion (AIC) to compare and parameterize the ARIMA model. In order to better predict the future economic trend and deploy the strategic point of economic growth, we hope to build and forecast the GDP growth situation in the future through these models.

3. Method

3.1. Linear Regression Model

Regression analysis is one of the most widely used technique for analyzing multi-factor data Its broad appeal and usefulness result from the conceptually logical process of using an equation to express the relationship between a variable of interest (the response) and a set of related predictor variable [8]. Generally speaking, linear regression can solve its equation through the least squares method, which can calculate the straight line for y=bx+a [9]. To derive the equation y = bx+a:

DOI: 10.54254/2754-1169/101/20231594

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{x} x_i{}^2 - n\overline{x}^2} \quad (3.1.1)$$

Bringing $\hat{b}$, $\overline{x}$ and $\overline{y}$ value from (3.1.1) into formula:

$$\hat{a} = \bar{y} - \hat{b}\overline{x} \quad (3.1.2)$$

Deriving the equation from (3.1.2):

$$\hat{y} = \hat{b}x + \hat{a} \quad (3.1.3)$$

However, there is often more than one factor that affects Y. Assuming there are X_1, X_2, X_3,.....,k factors. The linear relationship:

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \text{---} + \beta_k X_k + \varepsilon \quad (3.1.4)$$

The ε in the equation (3.1.4) means the random variable.

3.2. ARIMA Algorithm

This time series technique makes very few assumptions and is very flexible. It is theoretically and statistically sound in its foundation and no a priori postulation of models is required when analyzing failure data [10]. The ARIMA model has three parameters: p, d, and q.

p - represents the lag number (lags) of the time series data itself used in the prediction model, also known as the AR/Auto Regression term

d - represents the number of orders of differential differentiation required for temporal data to be stable, also known as the Integrated term.

q - represents the lag number (lags) of the prediction error used in the prediction model, also known as the MA/Moving Average term

3.2.1. AR Model

The Auto Regressive model in order p can be written as AR(p) and the equation to show the AR(p):

$$x_t = \Phi_1 x_{t-1} + \Phi_2 x_{t-2} + \text{---} + \Phi_p x_{t-p} + w_t \quad (3.2.1.1)$$

The x_t in the equation (3.2.1.1) is stationary, the Φ_1 , Φ_2 , ... Φ_p w_t are the constant.

3.2.2. MA Model

A discrete linear system with input u (n) having a mean of zero and a variance of σ The white noise sequence of the discrete linear system has an output of x (n), and the relationship between the output and input can be represented by the following difference equation:

$$X(n) = \sum_{r=0}^{M} b_r\, u(n-r) \quad (3.2.2.1)$$

b_r (r=0,... M) is the coefficient, the model represents that the current output is the weighted sum of M inputs from the present and past.

DOI: 10.54254/2754-1169/101/20231594

3.3. Data Collection

The data we are using is from the website “Our World in Data” (OWiD) as original data, the research is based on Hannah et al.'s data(2020), which contains 60 columns and 29,589 rows about CO_2 emission, GDP, population and etc. in different country worldwide and in different years. The data is in the website “CO_2 and Greenhouse Gas Emissions - Our World in Data”. However, the data we used was rewritten based on OWiD data, which contains 15 columns and 3281 rows.

3.3.1. Data Selection

The data we use has many parameters, and in the table 1 below are the names of the columns that we used in the data set.

Table 1: Target Columns in Data Set

Columns Names	Description
year	Year between 2001~2020
iso_code	Three-digit country code, defined by ISO 3166-1
GDP	GDP measured in international dollars.
population	Population of each country
CO_2	Annual production-based CO_2 emissions (million tonnes).

3.3.2. Features Selection for ARIMA Algorithm and Linear Regression Model

Our target country is China, so we only selected the "CHN" item for analysis in iso_code. Although we selected three factors: GDP, population, and GDP for analysis, we found that the ARIMA model cannot clearly describe the relationship between the three factors, so we only made predictions for GDP, while CO_2 and population showed the predicted results in the Linear model.

4. Result, Analysis and Discussion

4.1. Linear Regression Model Prediction

We first use the Linear Regression model to predict the value of CO_2, GDP and population.

4.1.1. Explanation of Coding for Linear Regression Model

The coding is shown in the appendix “predict”.

4.1.1.1. Importing and Selecting Data

Firstly, the data.csv file was imported and defined to the corresponding year 2017, and then the year and GDP of China were selected for trinomial analysis.

4.1.1.2. Training Linear Regression Model

After importing the data set the model, the model was constructed using linear regression.

DOI: 10.54254/2754-1169/101/20231594

4.1.1.3. Result of Linear Regression Model Prediction of GDP

Output the model to predict subsequent GDP data from 2000 to 2017 and three years later, plot the graph. The other two factors can be predicted by change word "GDP" to "CO_2" or "population", here we import the matplotlib.pyplot as plt.

4.2. Result of Linear Regression Model

We used a linear regression model to predict GDP, CO_2, and population, and plotted these three factors on the below figure 1, figure 2 and figure 3.

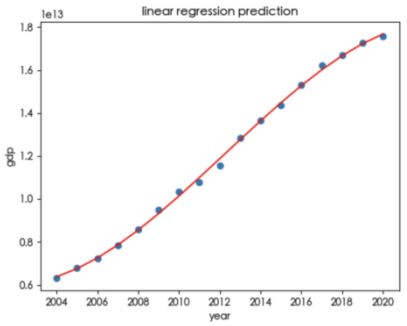

Figure 1: Prediction of GDP

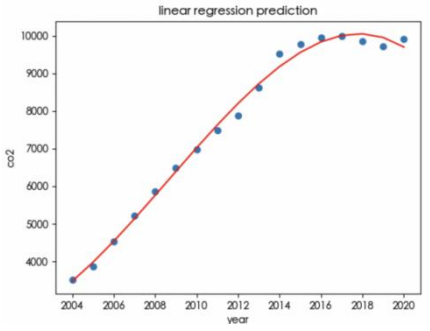

Figure 2: Prediction of CO_2

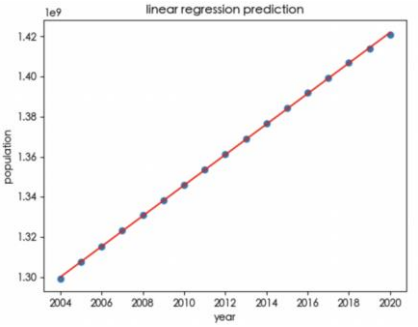

Figure 3: Prediction of Population

China's GDP from 2000 to 2020, which is closer to a linear function of one variable, but still has a certain curve. The pace of GDP growth picked up slightly from 2008 to 2016 and slowed slightly around 2018.

The second chart shows the projected CO_2 change in China from 2000 to 2020, with a larger curve. The growth rate of CO_2 increased steadily from 2000 to 2016, and slowed down or even decreased after 2018.

The third graph shows the projected population change of China from 2000 to 2020, which is closer to the linear function of one variable. The rate of population growth will increase steadily from 2000 to 2020.

However, we think that this model's goodness of fit is not so ideal. Thus, we try to use ARIMA model to do the prediction.

4.3. ARIMA Model Prediction

We think that the ARIMA model might be more precise when predict the GDP.

4.3.1. Coding and analyzing for ARIMA Model

The coding for this part is shown in the appendix "Predict-China".

4.3.1.1. Importing and Selecting Data

In the ARIMA model we want to use the data from 2001 to 2015 to predict the GDP value in the 2016 and 2017 to find the goodness of fit for this model. Thus, we need to import data from 2001 to 2015 and choose GDP as target.

4.3.1.2. Ploting the Time Series Graph

The two graphs are on the below, figure 4 and figure 5, figure 6, figure 7,

DOI: 10.54254/2754-1169/101/20231594

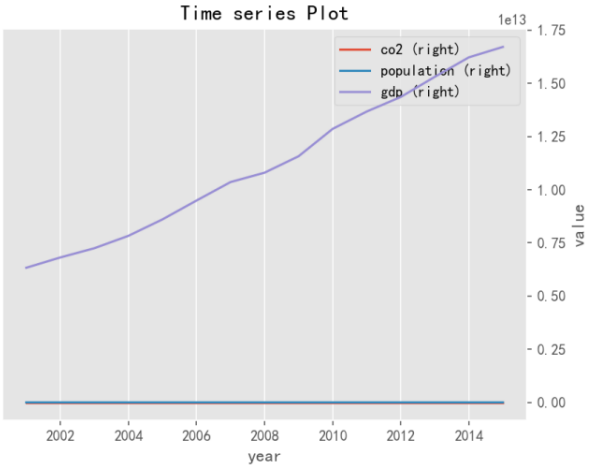

Figure 4: Time Series Plot

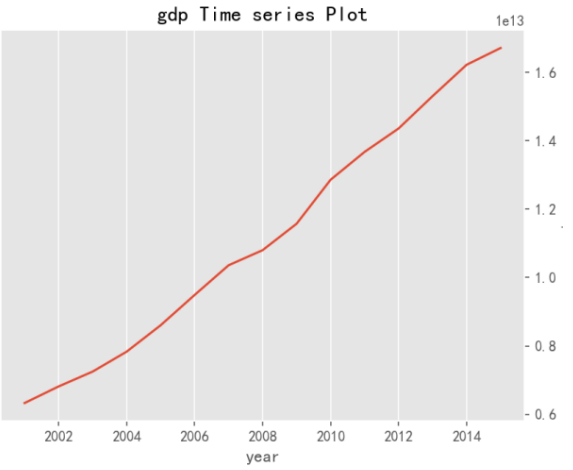

Figure 5: Time Series Plot for GDP

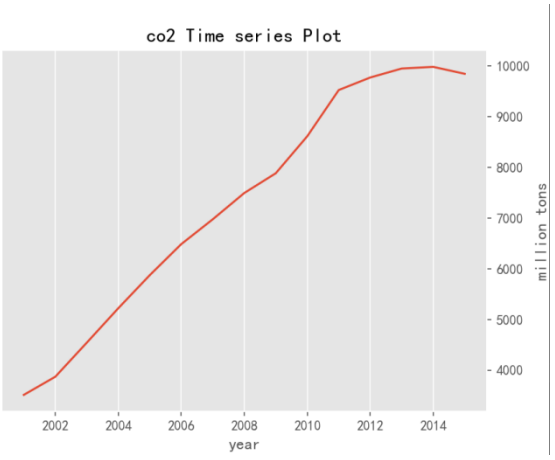

Figure 6: Time Series Plot for CO_2

Figure 7: Time Series Plot for Population

For time series, GDP is huge compared to carbon dioxide and the number of people, so this graph is mainly used to find a time series of GDP. In this chart, GDP growth is more stable and less volatile. Therefore, the second chart presents the GDP time series separately, and it can be seen more clearly that the growth curve is like a linear function of one variable, so the growth is relatively stable. It can be seen from the picture that the data model of population growth is very close to the linear equation of one variable, with a constant and steady growth from 2001 to 2015.Through the carbon dioxide data model, it can be seen that the annual growth rate of carbon dioxide is relatively constant from 2001 to 2009, but the increase rate of carbon dioxide is faster from 2009 to 2011. Then the rate of carbon dioxide slowed again from 2011 to 2015. Through the carbon dioxide data model, it can be seen that the annual growth rate of carbon dioxide is relatively constant from 2001 to 2009, but the increase rate of carbon dioxide is faster from 2009 to 2011. Then the rate of carbon dioxide slowed again from 2011 to 2015.

4.3.1.3. Testing for Stability of Time Series

This is the part of the for cycle, it import adfuller from statsmodels.tsa.stattools, which called ADF statistic. ADF statistic is a key indicator for determining whether a time series has a unit root. The result of the cycle give out that ADF statistic: 1.017281, p-value: 0.994445, critical value:1%: -

DOI: 10.54254/2754-1169/101/20231594

4.138, 5%: -3.155, 10%: -2.714. The p-value is greater than critical value 0.05, which means that the GDP's time series is not stable in the original data set.

4.3.1.4. Differential and Plotting GDP Time Series

The graph is shown below, figure 8 figure 9 figure 10.

Figure 8: Differential GDP Time Series Plot

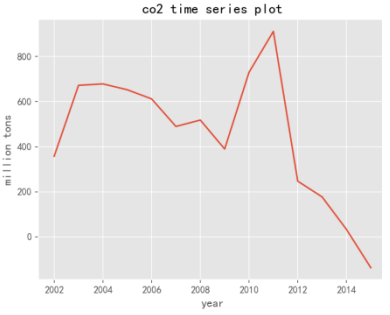

Figure 9: Differential CO_2 Time Series Plot

Figure 10: Differential Population Time Series Plot

To better see the more subtle changes in GDP growth, here is our GDP data after the first-order difference. As you can see, GDP grew from 2003 to 2006,2008 to 2010,2012 to 2013, and declined from 2006 to 2008,2010 to 2012,2013 to 2015. In the differential data model of carbon dioxide, the overall trend is decreasing, but in 2001 to 2023, 2007 to 2008, and 2009 to 2011, the difference data is increasing. In the differential model of population growth, its overall trend is downward curve. It showed a decreasing trend from 2001 to 2007 and from 2012 to 2015, and the slope was large. From 2007 to 2012, the data showed a brief upward trend.

4.3.1.5. Testing for Stability of Differential Time Series

Way of testing for the differential time series is the same as the way in the 4.2.1.3, the result after differential is ADF statistic: -3.653241, p-value: 0.004821, critical value:1%: -4.138, 5%: -3.155, 10%: -2.714. The p-value is much smaller than the critical value 0.05, this indicates that the time series of GDP data after first order differential is stationary. Then apply the acorr_ljungbox in the statsmodels.stats.diagnostic to do White noise test with lags (1, 3, 4, 5, 6) and the result show on the table 2.

Table 2: White Noise Test Value

lag	lb_stat	lb_pvalue
1	12.190003	0.000480
2	12.190003	0.000051
3	23.614178	0.000030
5	23.614178	0.000137
6	25.299453	0.000300

The table shows that the differential data also pass the white noise test.

DOI: 10.54254/2754-1169/101/20231594

4.3.1.6. Selfcorrelation and Model Order Determination

We import plot_acf from statsmodels.graphics.tsaplots to plot the graph the selfcorrelation of GDP data, the graph is shown below, figure 11. Then import statsmodels.api as sm to determine the order of the model.

The best _pq give out the result: [2, 1, 780.6117019240437, 783.1679312425048], thus, the order of the ARIMA model is (2, 1, 1)

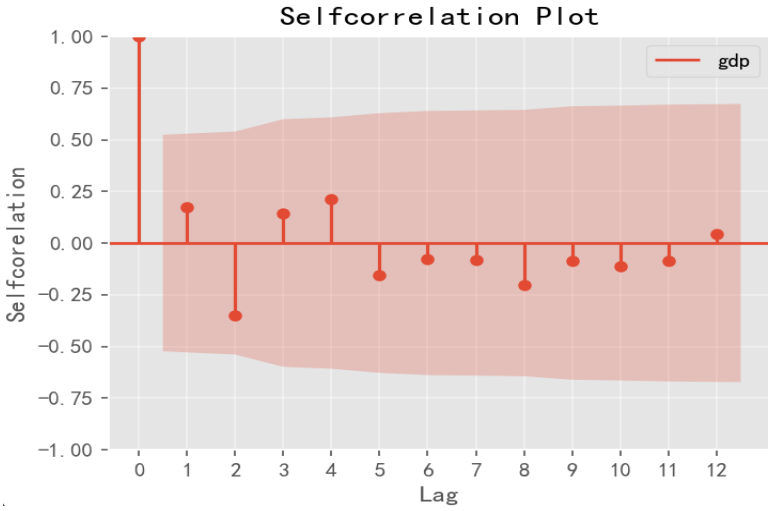

Figure 11: Selfcorrelation Plot

In this graph we can find the ACF graph, the shape of the graph is a decreasing cosine curve which means that the autocorrelation coefficient is trailing. The graph shows that the it decreases very fast and fluctuates around zero, which can also shows that the trailing.

4.3.1.7. Residual of Result

we plot the "ARIMA(2,1,1)Residual curve" and “dense curve”, which is shown below in figure 12 and figure 13.

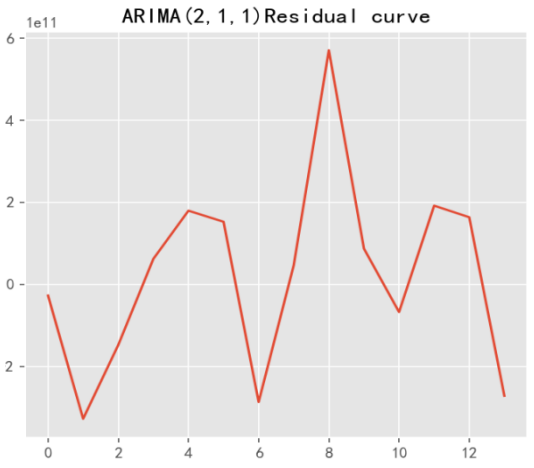

Figure 12: ARIMA(2,1,1)Residual Curve

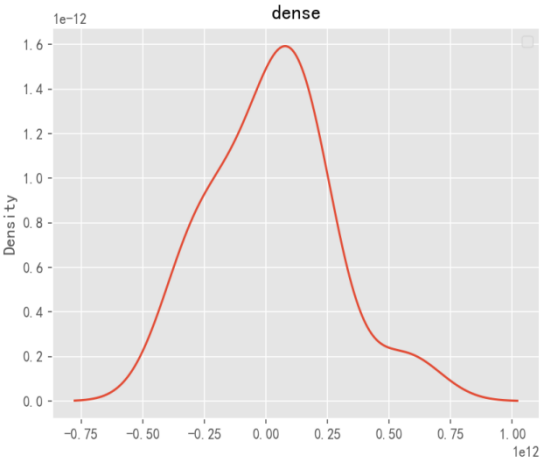

Figure 13: Dense Curve

In the Figure 8 we can find that the residual value is very large, we suspect that it is because of the original data, the value of original data is very huge. In figure 9, density curve presents a bell shape,

DOI: 10.54254/2754-1169/101/20231594

that is, a normal distribution, which means that the data obtained are stable and reliable, and can be used as a conclusion.

4.3.1.8. Prediction

Using ARIMA (2, 1, 1) to predict the GDP. The result of prediction is the 6.84172557e+12, 7.57635921e+12, 7.97881479e+12, 8.54082724e+12, 9.30982640e+12, 1.02055345e+13, 1.10860345e+13, 1.15258858e+13, 1.22877751e+13, 1.35859696e+13, 1.44255969e+13, 1.51124653e+13, 1.60571199e+13, 1.69826794e+13, 1.74785869e+13, 1.82341198e+13. The graph plot to compare about the predict value and original value is shown below, figure 10

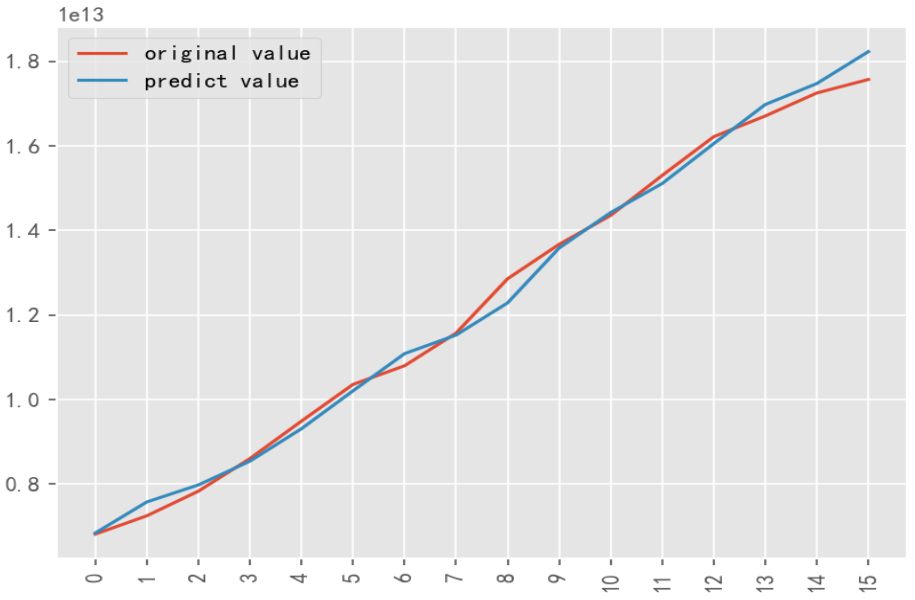

Figure 14: Comparison between Original Value and Predict Value

In the figure 14, the red line is the original data and the blue line is the predicted data. As can be seen from the graph, the difference between the two is small, so the prediction result is more accurate than linear regression model.

4.3.1.9. Goodness of Fit

Importing r2_score from sklearn.metrics to test the goodness of fit for this model. The value for the goodness of fit is 0.994021428860557 that means the model is very fit for predicting Chinese GDP. Then we test the corelation the result is 0.99720553 which is very impressive.

5. Conclusion

5.1. Conclusion

In this paper, we forecast and fit China's GDP quantity through linear regression model and ARIMA model. Through fitting, we find that the predicted value of ARIMA model is closer to the actual value, and its goodness of fit value is very high. Therefore, we believe that ARIMA model can more accurately predict the GDP of the coming years and better assist countries in formulating economic policies. Both ARIMA model and linear regression model are relatively easy to interpret. Linear models are simple and easy to use and can resolve multiple variables. However, it requires high data

DOI: 10.54254/2754-1169/101/20231594

accuracy and is almost exclusively suitable for linear problems. The advantage of ARIMA is that it does not need to rely on other exogenous variables, but it can only make short-term predictions and cannot use nonlinear functions.

5.2. Expectation

In this paper, we calculate the three data separately and obtain the predicted data, but there may be some connection among the three data, and these connections may affect each other's predicted value. In the future, we may find a more suitable data model to predict and fit the three data at the same time, so as to obtain the correlation among the three data and make better prediction. In addition, we only forecast for the next few years, if we can find a more stable numerical model, we can put in a larger number of data to forecast for the next few decades so that countries can prepare earlier. In this paper, we calculate the three data separately and obtain the predicted data, but there may be some connection among the three data, and these connections may affect each other's predicted value. In the future, we may find a more suitable data model to predict and fit the three data at the same time, so as to obtain the correlation among the three data and make better prediction. In addition, we only forecast for the next few years, if we can find a more stable numerical model, we can put in a larger number of data to forecast for the next few decades so that countries can prepare earlier.

References

[1] *Yu Lianmin, The Application of ARIMA Model in China's GDP Forecast, Times Finance, no. 21, pp. 1, 2017*

[2] *Rawski, T. G. (2001). What is happening to China's GDP statistics?, China Economic Review, 12(4), 347–354. doi:10.1016/s1043-951x(01)00062-1*

[3] *Chen, M., Kwok, C. L., Shan, H., & Yip, P. S. F. (2018). Decomposing and Predicting China's GDP Growth: Past, Present, and Future, Population and Development Review, 44(1), 143–157. doi:10.1111/padr.12129*

[4] *Asif Mahmud, Ian Hamilton, Vikash V. Gayah, Richard J. Porter, Estimation of VMT using heteroskedastic log-linear regression models, Transportation Letters, 2023, ISSN 1942-7867,*

[5] *Chun-Hua Chien, Amy J.C. Trappey, Chien-Chih Wang, ARIMA-AdaBoost hybrid approach for product quality prediction in advanced transformer manufacturing, Advanced Engineering Informatics, Volume 57, 2023, 102055, ISSN 1474-0346*

[6] *HHXUN, ARIMA model introduction, 2018-04-08 21:52:10*

[7] *Jun Luo, Yaping Gong, Air pollutant prediction based on ARIMA-WOA-LSTM model, Atmospheric Pollution Research,Volume 14, Issue 6,2023,101761,ISSN 1309-1042,*

[8] *Douglas C. Montgomery, Elizabeth A. Peck,G. Geoffrey Vining, Introduction to linear regression analysis, pp. 13 , 2021*

[9] *Draper, N.R. and Smith, H. Applied Regression Analysis. Wiley Series in Probability and Statistics. 1998.*

[10] *S.L. Ho and M. XIE, THE USE OF ARIMA MODELS FOR RELIABILITY FORECASTING AND ANALYSIS, Computers & industrial engineering, 1998 - Elsevier*