International Journal of Mathematics Trends and Technology
ISSN: 2231-5373/ https://doi.org/10.14445/22315373/IJMTT-V69I7P503

Volume 69 Issue 7, 17-31, July 2023

Original Article

Forecdiction of GDP Appreciation Ratio based on ARMA Model and Multiple Regression

Zhi Zhang[1], Hang Zhang[2], Xiang Han[3]

[1,2,3]*School of Mathematics and Information Science, Henan Polytechnic University, Henan, China*

Received: 30 May 2023 Revised: 03 July 2023 Accepted: 15 July 2023 Published: 28 July 2023

Abstract *- This paper selects the national GDP data from 2000 to 2021, using the method of multiple linear regression and ARMA model, analyze and forecast the national GDP value-added ratio, select the eight GDP value ratio influencing factors, the empirical analysis is conducted on the influencing factors of the GDP increment ratio through gradual regression, find out the influencing factors significantly affect the GDP value-added ratio. On this basis, we found a long-term stable relationship between the GDP value-added ratio and the related influencing variables through the co-integration test. By fitting the ARMA model in R language, we finally obtained the forecast value of GDP appreciation ratio in the next five years, analyzed the change trend, and combined with the current situation of the influencing factors of GDP increment ratio, analyze the future development of various industries.*

Keywords *- Multiple linear regression, Stepwise regression, ARMA model, Co-integration test.*

1. Introduction

We have collected data on China's Gross Domestic Product (GDP) from 2000 to 2021, as well as the GDP of various industries in China in recent years. We conducted an empirical analysis on these data, taking the GDP value-added ratio as the response variable and selecting the GDP value-added ratios of eight other industries as predictor variables. This analysis allows us to develop a mathematical model to predict the GDP value-added ratio and provide guidance for various industries[10].

Gross domestic product (hereinafter referred to as GDP in our article) is the final result of the production activities of all resident units in a country (or region) over a certain period of time. GDP is the core index of national economic accounting, and also an important index to measure the economic situation and development level of a country or region. According to data analysis, we have found that China's GDP has been steadily increasing in recent years. However, compared to developed countries, there is still a certain gap that exists.. China is the largest developing country in the world, and we are still far from developed countries.China's GDP has shown an overall upward trend, but the annual GDP value-added ratio varies. When the value-added ratio is higher, we can infer that China's domestic GDP has experienced significant growth in that year, bringing it closer to the GDP levels of developed countries. From the perspective of annual GDP value-added ratio, we will analyze the changes in China's domestic GDP in recent years, identify the predictor variables that have a significant impact on the GDP value-added ratio, and construct a model to analyze the future trend of GDP value-added ratio.

We collect some of the influence of GDP variables, through the data calculation of the corresponding predictive variables, prediction variables include: construction industry value ratio, industrial value ratio, industry added ratio, animal fishery value ratio, wholesale and retail value ratio, transportation (warehousing and postal service) value ratio, accommodation and catering industry, the real estate industry value added ratio. Through univariate logistic regression analysis and gradual regression, the significance of influencing factors was analyzed, and the multivarilinear model of GDP value-added ratio and predictor variables was obtained, and co-integration test. It was found that the GDP value-added ratio and predictor variables in multivariate linear model showed a long-term stable trend. By constructing the ARMA model on the predictor variables in the multivariate linear model, we obtained the value of the predictor variables in the next 5 years, and put it into the multivariate linear model we constructed to obtain the change trend of GDP appreciation compared to the next 5 years. We apply the method of combining the time series with the multivariate linear regression model, which can well explore the significant factors affecting the response variable, and predict the future value of the predictive variable. Combined with the multivariate linear model, we can well predict the future value of the response variable[13][11].

2. GDP Value-Added Ratio Prediction

2.1. Data Visibility Analysis of the Variables

All the data in this paper are obtained from the China Statistical Yearbook. Through the China Statistical Yearbook, we get the relevant data on nine variables from 2000 to 2021, GDP, agriculture, forestry, animal husbandry, fishery, industry, industry, transportation, transportation, finance and real estate. Except for the accommodation and catering industry, the other eight variables showed an overall upward trend from 2000 to 2021. We first gave a line chart on GDP.

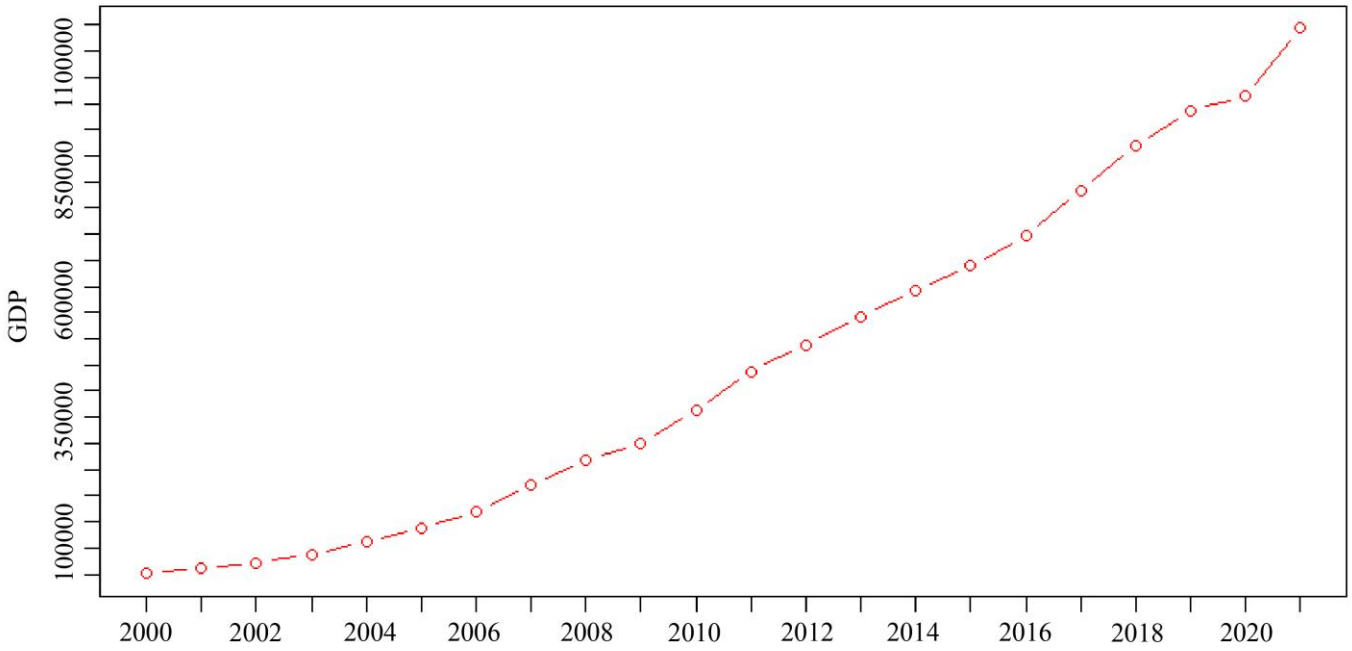

Fig. 1 GDP trends from 2000 to 2021

From Figure 1, we can see that the overall GDP increases monotonically, but the magnitude of the change from year to year is different. However, what we are interested in is not the actual value of GDP in each year, but mainly focuses on the annual added value of GDP. Through the analysis of the added value, we can know the development of various industries in China, and understand whether the pace of China's modernization is accelerated or slow, so we processed the data we collected accordingly. We give an expression for the value added ratio: (this year GDP-last year GDP) / previous year GDP, we calculate the value added ratio for nine variables, for 2000 by collecting data from 1999.

We give the actual GDP data and value-added ratio from 2000 to 2021. The specific information is given in the following table (Table 1).

Table 1. Actual GDP data and value-added ratio

Year	Value-added ratio%	Year	Value-added ratio%
2000	10.7	2011	18.0
2001	11.0	2012	10.0
2002	9.7	2013	10.0
2003	13.0	2014	8.5
2004	18.0	2015	7.0
2005	16.0	2016	8.4
2006	17.0	2017	11.0
2007	23.0	2018	10.0
2008	18.0	2019	7.3
2009	9.2	2020	2.7
2010	18.0	2021	13.0

The change trend of the GDP value-added ratio is shown in Figure 2:

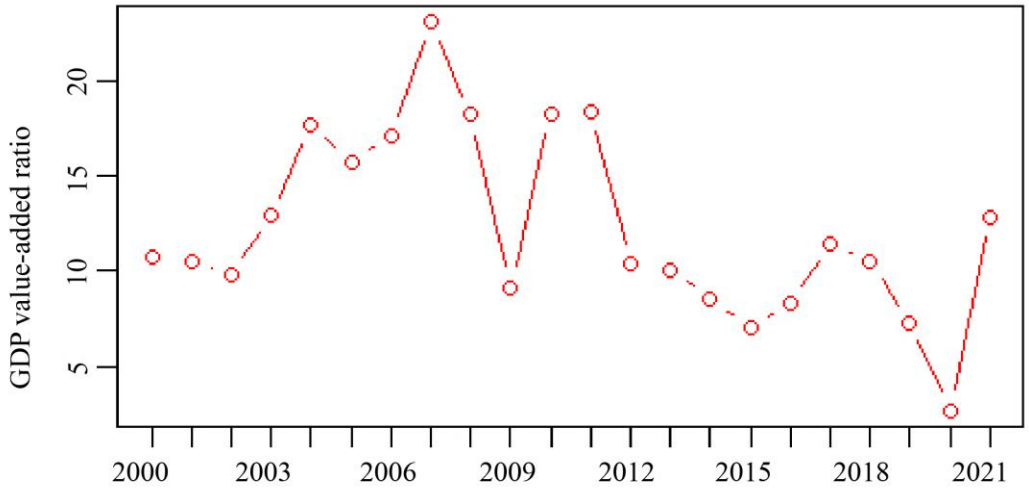

Fig. 2 Trend chart of GDP value-added ratio

As can be seen from Figure 2, China's GDP appreciation ratio from 2000 to 2021 is less than 5% in 2020, and the average GDP appreciation ratio in other years is above 5%, especially the GDP appreciation ratio in 2007, which may be related to the relevant policies implemented in China in 2007.

Due to the large number of 8 variables, it is too complicated to give the trend chart of the value-added ratio of each variable one by one. We give the comprehensive trend chart of 8 variables in Figure 3, as shown on the next page.

As seen from the figure below, we can clearly see that the increase ratio of the accommodation industry and catering industry is lower than 0 in 2020, indicating that China's catering industry and accommodation industry will show a negative growth state in 2020, which may be due to the impact of the epidemic, resulting in a serious impact on the catering industry and the accommodation industry in China.

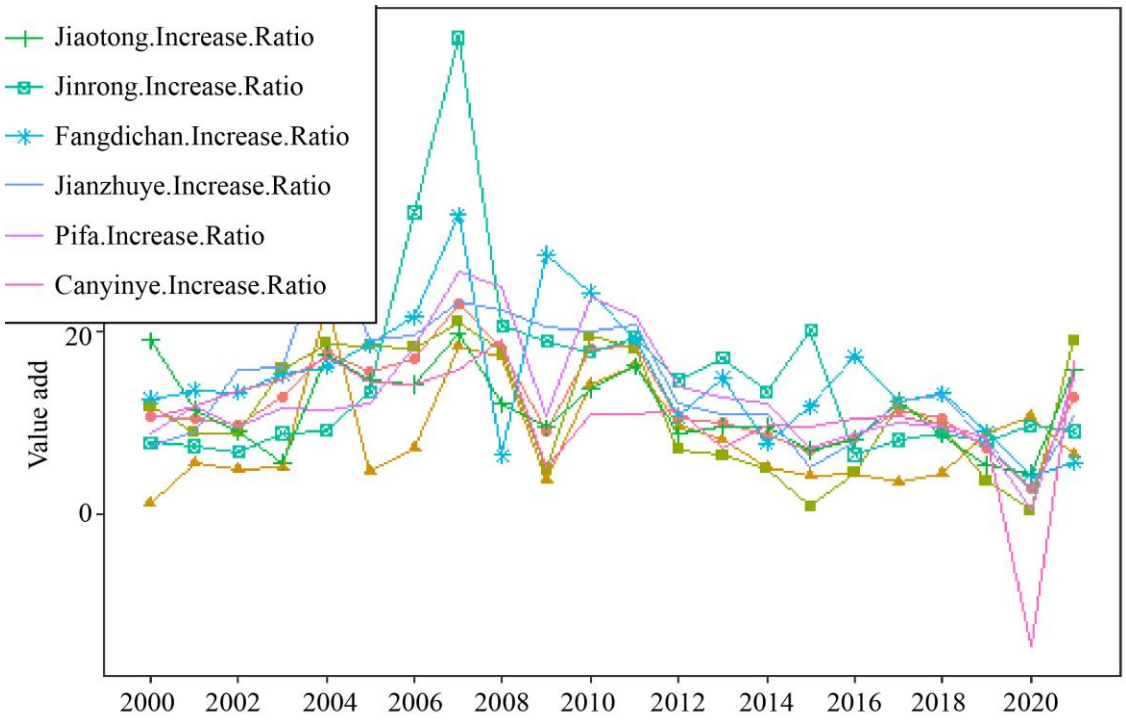

Fig. 3 Comprehensive trend diagram

The value-added ratio of the remaining variables is greater than zero, It shows that from 2000 to 2021, the GDP of China's eight sectors, namely construction, agriculture, forestry, fishing, industry, transportation, storage, postal services, finance, real estate, and wholesale and retail industries, exhibited an upward trend. However, there were differences in the development speeds among these eight industries.For instance, in 2007, there was a particularly significant development gap among the industries. The financial industry's value-added ratio reached 52.42% in 2007, showing a rapid growth rate, far higher than the value-added ratios of the other seven industries.The larger difference in the value-added ratio may be the impact of the 2008 financial crisis.

2.2. Construction of a Multiple Linear Regression Model

2.2.1. Univariate Regression Analysis

We give the variable in the paper, the previous section, we this paper is not mainly interested in the actual value of each variable, the actual value can only see the GDP in this year, does not reflect more meaningful information, so our main purpose is to analyze the change trend of GDP value ratio, it can reflect the change of GDP every year, which can reflect the development of modernization process in our country. We took the GDP value added ratio as the response variable and the 8 industry value added ratio as the predictor variable. x_1 value-added ratio of agriculture,forestry,animal husbandry and fishery; x_2 industrial value-added ratio; x_3 construction industry value-added ratio; x_4 wholesale and retail value-added ratio;$x_5$transportation value-added ratio;$x_6$housing, accommodation and catering industry value-added ratio;$x_7$value-added ratio of financial industry;$x_8$real estate industry appreciation ratio[6].

The specific idea of univariate regression analysis is extremely familiar to every statistician, and we will not describe the idea of univariate regression analysis.Through the analysis of R software, we obtained the influence of each predictor variable on the response variable respectively, and we calculated the correlation value of each predictor variable for the response variable. We will present it to you in the form of a table (as Table 2).

Table 2. Univariate regression analysis table

Variable	Coefficient	Standard error	t value	P value
x_1	0.5312	0.1436	3.699	0.00142
x_2	0.6618	0.0604	10.942	< 0.0001
x_3	0.5829	0.0862	6.758	< 0.0001
x_4	0.6769	0.0782	8.648	< 0.0001
x_5	0.8357	0.1571	5.318	< 0.0001
x_6	0.4986	0.1205	4.139	0.000509
x_7	0.2835	0.0806	3.514	0.00218
x_8	0.3864	0.1243	3.107	0.00555

Through the univariate regression analysis in Table 2, we could intuitively observe the P-value of the linear regression model of each predictor variable and the response variable, and could find that the effects of 8 predictor variables on the response variables were statistically significant.

2.2.2. *Multivariate Linear Regression Model*

First examining the correlation between individual predictor variables, the correlation coefficient Table 3.

Table 3. The correlation coefficient table

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	1.0000	0.4985	0.6519	0.5782	0.3934	0.1912	0.4250	0.2071
x_2	0.4985	1.0000	0.7472	0.7566	0.7680	0.6681	0.3879	0.3907
x_3	0.6519	0.7472	1.0000	0.6628	0.5589	0.5689	0.4314	0.5763
x_4	0.5782	0.7566	0.6628	1.0000	0.5958	0.6002	0.6768	0.4863
x_5	0.3934	0.7680	0.5589	0.5958	1.0000	0.5248	0.4200	0.4224
x_6	0.1912	0.6681	0.5689	0.6002	0.5248	1.0000	0.1987	0.2250
x_7	0.4250	0.3879	0.4314	0.6768	0.4200	0.1987	1.0000	0.6415
x_8	0.2071	0.3907	0.5763	0.4863	0.4224	0.2250	0.6415	1.0000

We generally believe that the correlation coefficient between variables is greater than 0.3 to consider a weak correlation between variables. We can find through the data of Table 3 that the problem of possible multicollinearity between predictive

variables. We used a stepwise regression to construct a multivariate linear model between the predictor variables and the response variables. Through gradual regression, we choose: $x_1$value-added ratio of agriculture, forestry, animal husbandry and fishery;$x_2$industrial value-added ratio;$x_4$wholesale and retail value-added ratio;$x_6$accommodation and catering industry value-added ratio;$x_7$value-added ratio of financial industry;$x_8$real estate industry appreciation ratio as predictive variables to construct multiple linear regression models. The specific information of multiple linear model is displayed in tables, as shown in Table 4[2].

Table 4. Multiple linear model information table

Variable	Coefficient	Standard error	t value	P value
x_1	0.1554	0.0317	4.896	0.000194
x_2	0.3666	0.0367	9.983	< 0.0001
x_4	0.1171	0.0486	2.409	0.02933
x_6	0.1122	0.0317	3.541	0.002962
x_7	0.0487	0.0225	2.166	0.046795
x_8	0.1044	0.0265	3.943	0.001301

After analysis in Table 4, the multiple linear regression equation is:

$$y = 1.7448 + 0.1554x_1 + 0.3666x_2 + 0.1171x_4 + 0.1122x_6 + 0.0487x_7 + 0.1044x_8 \quad (1)$$

Table 5. The ADF test form

Variable	ADF value	1%	5%	10%	Smooth type
y**contains intercept terms and time trend terms**	-2.9471	-4.38	-3.60	-3.24	non-stationary
y**contains intercept terms**	-2.2124	-3.75	-3.00	-2.63	
y**no intercept items and time trend items**	-0.6164	-2.66	-1.95	-1.6	
dy**contains intercept terms and time trend terms**	-5.8299	-4.38	-3.60	-3.24	steady
x_1**contains intercept terms and time trend terms**	-3.4376	-4.38	-3.60	-3.24	non-stationary
x_1**contains intercept terms**	-3.2324	-3.75	-3.00	-2.63	
x_1**no intercept items and time trend items**	-1.0763	-2.66	-1.95	-1.6	
dx_1**contains intercept terms and time trend terms**	-5.8299	-4.38	-3.60	-3.24	steady
x_2**contains intercept terms and time trend terms**	-3.0298	-4.38	-3.60	-3.24	non-stationary
x_2**contains intercept terms**	-2.4033	-3.75	-3.00	-2.63	
x_2**no intercept items and time trend items**	-0.7835	-2.66	-1.95	-1.6	
dx_2**contains intercept terms and time trend terms**	-4.6932	-4.38	-3.60	-3.24	steady
x_4**contains intercept terms and time trend terms**	-2.1935	-4.38	-3.60	-3.24	non-stationary
x_4**contains intercept terms**	-1.9522	-3.75	-3.00	-2.63	
x_4**no intercept items and time trend items**	-2.1935	-2.66	-1.95	-1.6	
dx_4**contains intercept terms and time trend terms**	-4.7667	-4.38	-3.60	-3.24	steady
x_6**contains intercept terms and time trend terms**	-2.9619	-4.38	-3.60	-3.24	non-stationary
x_6**contains intercept terms**	-1.24	-3.75	-3.00	-2.63	
x_6**no intercept items and time trend items**	-0.9607	-2.66	-1.95	-1.6	
dx_6**contains intercept terms and time trend terms**	-4.8264	-4.38	-3.60	-3.24	steady
x_7**contains intercept terms and time trend terms**	-2.4242	-4.38	-3.60	-3.24	non-stationary
x_7**contains intercept terms**	-2.2972	-3.75	-3.00	-2.63	
x_7**no intercept items and time trend items**	-1.0566	-2.66	-1.95	-1.6	
dx_7**contains intercept terms and time trend terms**	-3.8423	-4.38	-3.60	-3.24	steady
x_8**contains intercept terms and time trend terms**	-2.2871	-4.38	-3.60	-3.24	non-stationary
x_8**contains intercept terms**	-1.4934	-3.75	-3.00	-2.63	
x_8**no intercept items and time trend items**	-0.8136	-2.66	-1.95	-1.6	
dx_8**contains intercept terms and time trend terms**	-5.2354	-4.38	-3.60	-3.24	steady

2.3. Test of Variables

2.3.1. Stability Test

When modeling the ARMA model, the time series we use needs to meet the stability, so the stability test of the sequence is conducted first. In this paper, ADF stability test is used to test the stability of the sequence. ADF stability test is to judge whether the time series has a unit root: if the time series is stable, there is no unit root; otherwise, there is a unit root. The results of the tests are shown in Table Table 5[12].

According to Table 5 above, we can see that the original sequence of the above seven variables with intercept terms and time trend terms, intercept terms, and ADF values without intercept terms and time trend terms have no critical point of less than 1%. Therefore, according to the results of the test of ADF, the original sequence is not stable, but after the first order difference, we can find that the ADF value of the intercept term and the first order difference are less than the critical point of 1%. Because of this, we can conclude that the above seven variables are stationary sequences after the first order difference, the$I(1)$ sequence. All seven sequences are first-order stationary sequences, the same order of single integration, can be co-integrated test.

2.3.2. Co-Integration Test

Co-integration is a statistical description of the long-term equilibrium relationship of non-stable economic variables. The primary task of co-integration test is to test the single integration of time series, that is, to test whether a non-stable sequence can become a stable sequence after the difference; the purpose of co-integration test is to test whether there is a long-term stable relationship between non-stable time series. In other words, the role of the cointegration test is to test whether there is a pseudo-regression between the variables of the regression equation that we obtained before[5].

Co-integration test usually uses two methods: EG two-step co-integration test and Johansen co-integration test. The EG two-step co-integration test is actually the unit root test of the residuals of the fitted regression model.

In this paper, EG two-step method is used to test.The first step is a multiple linear regression model, and the second step is the unit root test of residuals. The residuals were found to be unconstant first order autocorrelation stationary sequence, so the co-integration test was passed. Multivariate linear models can be performed without pseudo-regression.

Table 6. The ADF test for the residuals

	ADF value	**1%**	**5%**	**10%**	**Smooth type**
Residues contain an intercept term and a time-trend term	-2.5691	-4.38	-3.60	-3.24	non-stationary
The residue contains an intercept term	-2.7749	-3.75	-3.00	-2.63	non-stationary
Residues do not contain intercept terms and time trend terms	-2.9596	-2.66	-1.95	-1.6	steady

According to Table 6, we clearly see about the residual test results of formula (5), the residual with the ADF value of intercept and time trend term greater than 1% critical point, the residual with intercept ADF value is also greater than 1% critical point, but the residual without intercept and time trend term ADF value is less than 1% critical point, we can conclude that the residual sequence (5) meet the ADF test, the formula (5) multiple linear regression model through the co-integration test.

2.4. The ARMA Model for the Predictor Variables

2.4.1. White Noise Test on the Variables

In the first place,a brief introduction of the white noise sequence:

The popular explanation of white noise is that there is no information in the accepted signal and all white noise. The characteristic of the white noise sequence is that the random variables of any two time points in the sequence are not correlated, and no dynamic law can be found in the sequence. Due to this characteristic, the white noise sequence cannot predict in the future with previous data.

If the time series$\{\varepsilon_t\}$ meets:

1） $E(\varepsilon_t) = 0$， $Var(\varepsilon_t) = \sigma^2$,

2） $Cov(\varepsilon_t \varepsilon_s) = 0$， That is, for arbitrary$s \neq t$, $\varepsilon_t$and $\varepsilon_s$irrelevant.

This is called$\{\varepsilon_t\}$ a white noise sequence.

Not all stationary sequences are worth modelling.Only those sequences that exhibit strong and meaningful correlations between their values, and where historical data have a certain impact on future developments, are worth our time to explore valuable information from historical data for predicting the future development of the series.

If the sequence values do not have any correlation with each other, it means that the sequence is a sequence without memory, and the past behavior has no effect on future development, and this sequence is called pure random sequence. From the perspective of statistical analysis, pure random sequences are sequences without any analytical value.

We conducted the white noise test for the above eight predictors respectively, and found that only the first order difference predictor variables, x_1,$x_2$andx_8 are non-pure random sequence. Therefore, we build ARMA models for these three variables respectively, and show the specific analysis in the subsequent sections.

2.4.2. Models for the Predictor Variables x_1

For time series building ARMA model, we first require the time series is a smooth sequence. Through the difference in the previous order, we can know the predictor variablex_1ADF test found that the original sequence is non-stationary, but the prediction variablex_1, after the first order difference sequence is smooth sequence. We constructed the ARMA model by considering the sequence after the differencedx_1,and restore the difference after the equation, get the equation about the original sequence.

For the above obtained stationary sequence, the white noise test is required. If the sequence is white noise sequence, the above characteristics of the white noise sequence show that the previous data have no influence on the future prediction, so there is no further analysis necessary; If the sequence is non-white noise sequence, the ARMA(p,q) model can be used for the prediction.

We give autocorrelation and partial autocorrelation maps after the first order difference, as shown in the following below.

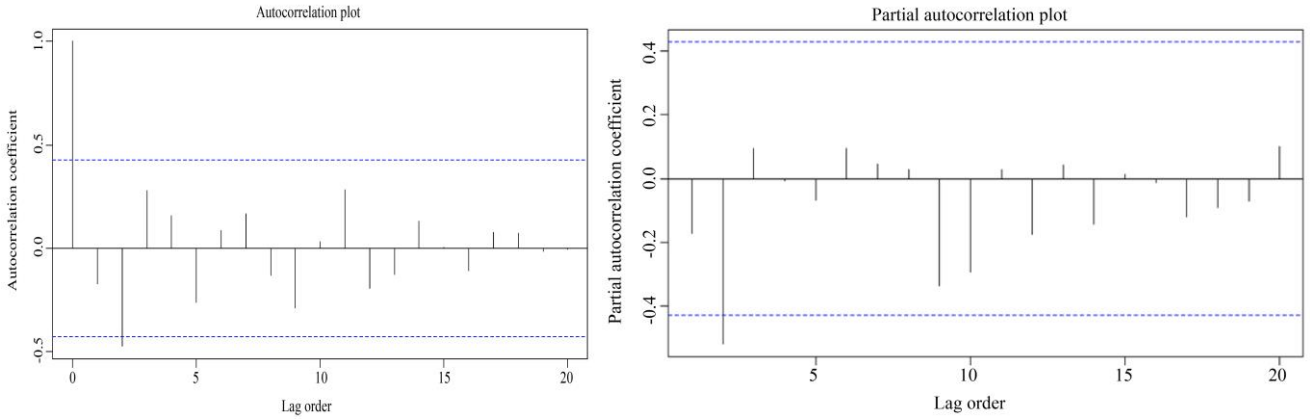

Fig. 4 dx_1autocorrelation and partial autocorrelation graphs

It can be seen from the figure that the first order difference is stable, and the white noise test is conducted on the sequence after the differencex_1, and the p-value is found that the p-value of the difference sequence is less than 0.05.We can conclude that the sequence after the first order difference is a non-white noise sequence, which can be studied later. We can build the AR (1) model by adopting the minimum AIC criterion[8].

Calculate the data, and get the equation restored after the first order difference is:

$$x_t = 0.0196 + 0.7929x_{t-1} + 0.2071x_{t-2} + \varepsilon_t \quad (2)$$

Below we make white noise diagnosis of residual, and the results are shown in the following figure.

For the below figure, through the observation of the lower left figure, we can see that each point is on the dotted line, and the p-value of the statistic is significantly greater than 0.05 (the dotted line is 0.05 reference line). We can think that the residual sequence of this fitted model belongs to the white noise sequence, that is, the significance of the fitted ARMA model is established[9].

We learned that about building type (2) of ARMA model residual meet the white noise test, we are about the model coefficient test, we found through R software test, meet the coefficient test the p value is less than 0.05, so we can conclude that the model meet the model coefficient test, can predict the future value.

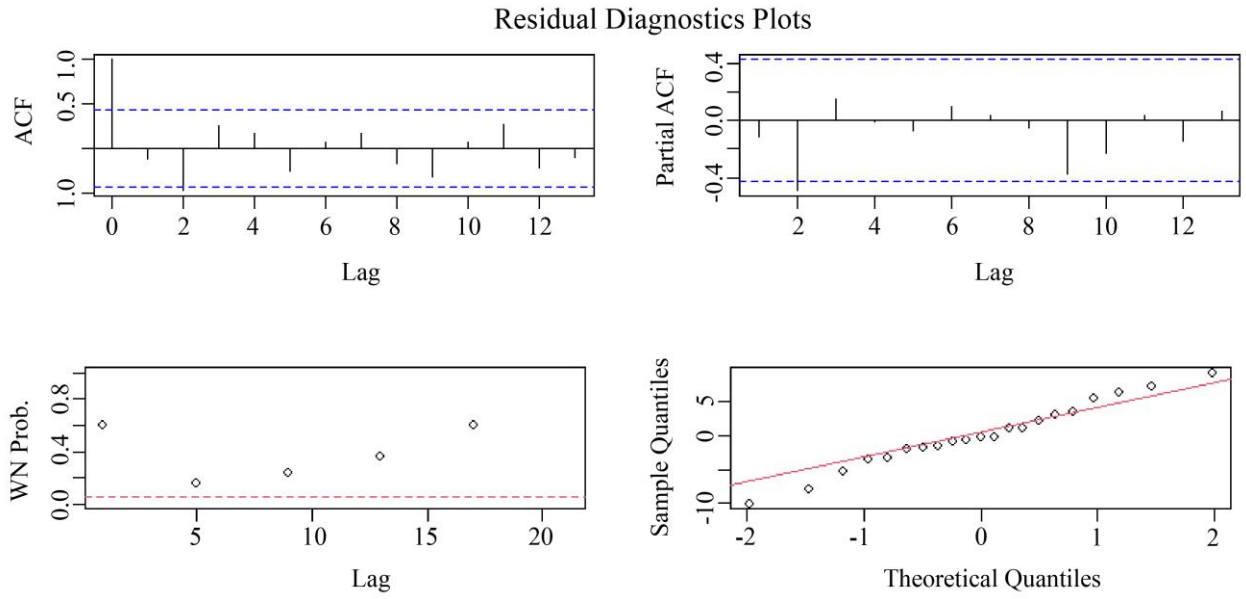

Fig. 5 White noise diagnosis of x_1 residuals

The constructed model (2) meets the residual white noise test and the significance test of the model coefficient. We can use the model (2) to predict the future value. We use the model (2) to predict the value-added ratio of agriculture, forestry, animal husbandry and fishery in the next 5 years. The details are shown in the table below.

Table 7. Relevant values of agriculture, forestry, animal husbandry and fishery in the next five years

Year	First order difference results%	Restore the results%	Forecast value (RMB 100 million)
2022	-2.07	4.51	90667.7
2023	0.45	4.96	95164.8
2024	-0.07	4.89	99818.4
2025	0.03	4.92	104729.5
2026	0.02	4.94	109903.1

From the relevant information in Table 7, we can see that the value-added ratio of agriculture, forestry, animal husbandry and fishery in the next five years is greater than zero, indicating that the total output value of agriculture, forestry, animal husbandry and fishery in the next five years shows an overall upward trend. However, the value-added ratio of agriculture, forestry, animal husbandry and fishery in 2022 has decreased by 2.07 compared with 2021, and the growth rate has decreased. Nonetheless, in the four years after 2022, we can find that the value-added ratio of agriculture, forestry, animal husbandry and fishery is stable around 4.9, and the growth rate is relatively stable. We can know that the growth rate of China's agriculture, forestry, animal husbandry and fishery in the next five years is roughly similar, but the growth rate is only about 4.9%.We know that China is a major agricultural country, and its population ranks among the highest in the world. The annual grain production and consumption in China are enormous. However, with the development of science and technology, China's food output has been able to meet the needs of the Chinese people and meet the requirements of a well off society in an all-round way.In recent years, the main direction of China is the development of industry and science and technology, and the demand for agriculture, forestry, animal husbandry and fishery has decreased, which may lead to the relatively slow growth rate of China's lunar animal husbandry and fishery. However, it is worth celebrating that China's agriculture, forestry, animal husbandry and fishery are still on the rise trend, which can also drive the development of China's total GDP and contribute to China's socialist modernization.

2.4.3. Models for the Predictor Variables x_2

Through the analysis of the previous section, we can know, through the predictor variable x_2 ADF test found that the original sequence is non-stationary, but the prediction variable x_2, after the first order difference sequence is smooth sequence.We constructed the ARMA model by considering the sequence after the difference dx_2, and restore the difference

after the equation, get the equation about the original sequence.

The autocorrelation plots and the partial autocorrelation plots after the first order difference are given, as shown in the following figure.

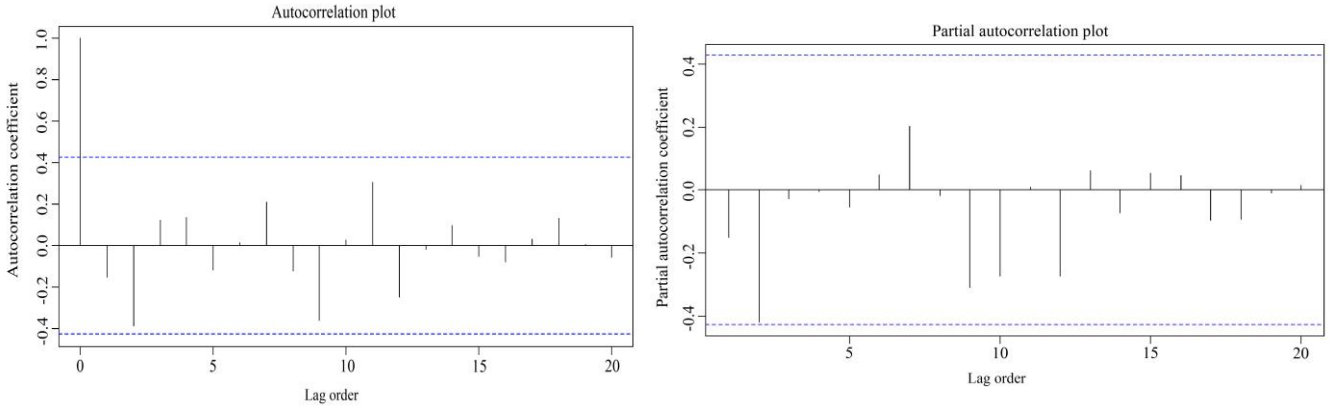

Fig. 6 dx_2autocorrelation and partial Autocorrelation Graphs

We know that the first order difference of sequence x_2 is stable, so we test the white noise, and then the p value is obtained. We find that the p value of the sequence x_2 is less than 0.05. We can conclude that the sequence after the first order difference is a non-white noise sequence, which can be further studied. We can build the AR (1) model by adopting the minimum AIC criterion.

Calculate the data, and get the equation restored after the first order difference is:

$$x_t = 0.2201 + 0.7916x_{t-1} + 0.2084x_{t-2} + \varepsilon_t \tag{3}$$

The results of white noise diagnosis of residuals are shown in Figure 7.

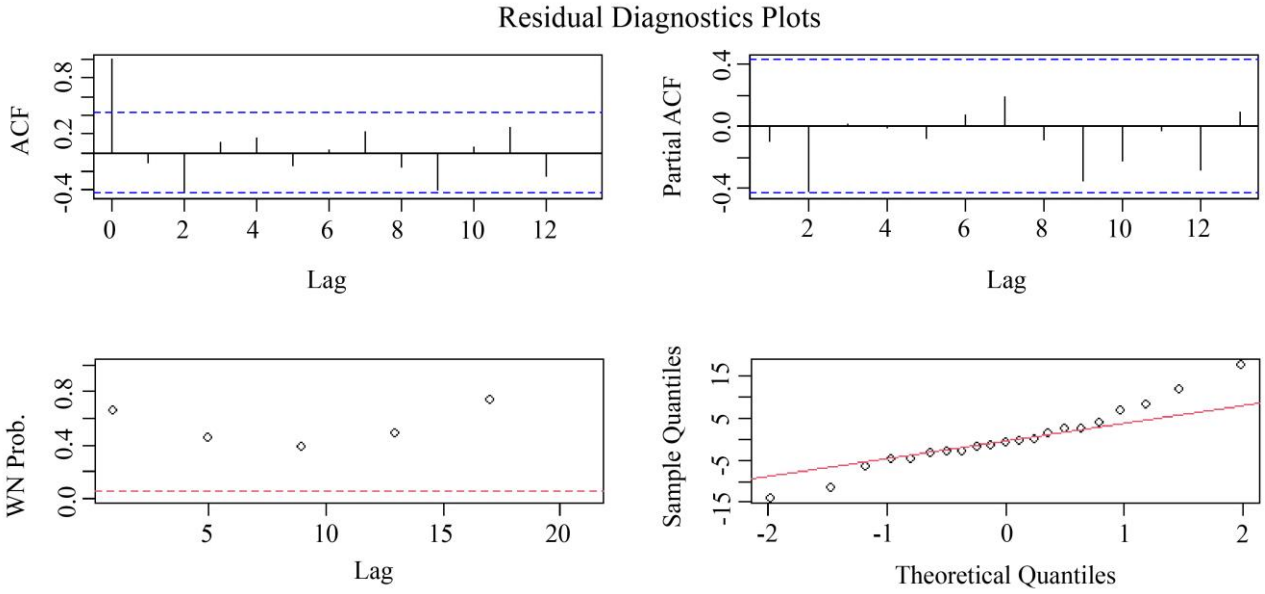

Fig. 7 White noise diagnosis of $x_2$residuals

We learned that about the (3) ARMA model residual meet the white noise test.Next, We will conduct the relevant model coefficient test. Through the analysis using R software, we find that the coefficient tests satisfy the condition of having values less than 0.05. Therefore, we can conclude that the constructed model meets the criterion for model coefficient significance and is suitable for predicting future values.

The constructed model (3) meets the residual white noise test and the significance test of the model coefficient. We can use the model (3) to predict the future value. We use the model (3) to predict the specific value-added ratio for the industry over the next five years, as shown below.

Table 8. Relevant values of industry in the next five years

Year	First order difference results%	Restore the results%	Forecast value (RMB 100 million)
2022	-3.63	15.44	430100.9
2023	1.02	16.46	500895.5
2024	0.05	16.51	583593.3
2025	0.25	16.76	681430.5
2026	0.21	16.97	797069.3

From Table 8, we can see that in the second column of Table 8, the first order difference of 2022 is-3.63, indicating that the industrial value added ratio of 2022 is lower than that of 2021, the first order difference of the remaining four years are greater than 0, indicating that the industrial value added ratio in 2023-2026 industrial value added ratio increased year by year, but the increase is not very large. Through the analysis of the third column of the table 8, China's industrial value added ratio is greater than zero, that our country's industrial output value is increasing every year, and our country industrial value added than more than 15%, there is a big growth trend.It shows that the growth rate of China's total industrial output value is faster, which is far greater than the value-added ratio of agriculture, forestry, animal husbandry and fishery in China. However, we find that from 2021 to 2022, the value increase ratio will have a significant downward trend. One likely reason is that the epidemic has brought a huge impact on all industries in China in the past three years, leading to an obvious downward trend of industrial GDP in 2022.China is a big agricultural country, but in recent decades, China has been committed to improving China's industrial level, the national government invested a lot of experience to develop industrial technology, so that China's industrial GDP has been showing a trend of rapid growth [6]. Industry is a concrete display of the country's hard power, our country should vigorously develop industrial technology, enhance China's industrial level, so that the industrial level can exceed the level of developed countries as soon as possible, as soon as possible to realize China's industrial modernization.

2.4.4. Models for the Predictor Variables x_8

Through the analysis of the previous section, we can know, through the predictor variablex_8ADF test found that the original sequence is non-stationary, but the prediction variable x_8, after the first order difference sequence is smooth sequence.We constructed the ARMA model by considering the sequence after the differencedx_8,and restore the difference after the equation, get the equation about the original sequence.

The autocorrelation plots and the partial autocorrelation plots after the first order difference are given, as shown in the following figure.

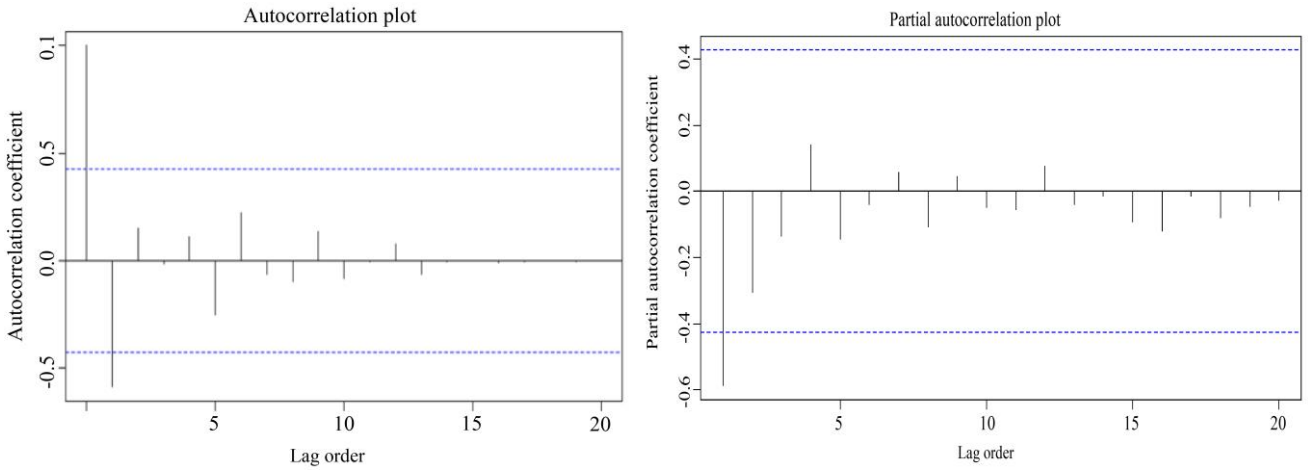

Fig. 8 dx_8autocorrelation and partial autocorrelation graphs

We know that the first order difference of sequencex_8 is stable, which we test with white noise, and the p value is obtained. We find that the p value of the sequencex_8 is less than 0.05. We can conclude that the sequence after the first order

difference is a non-white noise sequence, which can conduct subsequent study. We can build the AR (1) model by adopting the minimum AIC criterion.

Calculate the data, and get the equation restored after the first order difference is:

$$x_t = -0.3899 + 0.4354x_{t-1} + 0.5646x_{t-2} + \varepsilon_t \quad (4)$$

The results of white noise diagnosis are shown in Figure 9:

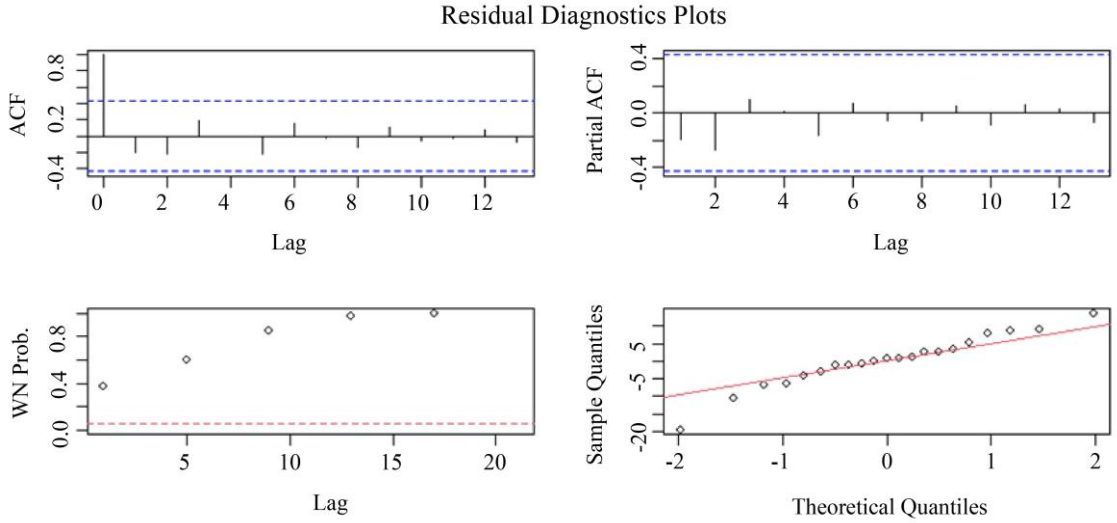

Fig. 9 White noise diagnosis of $x_8$residuals

For Figure 9 above, through the observation of the lower left figure, we can see that each point is on the dotted line, and the p-value of the statistic is significantly greater than 0.05 (the dotted line is 0.05 reference line). We can think that the residual sequence of this fitted model belongs to the white noise sequence, that is, the significance of the fitted ARMA model is established.

We learned that about build (3) ARMA model residual meet the white noise test.Next,we will conduct the relevant model coefficient test.We found through R software test, meet the coefficient test the p value is less than 0.05, so we can conclude that the model meet the model coefficient test, can predict the future value.

The constructed model (3) meets the residual white noise test and the significance test of the model coefficient. We can use the model (3) to predict the future value. We use model (3) to predict the value-added ratio of the real estate industry in the next five years, shown in the table below.

Table 9. Forecforecast of real estate industry in the next five years

Year	First order difference results%	Restore the results%	Forecast value (RMB 100 million)
2022	-1.40	4.23	80841.6
2023	0.18	4.41	84406.7
2024	-0.71	3.70	87529.7
2025	-0.20	3.50	90593.2
2026	-0.49	3.01	93320.1

We can see from table 9, for table 9 column 2, we can know that in 2022,2024,2025,2026, the first order difference results are less than 0, shows that the real estate industry in these years are lower than the previous year, but it is worth noting that the third value of table 9 is more than 0, shows that the real estate industry GDP is still the overall upward trend, but the growth ratio is worrying, growth than every year less than 5%, far lower than China's industrial growth ratio, the growth rate is slow.About the real estate industry GDP growth is slow, may be our country now rapidly aging population, birth population compared with previous years, lead to the gradual decline of the real estate industry in China's rapid development in previous years, the real estate industry is an important pillar of China's GDP growth, for China's GDP growth contributed great power,

but because now the new population decline, the aging of serious problems, lead to the real estate industry development is slow, the prospect is worrying. Moreover, due to the huge impact of the epidemic in the past three years, the income of Chinese people is unstable, and very few people spend money on the real estate industry, which is also a reason for the decline of the real estate industry. With the passing of the epidemic, the real estate industry may still have the opportunity to develop rapidly in the future.

2.4.5. Reconstruct the Multiple Linear Regression Model

For the multivariate model of the above three variables, we found that the linear model residual still holds through the co-integration test.

Table 10. The residual ADF test

	ADF值	1%	5%	10%	Smooth type
Residues contain an intercept term and a time-trend term	-2.4424	-4.38	-3.60	-3.24	non-stationary
The residue contains an intercept term	-2.6019	-3.75	-3.00	-2.63	non-stationary
Residues do not contain intercept terms and time trend terms	-4.7667	-2.66	-1.95	-1.6	steady

Therefore, ternary models can be constructed. The specific three-way linear model is as follows:

$$y = 2.3695 + 0.1917x_1 + 0.5119x_2 + 0.1646x_8 \tag{5}$$

We can get the results by substituting the values of each variable for the next 5 years.

Table 11. Relevant predicted values of GDP in the next five years

Year	The GDP value-added ratio is forecast%	GDP forecast value (RMB 100 million)
2022	11.83	1278965.8
2023	12.47	1438452.8
2024	12.36	1616245.6
2025	12.46	1817638.8
2026	12.49	2044661.9

Through inquiry, we know that the actual GDP value of 2022 is 12.1 billion yuan, which is quite different from the predicted GDP value of 12.7 billion yuan in this paper. It shows that the model of this paper has good prediction ability and can provide a certain prediction and guidance for the future development and change of China's GDP.

3. Method Introduction

3.1. ARMA Model

ARMA model is a time series analysis method proposed on the basis of AR model and MA model, which is widely used in the prediction problem[1].

① AR model

For the AR model, the mathematical expression satisfied by the time series$\{x_t\}$is the:

$$\begin{cases} x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + \varepsilon_t, \\ \phi_p \neq 0, \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t, \\ E(x_s \varepsilon_t) = 0, \forall s < t, \end{cases} \tag{6}$$

Where: $\phi_0$represents the constant of the AR model; $\phi_1, \phi_2, \ldots, \phi_p$represents the coefficient of the AR model; $x_t, x_{t-1}, \ldots, x_{t-p}$ represents sequence values of time of $t, t-1, \ldots, t-p$, respectively; ε_t represents random interference sequence value of time t;Bceause$\phi_p \neq 0$, the highest order of the AR model is p, recorded as the AR(p) model; $E(\varepsilon_t)$ represents the expectation of the random interference sequence$\{\varepsilon_t\}$; $Var(\varepsilon_t)$ represents the variance of the random interference sequence$\{\varepsilon_t\}$;$E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_s \varepsilon_t) = 0, s \neq t$ indicates that the random interference sequence $\{\varepsilon_t\}$ is zero mean white noise sequence; $E(x_s \varepsilon_t) = 0, s < t$indicates that the current random interference is unrelated to the past sequence value.

② MA model

For the MA model, the mathematical expression satisfied by the time series$\{x_t\}$is the:

$$\begin{cases} x_t = \mu + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \ldots - \theta_q\varepsilon_{t-q}, \\ \theta_q \neq 0, \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t\varepsilon_s) = 0, s \neq t, \end{cases} \tag{7}$$

Where: μ represents the constant of the moving average model; $\theta_1, \theta_2, \ldots, \theta_q$ represents the coefficient of the moving average model; $\varepsilon_t, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q}$ represents the random interference sequence value $t, t-1, \ldots, t-q$ in time. Since $\theta_q \neq 0$, therefore, the highest order of the MA model is q, denoted as the MA(q) model.

③ ARMA model

For the ARMA model, the mathematical expression satisfied by the time series $\{x_t\}$ is the:

$$\begin{cases} x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \ldots - \theta_q\varepsilon_{t-q}, \\ \phi_p \neq 0, \theta_q \neq 0, \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t\varepsilon_s) = 0, s \neq t, \\ E(x_s\varepsilon_t) = 0, \forall s < t, \end{cases} \tag{8}$$

Where: ϕ_0 represents the constant of the model; $\phi_1, \phi_2, \ldots, \phi_p$ represents the autoregressive coefficient; $\theta_1, \theta_2, \ldots, \theta_q$ represents the moving average coefficient. Since $\phi_p \neq 0, \theta_q \neq 0$, therefore, the highest order of the autoregressive part of the model ARMA is p, and the highest order of the moving average part is q, denoted as the ARMA(p,q) model. As can be seen from the above equation, the ARMA(p,q) model is built based on the AR(p) model and the MA(q) model.

Using the ARMA(p,q) model, the sequence used is stable, in other words, the affected factors are basically the same in the time frame of the study. If the given sequence is not stationary, it is essential to preprocess the sequence to make it stationary before modeling using the ARMA model[3].

3.2. ARMA Model Modeling Steps

- Preprocessing of the sequence to determine whether the sequence is a smooth and non-pure random sequence. If it is a non-smooth sequence, the sequence is processed to make it meet the condition of the model ARMA(p,q) modeling, that is, the processed sequence is a smooth non-white noise sequence;
- The sample autocorrelation coefficient (ACF) and the sample partial autocorrelation coefficient (PACF) diagram of the observed value sequence are obtained;
- Fit the sample autocorrelation coefficient and correlation coefficient of partial autocorrelation coefficient ARMA(p,q);
- Estimate the unknown parameters in the model;
- Test the validity of the model. If the fitting model fails to pass the test, turn to step 3, re-select the model for fitting;
- Model optimization. If the fitted model passes the test, it still turns to step 2, fully considers various possibilities, build multiple fitted models, and select the optimal model from all the models that pass the test;
- Use the fitted model to predict the future trend of the sequence.

3.3. Multivariate Linear Regression Model

Multiple linear regression can reflect the correlation between variables, which assumes a linear correlation between dependent and independent variables, uses a linear regression model to fit the data of dependent and independent variables, and determines the model parameters through the least squares method to obtain the regression equation. From the regression equation, we can predict the for the future data[8].

The mathematical expression of the multivariate linear regression is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon, i = 1,2, \cdots n. \tag{9}$$

Where: y is a response variable, $x_1, x_2, \ldots, x_n$ is a measurable n predictor variable, β_0 is a constant term, β_i is a partial regression coefficient, $i = 1,2, \ldots, n, \varepsilon$ is a random error term. The above formula indicates that the change of dependent variable y is affected by two parts: first, the linear change part affected by the change of independent variables x; second, the change part affected by other random factors ε.

Before determining the model parameters, univariate regression analysis and multicollinearity analysis are needed on the independent variables. After selecting the independent variables, we start to determine the model parameters, and then get the regression equation.

The indicators to measure GDP should be selected according to the principles of science, comprehensiveness and data reliability. GDP of a country depends on its national income, the total production of agriculture and industry, as well as the overall consumption. This article selects nine variables to analyze the data: GDP contribution ratio, agricultural and forestry industry contribution ratio, industrial contribution ratio, construction industry contribution ratio, wholesale and retail industry contribution ratio, transportation (storage and postal services) contribution ratio, accommodation and catering industry contribution ratio, financial industry contribution ratio, and real estate industry contribution ratio[4].

4. Conclusion

Data in this paper through the China statistical yearbook, China GDP value-added ratio as the response variable, select eight predictive variables for empirical analysis, using ARMA model and multiple linear regression model, explore the influencing factors affecting the development of GDP, and the application of ARMA model, predict the possible value of predictive variables in the next five years, constructed into the multiple linear regression model, get the response variable GDP value-added value in the next five years.

From the above four tables in Tables 7,8,9 and 10, we can know that agriculture, forestry, animal husbandry, fishery, industry, real estate and GDP will continue to increase in the next five years. We can intuitively see that the development speed of the four variables is different according to the value-added ratio of each variable. Our primary goal is GDP growth, with value-added ratio of over 10% for the next 5 years. The pace of development is rapid, and it is possible that in the coming decades, we may surpass the GDP levels of developed countries.We select influence GDP value-added ratio of eight predictor variables, through univariate regression analysis, verified the eight predictor variables have significant influence on the response variables, then we through gradual regression, build a multiple linear model, but because there are three predictor variables can not meet the white noise sequence test, we again through the white noise sequence of three multivariate linear model (5), found that it can still through the co-integration test, so we think the construction of multivariate linear model (5) is meaningful. We constructed models for the three predictive variables to predict the GDP in the next 5 years and put them into the model (5) to obtain the forecast value of China's GDP in the next 5 years. We combined the ARMA model with the multivariate linear regression model, and applied this method to the application analysis of practical problems, which can well predict the future values of the corresponding variables.We collected China's GDP data for 2022 and found that the difference between the actual and predicted values was very small, indicating that the model can predict well and that the method has good feasibility in practice. Similarly, the method can be applied to other scenarios to provide statistical solutions to practical problems.

References

[1] R.I. Jennrich, and P.F. Sampson, “Application of Stepwise Regression to Non-linear Estimation,” *Technometrics*, vol. 10, no. 1, pp. 63-72, 1968. [Google Scholar] [Publisher Link]

[2] H.L. Gray, G.D. Kelley, and D.D. Mc Intire, “A New Approach to ARMA Modeling,” *Communications in Statistics-Simulation and Computation*, vol. 7, no. 1, pp. 1-77, 1978. [CrossRef] [Google Scholar] [Publisher Link]

[3] Leland Wilkinson, “Tests of Significance in Stepwise Regression,” *Psychological Bulletin*, vol. 86, no. 1, pp. 168-174, 1979. [CrossRef] [Google Scholar] [Publisher Link]

[4] Elias Samaras et al., “AARMA Representation of Random Processes,” *Journal of Engineering Mechanics*, vol. 111, no. 3, 1985. [CrossRef] [Google Scholar] [Publisher Link]

[5] Ludwig Fahrmeir, and Gerhard Tutz, *Multivariate Statistical Modelling Based on Generalized Linear Models*, New York: Springer-Verlag, 1994. [Google Scholar]

[6] Jin-Li Hu, and Cheng-Hsun Lin, “Disaggregated Energy Consumption and GDP in Taiwan: A Threshold Co-integration Analysis,” *Energy Economics*, vol. 30, no. 5, pp. 2342-2358, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[7] Colin M. Beale et al., “Regression Analysis of Spatial Data,” *Ecology Letters*, vol. 13, no. 2, pp. 246-264, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[8] John Fox, and Sanford Weisberg, “Multivariate Linear Models in R, An R Companion to Applied Regression,” *Los Angeles: Thousand Oaks*, 2011. [Google Scholar]

[9] Byoung Seon Choi, *ARMA Model Identification*, Springer Science & Business Media, 2012. [Google Scholar] [Publisher Link]

[10] Nelson Fumo, and M.A. Rafe Biswas, “Regression Analysis for Prediction of Residential Energy Consumption,” *Renewable and Sustainable Energy Reviews*, vol. 47, pp. 332-343, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[11] L. L, "Application of Time Series Analysis in GDP Prediction in China," *Time Finance*, vol. 670, no. 24, pp. 20-24, 2017.

[12] H.R. Zhao, "Measurement Analysis of Factors Influencing GDP in China," *China Collective Economy*, vol. 686, no. 30, no. 12-14, 2021.

[13] Huanyu Zheng, Malin Song, and Zhiyang Shen, "The Evolution of Renewable Energy and Its Impact on Carbon Reduction in China," *Energy*, vol. 237, p. 121639, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Chunmiao Jia et al., "Prediction of Railway Freight Volume in Ganning District based on ARMA Model and Multiple Regression," *Integrated Transportation*, vol. 44, no. 9, pp. 147-154, 2022. [Google Scholar]