Efficacy Evaluation of Statistical Prediction Models: A Comparative Analysis based on ARIMA Model, Grey Model and Polynomial Regression Model

Haoru Du

Meihua school, Suzhou, 215000, China

Abstract. This paper uses ARIMA model, grey model and polynomial regression model to estimate and forecast two important indicators of China's real GDP growth rate and consumer price index (CPI). The results show that the sequence predicted by the polynomial regression model has the highest degree of agreement with the actual value, and has the lowest prediction error and the best prediction performance, while the other two types of models are not suitable for long-term prediction.

Keywords: ARIMA Model; Grey Model; Polynomial Regression Model; Efficacy Evaluation.

1. Introduction

Statistical forecasting refers to a statistical technique that uses the existing trends of the development of things to simulate, predict their future development trends, and identify the changing laws and changing trends. There are many kinds of existing statistical prediction models, and the mainstream methods include ARIMA model, grey model and polynomial regression model. From the research purpose, all three types of models can achieve the effect of statistical prediction. From a statistical point of view, there is a big gap in their prediction performance and scope of application. ARIMA model (Autoregressive Integrated Moving Average Model) is a prediction model based on time series proposed by Box & Jenkins (1970), which estimates its future behavior trend through the characteristics of past behavior in time series, which belongs to a linear Time series forecasting models. Grey model is a model that predicts unknown information based on known information, and makes mathematical inference of unknown information by extracting the general laws of known information as effective components (Deng, 1982). Polynomial Regression Model (Polynomial Regression Model) is an extended form of simple linear regression model, which aims to incorporate explanatory variable sequences of different orders into the model at the same time, and determine the order through the principle of minimum residual error. The predictive power of linear regression. There are many empirical cases using these types of models for applied research in the research community. Aslanargun *et al.* (2007) used the ARIMA model to dynamically predict the number of tourists to Turkey; Liu *et al.* (2010) based on similarity and proximity. From different perspectives, the gray model is extended to a gray similarity correlation model and a gray proximity correlation model to predict the trend of series from two different perspectives, as well as the relationship and influence between series; Tao & Cao (2020) Polynomial regression model. The predictive ability of the model was tested, and the extension of the polynomial regression model with moderator variables and mediator variables was discussed. Finally, an empirical demonstration was carried out through an example. In view of the above research background, this paper will use the above three types of models to predict the series based on China's economic output (GDP) and price index series (CPI), compare the pros and cons of different models through feature observation and error analysis, and then summarize them. The scope of application and application scenarios can provide empirical reference for future statistical research.

2. Comparison of Forecasting Ability of Different Time Series Forecasting Models

In this paper, the real GDP growth rate, which measures the economic growth rate, and the Consumer Expenses Price Index (CPI), which measures the inflation level, are selected as the data

series for testing the prediction model. The sample period is from 1953 to 2021. The data comes from the official website of the National Bureau of Statistics of China (http://www.stats.gov.cn/).

2.1 Model Prediction

1. ARIMA Model

The full name of the ARIMA model is the differential autoregressive moving average model. It starts from the time series itself, establishes a corresponding model for analysis, draws relevant conclusions about its past behavior, and predicts and infers its future behavior. Because of its thorough theoretical analysis and simple and effective application analysis, it is one of the main tools for linear time series forecasting. The basic idea and principle of the ARIMA model is to approximate a stationary time series with a mathematical model. Once the model is identified, it can predict future values of the time series based on past and present values. The ARIMA model consists of three parts: autoregressive process, moving average process, and single integral. The model expression is as follows:

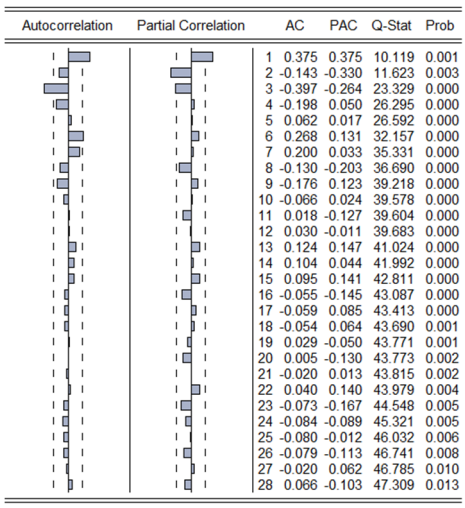
$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \quad (\phi_p \neq 0,\ \theta_q \neq 0) \tag{1}$$

Below is the model prediction process. First, judge whether the sequence is monolithic. Here, the ADF unit root test method is used to judge the stationarity of the sequence. The test results are shown in Table 1. It is not difficult to find that both the real GDP growth rate series and the CPI series passed the unit root test at the 99% significance level, proving that the two series are stationary and can be directly used for model prediction.

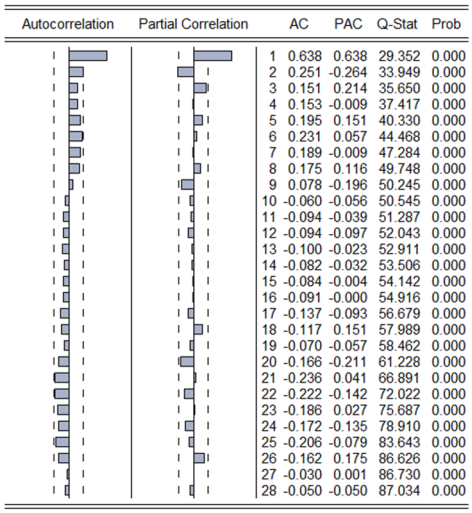
Table 1. ADF unit root test results

		Real GDP Growth Rate		CPI	
		t-Statistic	Prob.*	t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-6.348732	0.0000	-4.460380	0.0006
Test critical values	1% level	-3.531592		-3.531592	
	5% level	-2.905519		-2.905519	
	10% level	-2.590262		-2.590262	

Second, construct an autoregressive process and a moving average process. This step needs to follow the autocorrelation plot (ACF) and partial autocorrelation plot (PACF) setup, as shown in Figure 1. According to the censoring of the figure, the two models can be set as ARIMA (1,0,1) and ARIMA (1,0,3).



(a) Real GDP Growth Rate



(b) CPI

Figure 1. Autocorrelation plot (ACF) and partial autocorrelation plot (PACF) of two time series

Finally, the estimation and prediction are completed based on the set model, and the parameter estimation results are shown in Table 2:

Table 2. ARIMA model estimation results

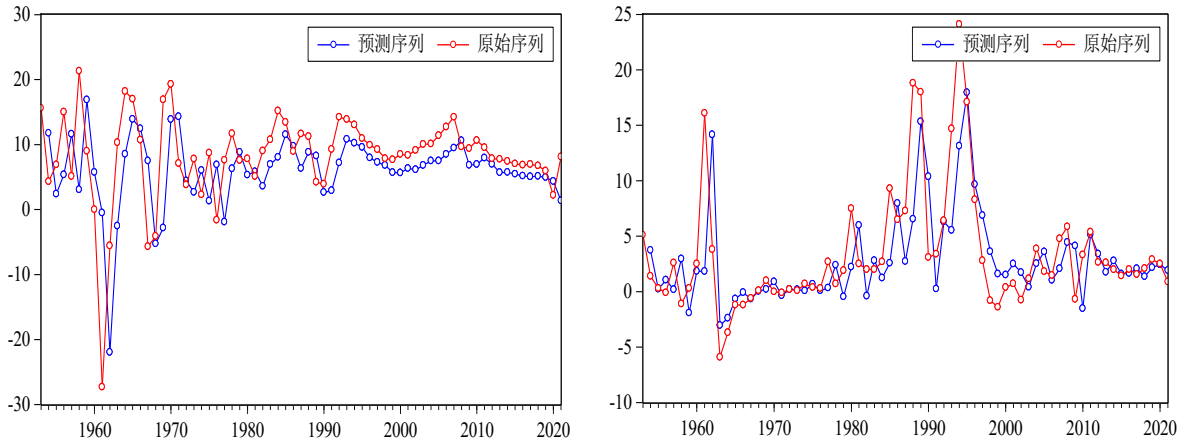
Real GDP Growth Rate				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(1)	0.715832	0.119486	5.990929	0.0000
MA(1)	0.090225	0.116473	0.774640	0.4413
CPI				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(1)	0.950968	0.108778	8.742308	0.0000
MA(1)	-0.040807	0.165008	-0.247303	0.8055
MA(2)	-0.490826	0.190988	-2.569929	0.0125
MA(3)	-0.136932	0.144275	-0.949099	0.3461

According to the parameter estimation results, the prediction models shown in equations (2) and (3) can be obtained:

$$GDP_t = 0.716GDP_{t-1} + \varepsilon_t - 0.090\varepsilon_{t-1} \tag{2}$$

$$CPI_t = 0.951CPI_{t-1} + \varepsilon_t + 0.040\varepsilon_{t-1} + 0.491\varepsilon_{t-2} + 0.137\varepsilon_{t-3} \tag{3}$$

The predicted sequence and the original sequence estimated by the ARIMA model are shown in Figure 2:



(a) Real GDP Growth Rate (b) CPI

Figure 2. ARIMA model prediction results

2. Grey Model

The grey model is referred to as the GM model, which is to establish a grey differential prediction model through a small amount of incomplete information, and to make a fuzzy long-term description of the development law of things. If a system has the ambiguity of hierarchical and structural relationships, the randomness of dynamic changes, and the incompleteness or uncertainty of index data, these characteristics are called gray. A system with grayness is called a gray system. Grey models aim to predict future trends from known information. Taking the typical gray model GM (1,1) as an example, let the original sequence $X^{(0)} = (x^{(0)}(1), x^{(0)}(2), \ldots, x^{(0)}(n))$, $x^{(0)}(k) \geq 0$, $k = 1, 2, \ldots, n$. An accumulation sequence of the original sequence can be written as $X^{(1)} = (x^{(1)}(1), x^{(1)}(2), \ldots, x^{(1)}(n))$, where: $x^{(1)}(k) = \sum_{i=1}^{k} x^{(0)}(i)$, $k = 1, 2, \ldots, n$.

On the basis of the original sequence, generate the nearest neighbor mean equal weight sequence of $X^{(1)}$, referred to as $Z^{(1)} = (z^{(1)}(2), z^{(1)}(3), \ldots, z^{(1)}(n))$, where: $z^{(1)}(k) = 1/2\ (x^{(1)}(k)+x^{(1)}(k\text{-}1))$, $k = 1, 2, \ldots, n$.

Next, according to the grey system theory (Deng, 1982), a whitening differential equation for $x^{(1)}$ with respect to time t is established, that is, the GM (1,1) model $dx^{(1)}/dt + ax^{(1)} = b$, where: a, b are parameters to be determined. Using the least squares method to solve $(a, b)^T=(B^T B)^{-1}B^T Y_N$, and

$$B=\begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ \vdots & \vdots \\ -z^{(1)}(n) & 1 \end{bmatrix},\quad Y_N=\begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \vdots \\ x^{(0)}(n) \end{bmatrix}$$

After solving a and b, the time response formula of $x^{(1)}(k)$ can be obtained:

$$\hat{x}^{(1)}(k+1)=\left[x^{(0)}(1)-\frac{b}{a}\right]e^{-ak}+\left|\frac{b}{a}\right. \quad (k=0,1,\cdots n) \tag{4}$$

After cumulative reduction, we can get:

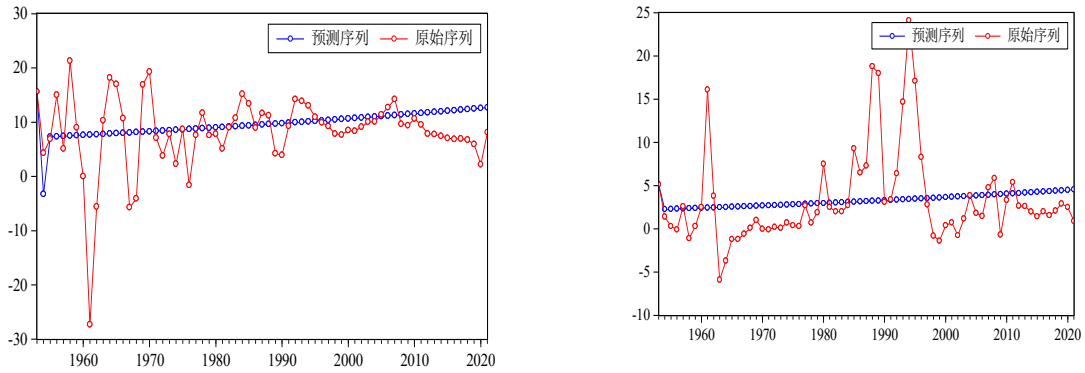
$$\hat{x}^{(0)}(k+1)=\hat{x}^{(1)}(k+1)-\hat{x}^{(1)}(k)=\left(1-e^{-a}\right)\left[x^{(0)}(1)-\frac{b}{a}\right]e^{-ak} \quad (k=0,1,\cdots,n-1) \tag{5}$$

The predicted formula obtained after the equation operation is:

$$\hat{GDP}^{(1)}(k)=(15.6+853.318)e^{0.008(k-1)}-853.318 \tag{6}$$

$$\hat{CPI}^{(1)}(k)=(5.1+212.918)e^{0.010(k-1)}-212.918 \tag{7}$$

The grey model can be used to predict the two sequences, and the results are shown in Figure 3:



(a) Real GDP Growth Rate (b) CPI

Figure 3. Prediction results of the grey model

3. Polynomial Regression Model

Linear regression studies the regression problem between a dependent variable and an independent variable, and studies a polynomial regression analysis method between a dependent variable and one or more independent variables, which is called a polynomial regression model. A polynomial regression model refers to a regression where the regression function is a polynomial of the regressor variables. The polynomial regression model is a type of linear regression model, in which the regression function is linear with respect to the regression coefficients. Since any function can be approximated by a polynomial, polynomial regression is widely used. The model expression is as follows:

$$y_t=\beta_0+\beta_1 x_i+\beta_2 x_i^2+\cdots+\beta_n x_i^n+\varepsilon_i \quad (i=1,2,\cdots,m) \tag{8}$$

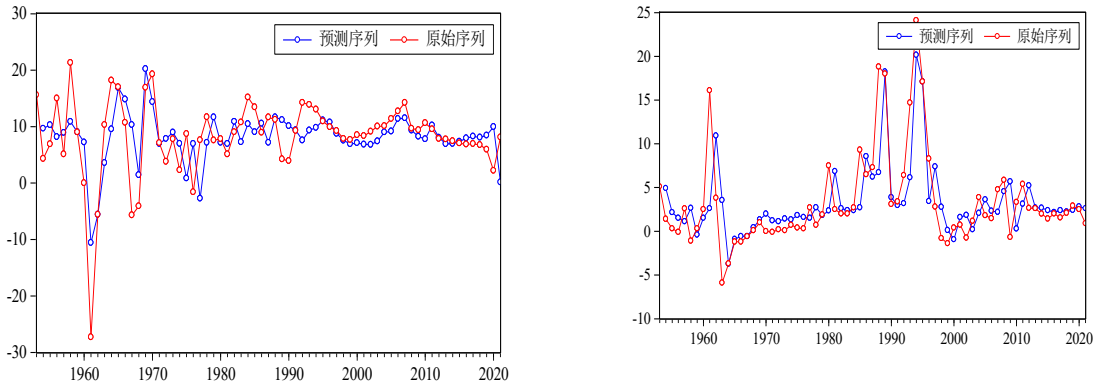
The biggest advantage of polynomial regression is that the measured points can be approximated by increasing the high-order term of x until the optimal state. In fact, polynomial regression can handle quite a class of nonlinear problems, and it plays an important role in regression analysis because any function can be approximated by a polynomial piecewise. Therefore, in common practical problems, regardless of the relationship between dependent variables and other independent variables, we can use polynomial regression models for predictive analysis.

Below is the estimation process. First, the independent variables of different orders are substituted in turn, and the order of the polynomial regression model is determined according to the principle of minimum residual error and maximum goodness of fit. Referring to the practice of Li & Liu (2019), the prediction formula after excluding items with coefficients less than 0.001 is:

$$GDP = -10.570 + 0.307x + 3.079x^2 - 0.311x^3 - 0.102x^4 + 0.016x^5 \tag{9}$$

$$CPI = 1.230 + 1.101x - 0.380x^2 + 0.016x^3 + 0.033x^4 - 0.006x^5 \tag{10}$$

The polynomial regression model can be used to predict the two sequences, and the results are shown in Figure 4:



(a) Real GDP Growth Rate (b) CPI

Figure 4. Prediction results of polynomial regression model

2.2 Model Evaluation

In order to compare the forecasting accuracy of the three types of forecasting models and clarify the applicability of different forecasting models, we calculated the error metrics to measure the accuracy of the forecasting models. Error indicators mainly include two categories: one is the root mean square error (*RMSE*), which is the square root of the ratio of the square of the deviation between the predicted value and the true value to the number of observations n, which can well describe the accuracy of the prediction; the second is the average absolute percentage error (*MAPE*), which can be used to compare predictions of different proportions, is one of the most popular metrics for evaluating prediction performance, because its calculation process is intuitive and easy to compare and interpret. The calculation expressions of the two indicators are as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(A_t - F_t\right)^2} \tag{11}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{12}$$

Among them, A_t represents the actual value, and F_t represents the predicted value. After calculation, the forecast error results of each statistical forecast model for economic indicators are shown in Table 3:

Table 3. Comparison of prediction errors (*RMSE*, *MAPE*)

Real GDP Growth Rate			CPI		
Predictive Model	*RMSE*	*MAPE*	Predictive Model	*RMSE*	*MAPE*
ARIMA Model	6.962	63.512	ARIMA Model	3.907	90.382
Grey Model	6.997	68.149	Grey Model	5.328	346.341
Polynomial Regression Model	5.066	48.363	Polynomial Regression Model	3.577	85.628

Looking at Table 3, it is not difficult to find that the polynomial regression model has the best forecasting performance, so it can play the biggest advantage in the long-term forecasting of China's economic series. In addition, it is worth mentioning that this paper also conducts a short-term prediction comparison after the study. The study shows that the advantage of the polynomial regression model in short-term estimation will decrease, and the accuracy of the ARIMA model will increase significantly, indicating that the polynomial regression model will increase significantly. The model is more suitable for large sample prediction. Finally, the performance of the grey model in this study is unsatisfactory, because the original grey model is more suitable for the prediction of cumulative data, such as GDP, resident income, etc., and is not suitable for the series forecast with strong volatility.

3. Conclusion

This paper systematically reviews the theoretical principles, forecasting process and characteristic differences of ARIMA model, grey model and polynomial regression model, and conducts a complete empirical calculation and comparison using China's typical economic indicators-real GDP growth rate and CPI . The results show that: first, the polynomial regression model has the highest prediction accuracy, which can better restore the trend of China's real GDP growth rate and CPI, and is an important method to achieve China's economic indicators forecast; second, the gray model is not suitable for fluctuations. The prediction of the prediction model is more suitable for the prediction of the cumulative data, which provides an important empirical reference for the statistical prediction work; finally, if you want to completely distinguish the pros and cons of different prediction models, you need to start with samples of different lengths, which are relatively large. The difference between sample prediction and small sample prediction is only briefly evaluated in this paper. It is found that when the sample period is shortened, the prediction ability of the polynomial regression model will also decline. This provides mathematical revision ideas and empirical evidence for the selection and expansion of various statistical prediction models.

References

[1] Box G. E. P., Jenkins G. M. Time series analysis: forecasting and control[M]. San Francisco: Holden-Day, 1970.

[2] Deng J. The grey control system[J]. Journal of Huazhong University of Science and Technology, 1982, 10 (3):9-18.

[3] Aslanargun A, Mammadov M, Yazici B, et al. Comparison of ARIMA, neural networks and hybrid models in time series: tourist arrival forecasting[J]. Journal of Statistical Computation & Simulation, 2007, 77 (1/2):29-53.

[4] Liu S, Xie N, Forrest J. On new models of grey incidence analysis based on visual angle of similarity and nearness [J]. Systems Engineering-Theory & Practice, 2010, 30(5):881-887.

[5] Tao H, Cao W. Principle and Application of Polynomial Regression and Response Surface Analysis[J]. Statistics & Decision, 2020, 36(08):36-40.

[6] Li Z, Liu S. Prediction Comparison Based on ARIMA Model, Grey Model and Regression Model[J]. Statistics & Decision, 2019, 35(23):38-41.