

Applications of three distinct regression models in GDP predication

Tiankai Duan^{1,3}, Wenbo Niu² and Dehan Zang²

¹School of mathematics and statistics, Lanzhou University, Lanzhou, Gansu, 730000, China

²Qingdao No. 19 high school, Qingdao, Shandong, 266200, China

³320220908671@lzu.edu.cn

Abstract. This paper introduces the basic theory and formula of linear regression, multiple linear regression, and nonlinear regression. Linear regression is one of the commonly used analysis methods in statistical analysis, which can predict the trend of model data change to a certain extent. Multiple linear regression involves more variables to predict and analyze the change trend of data, and can predict the change of data more accurately. Nonlinear regression can predict the model of arbitrary relationship between variables, thus obtaining more accurate prediction data. In the selection of regression analysis method, data characteristics and problem background should be considered, and model assumptions and validation should be paid attention to ensure accuracy and reliability. In the applications, the paper discusses the application of simple linear regression to Okun's law and delves into the complex relationship between multiple variables and gross domestic product (GDP). Finally, it uses nonlinear regression equations to analyze the global inflation rate and the annual data, and proves that there is a nonlinear relationship between the two and a downward trend, which is supported by analyzing the data of Australia and Canada.

Keywords: linear regression, multiple linear regression, nonlinear regression, Okun's law.

1. Introduction

Regression model is a mathematical model that quantitatively describes statistical relationships. It can be divided into multiple linear regression models, univariate scale models, and nonlinear regression models. They are often used for various data analysis. Regression model is a predictive modeling technique that studies the relationship between the dependent variable and the independent variable. This technique is commonly used for predictive analysis, time series modeling, and discovering causal relationships between variables. The important foundation or method of regression models is regression analysis, which is a calculation method and theory that studies the specific dependency relationship between one variable and another. It is an important tool for modeling and analyzing data. There are many benefits to use regression analysis. Specifically, it indicates a significant relationship between the independent and dependent variables. It indicates the strength of the influence of multiple independent variables on a dependent variable [1]. Regression analysis also allows people to compare the interrelationships between variables that measure different scales, such as the relationship between price changes and the quantity of promotional activities. These are beneficial for market researchers, data

analysts, and data scientists to exclude and estimate the best set of variables for constructing predictive models.

The main problem of the paper is the methods and theory of the linear regression model, multiple linear regression model and nonlinear regression model, and the application of these models in gross domestic product (GDP) prediction. Three distinct examples are included in this paper [2]. Firstly, the researchers collected data on GDP growth and unemployment rates in the United States and China to test the applicability of Okun's law. But not all countries conform to this law, and Indonesia's data changes do not conform to Okun's law. It follows from this that although Okun's law is valid in most cases, there is more specific information related to this data that needs to be considered. Next, through the construction of multiple linear regression model, the effects of total import and export volume, total energy consumption, total population, and total retail sales of consumer goods on GDP are analyzed, and empirical support is provided. This paper not only provides a valuable reference for policy makers and economists, but also provides entrepreneurs and market analysts with tools for in-depth understanding of economic phenomena. Finally, through covariance analysis, it is proved that although there are some abnormal changes in some data, this data still has a real prediction effect.

The layout of the paper is following. Firstly, this paper will describe the methods and theory of these three models sequentially, which is the second part of this article. Secondly, this paper will use the three models predicting Chinese GDP with some factors. this paper will research relationship between the unemployment growth rate and GDP growth rate with linear regression model, the relationship between GDP and four factors with multiple linear regression model, and analyze the global inflation rate and the annual data with nonlinear regression model. Lastly, the paper will get the conclusion.

2. Methods and theory

2.1. Linear regression model

In data analysis, regression analysis is an important way of analyzing result. It is an effective way to measure the degree of correlation between two continuous variables while making certain predictions about patterns and trends in the data. In this section, the focus is linear regression model. It is a kind of simple linear regression model that is a powerful tool in statistical analysis and is widely applied in virous field, including economics, marketing, sociology, and medicine.

Supposed that the data in a dataset is collected in pairs, $(x_1, y_1), (x_2, y_2), (x_3, y_3) \dots \dots (x_n, y_n)$. Here, let x_1 denotes explanatory variable or independent variable and let y_1 denotes response variable or dependent variable. When explanatory variable can be thought of as a potential predictor of the response variable, and the two variables satisfy [3]

$$\hat{y} = \beta_0 + \beta_1 x. \quad (1)$$

Thus, the two variables could be defined as having a simple linear regression relationship.

According to this equation, the least square regression line, or "the line of best fit", can be fitted. In the function of the regression line, the two parameters, β_0 and β_1 , represent the intercept and the slope of the linear regression function, respectively. The least square regression line can, to a large extend, accurately describe, and predict the trend of the data.

2.2. Multiple linear regression theory

In regression analysis, if there are two or more independent variables, it can be called multiple regression. In fact, a phenomenon is usually associated with many factors, so using the optimal combination of multiple independent variables to predict or estimate the dependent variable is more realistic than using only one independent variable for prediction or estimation. Therefore, multiple linear regression has greater practical significance than univariate linear regression. This article will use multiple linear regression to estimate and predict the GDP in Chinese with the data from 2014 to 2023 from National Bureau of Standards.

The model is defined by

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k, \quad (2)$$

and it can also be defined as [4]

$$\begin{cases} Y = X\beta + \varepsilon \\ E(\varepsilon) = 0, cov(\varepsilon, \varepsilon) = \delta^2 I_n \end{cases} \quad (3)$$

This is K-ary linear regression model, and it can be abbreviated as $(Y, X\beta, \delta^2 I_n)$. In this equation,

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix}, X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_k \end{bmatrix} \quad (4)$$

There are three main problems in the process of multiple linear regression. The first is making point estimation with β and δ^2 , and making quantitative relationship between y and x_1, x_2, \dots, x_k . The second is examining the parameters of the model and the results of the model. The third is predicting the numerical value of y , which is making point (interval) estimation with y .

The first step is parameter estimation, estimating the parameter of $\beta_0, \beta_1, \dots, \beta_k$ by least square method, which is $Q = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_k x_{ik})^2$. One can properly select the β_1, \dots, β_k to make the quantity of Q least. Then, solving the equation to obtain the solution as $\hat{\beta} = (X^T X)^{-1} X^T Y$ and substituting the solved $\hat{\beta}_i, i = 0, 1, \dots, k$ to $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$. It is named as empirical regression plane equation, and $\hat{\beta}_i$ are named as empirical regression coefficient.

The second step is the examining of multiple linear regression models and regression coefficients. The first method is F -test. When condition H_0 holds, then

$$F = \frac{U/k}{Q_e/(n-k-1)} \sim F(k, n-k-1). \quad (5)$$

In this equation, H_0 means $\hat{\beta}_i = 0$, k means the number of variables, n means the total number of samples, U means regression square sum, and Q_e means residual sum of squares. In addition, $U = \sum_{i=1}^n (\bar{y} - y_i)^2$, $Q_e = \sum_{i=1}^n (y_i - \hat{y}_i)^2$. If $F > F_{1-\alpha}(k, n-k-1)$, then the condition H_0 cannot be thought established, which means there is a strong linear relationship between y and x_1, x_2, \dots, x_k . If not, then the condition H_0 can be thought established, which means there is not a strong linear relationship between y and other variables x_1, x_2, \dots, x_k . The second method is r -test. Let $R = \sqrt{\frac{U}{L_{yy}}} = \sqrt{\frac{U}{U+Q_e}}$, $0 < R \leq 1$, and R is multiple correlation coefficients between y and x_1, x_2, \dots, x_k . F can also be expressed by R as [5]

$$F = \frac{n-k-1}{k} \frac{R^2}{1-R^2}. \quad (6)$$

Therefore, one can use methods F and R to examine is equivalent.

The last step is prediction with linear regression model. The first kind of prediction method point prediction. If the regression equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$ pass the examine, for the independent variables $x_1^*, x_2^*, \dots, x_k^*$ that are defined, using $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \dots + \hat{\beta}_k x_k^*$ to predict the equation $\hat{y} = \beta_0 + \beta_1 x_1^* + \dots + \beta_k x_k^* + \varepsilon$. And \hat{y}^* is the point prediction of y^* . The second method of prediction is interval prediction. Besides, the confidence interval of $1 - \alpha$ for y is (\hat{y}_1, \hat{y}_2) , and

$$\begin{cases} \hat{y}_1 = \hat{y} - \hat{\delta}_e \sqrt{1 + \sum_{i=0}^k \sum_{j=0}^k c_{ij} x_i x_j t_{l-\frac{\alpha}{2}}(n-k-l)} \\ \hat{y}_2 = \hat{y} + \hat{\delta}_e \sqrt{1 + \sum_{i=0}^k \sum_{j=0}^k c_{ij} x_i x_j t_{l-\frac{\alpha}{2}}(n-k-l)} \end{cases} \quad (7)$$

2.3. Nonlinear regression equation theory

Nonlinear regression is a method of modeling the nonlinear relationship between a dependent variable and a set of independent variables. Unlike linear regression, nonlinear regression is not limited to estimating linear models, but can estimate models with arbitrary relationships between independent variables and dependent variables. This estimation is usually achieved with iterative estimation algorithms. Common nonlinear regression models include polynomial regression, exponential regression, logarithmic regression, etc. These models can better capture the complex relationships between data, thereby improving prediction accuracy. The formula is given by [6]

$$y = f(x, \beta) + \varepsilon \quad (8)$$

where y is the dependent variable, x is the independent variable, β is the parameter to be estimated f is the nonlinear function, and ε is the random error term. The purpose of a nonlinear regression model is to estimate the parameter β and find the best fitting function f , which minimizes the error between the predicted value and the actual value.

The intricate process of nonlinear regression often entails multiple stages, each playing a crucial role in arriving at accurate and meaningful predictions. The initial stage is model definition, where the researcher must carefully select the dependent and independent variables based on the context of the problem. This selection is crucial as it determines the nature of the nonlinear relationship being explored. For instance, in economic modeling, the dependent variable might be GDP growth, while the independent variables could include interest rates, inflation, and other macroeconomic indicators.

Once the variables are chosen, the next step is parameter estimation. This involves the application of iterative estimation algorithms to determine the optimal values for the model's parameters. These algorithms, such as the least squares method or gradient descent, iteratively adjust the parameter values to minimize the difference between the predicted and actual outcomes. This process can be computationally intensive, especially for complex nonlinear models, but it's crucial to obtaining accurate predictions.

After parameter estimation, the model undergoes rigorous testing to assess its goodness of fit and prediction ability. Goodness of fit measures how well the model explains the observed data, while prediction ability evaluates the model's performance on new, unseen data. To assess these metrics, a range of statistical tests and diagnostic tools are employed, such as the R-squared value, adjusted R-squared, and residual analysis. If the model demonstrates satisfactory goodness of fit and prediction ability, it can be deemed suitable for addressing the practical problems at hand. However, if the results are not satisfactory, the model may require refinement or re-specification, possibly through the inclusion of additional variables or the adjustment of the functional form. In conclusion, the process of nonlinear regression is a sophisticated and iterative one, requiring careful consideration of variable selection, parameter estimation, and model testing. By following this rigorous framework, researchers can develop models that provide accurate and actionable insights into complex nonlinear relationships.

3. Results and Application

3.1. Application of simple linear regression to Okun's law

The data is collected from website Macrotrends, a research platform that record statistics about economics, stock, commodities. The statistics are record in Excel spreadsheet. In this section the researcher will discuss the application of simple linear regression to Okun's law. Second-hand data were collected, including the GDP growth and unemployment rate of the United States and China, to investigate if Okun's Law is suitable for these countries [7].

As labor force is a critical factor of economic growth, intuitively, there is a certain kind of relationship between economic growth and unemployment rate. Arthur Melvin Okun proposed Okun's law, clarifying the relationship. The law indicates that for every 1% increase in the unemployment rate, a country's GDP (an effective way to measure economic growth) will roughly decrease 2%. This law seems to be an effective measurement of relationship between GDP and unemployment rate. However, for some countries with different and complicated situations, this law may not be a perfect indicator.

As for the United States, according to statistical data from 1992 to 2022, using data visualization tools, the linear regression scatter plot and the least square regression line could be shown in Figure 1.

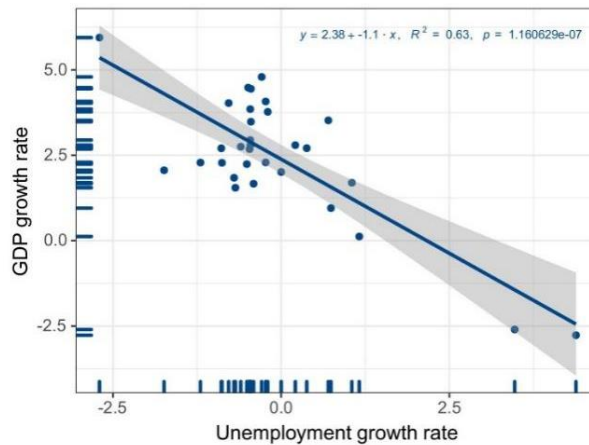


Figure 1. The relationship between the unemployment growth rate and GDP growth rate in the United States (in percent)

As the graph and the curve show, the two parameters of the function, β_1 and β_0 , are 2.38 and -1.1, respectively, indicating that in the United States, for every 1% of increase in unemployment rate, there are roughly about 1.1% decrease in GDP. This least square regression function shows that, in the United States, the relationship between unemployment rate and GDP roughly conforms to Okun's law. Moreover, as for China, according to statistical data from 1992 to 2022, using data visualization tools, the linear regression scatter plot and the least square regression line could be shown in Figure 2.

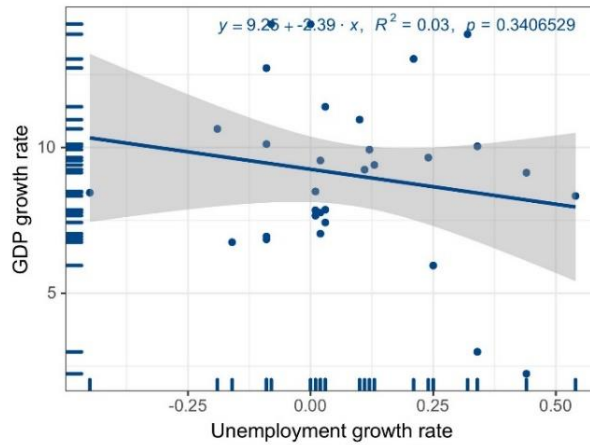


Figure 2. The relationship between the unemployment growth rate and GDP growth rate in China (in percent)

As the graph and the curve show, the two parameters of the function, β_1 and β_0 , are 9.26 and -2.39, respectively, indicating that in China, for every 1% of increase in unemployment rate, there are roughly about 2.39% decrease in GDP. This least square regression function shows that, in China, the relationship between unemployment rate and GDP roughly conforms to Okun's law. However, due to different labor force structures and different reemployment policies for the unemployed in different countries, Okun's law does not necessarily conform to the situation in all countries. Indonesia is a typical example. According to statistical data from 1992 to 2022, using data visualization tools, the linear regression scatter plot and the least square regression line could be shown in Figure 3.

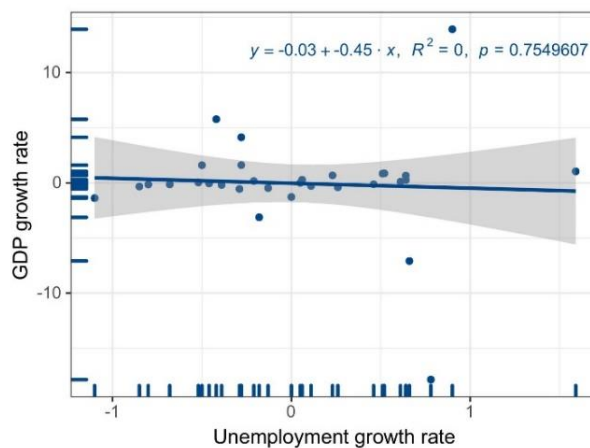


Figure 3. The relationship between the unemployment growth rate and GDP growth rate in Indonesia (in percent)

As the graph and the curve show, the two parameters of the function, β_1 and β_0 , are -0.45 and -0.03, respectively, indicating that in Indonesia, for every 1% of increase in unemployment rate, there are roughly about 0.45% decrease in GDP. This least square regression function shows that, in Indonesia, the relationship between unemployment rate and GDP doesn't conform to Okun's law. There are several possible reasons why Okun's law does not apply to certain countries. One of the reasons could be that due to differences between the ways countries measure their unemployment rate, the statistic of unemployment rate could have different standard. Another reason is that different countries, especially those less developed ones, have different labor force structure [8].

3.2. Application of multiple linear regression

This article will explore the relationship between GDP and total export-Import volume, total energy consumption, total population, and total retail sales of consumer goods, see Table 1 [9].

Table 1. Values of total export-import volume, total energy consumption, total population, total retail sales, and the GDP.

target	Total Export-Import Volume (100 million yuan) (x_1)	Total energy consumption (10 thousand tons of standard coal) (x_2)	Total population (10 thousand people) (x_3)	Total retail sales of consumer goods (100 million yuan)(x_4)	GDP (100 million yuan) (y)
2023	417568	572000	140967	471495	1260582
2022	418012	541000	141175	439732.5	1204724
2021	387415	525896	141260	440832	1149237
2020	322215	498314	141212	391981	1013567
2019	315627	487488	141008	408017	986515
2018	305010	471925	140541	377783	919281
2017	278099	455827	140011	347327	832036
2016	243387	441492	139232	315806	746359
2015	245503	434113	138326	286588	688858
2014	264242	428334	137646	259487	643563

Then getting the x_1, x_2, x_3, x_4 and Y . Supposed that $x_1 = [417568, 418012, 387415, 322215, 315627, 305010, 278099, 243387, 245503, 264242]$, then $X = [1 \ x_1 \ x_2 \ x_3 \ x_4]$. Substitute x_1, x_2, x_3, x_4 , and Y into the above equations, and the answers are as following:

$$\beta = [-5.2755 \times 10^6, 0.8297, 1.9357, 34.6123, 0.4389] \quad (8)$$

Then the fitting formula is $\hat{y} = -5.2755 \times 10^6 + 0.8297x_1 + 1.9537x_2 + 34.6123x_3 + 0.4389x_4$. The residual is thereby given by $r = [-3.8085 \ 6.7155 \ 2.4323 \ -2.6489 \ -3.2562 \ -2.1205 \ -4.1629 \ 7.5452 \ 6.7195 \ -7.4156] \times 10^3$. Using F-test and letting $\alpha=0.05$, one can get that $F = 2 \times 10^3 > 5.19 = F_{0.95}(4,5)$, so there is a strong linear relationship between y and x_1, x_2, x_3, x_4 . In addition, using r -test, and then $R=0.9994$, which is close to 1. This is also an argument to prove that there is a strong linear relationship between y and x_1, x_2, x_3, x_4 . Besides, $p = 3.3126 \times 10^{-8}$, which means there is little risk to use this model to predict the GDP.

Therefore, this section explored the relationship between GDP and total export-Import volume, total energy consumption, total population, and total retail sales of consumer goods with multiple linear regression model. From the equation $\hat{y} = -5.2755 \times 10^6 + 0.8297x_1 + 1.9537x_2 + 34.6123x_3 + 0.4389x_4$, all the four factors have a positive impact on GDP. Besides, the total population has the greatest impact on GDP. This may be helpful for simulating and predicting GDP.

3.3. Analysis and discussion of empirical results

Through the analysis of variance, the authors obtained the value of Df-Residuals. Due to the large amount of data, the sum sq of Residuals reached 325517. Therefore, the degree of freedom of data residuals 41 indicates that this set of data has a relatively ideal prediction effect and can verify the accuracy of the data. As can be seen from the data analysis chart and trend line shown in Figure 4, the global inflation rate presents a nonlinear regression downward trend during 1980-2024. Although some data show abnormal increases and decreases, the global inflation rate has dropped from 7.5% to nearly 0% overall. According to F value and $Pr(> F)$ in covariance analysis, the year has a significant impact on the inflation rate, and there is a nonlinear relationship between the two [10].

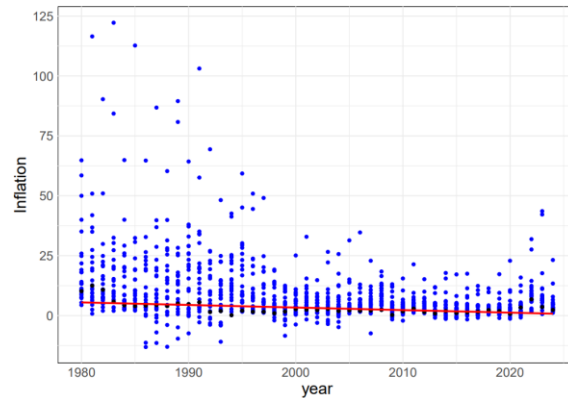


Figure 4. The distribution of global inflation rates for the period 1980-2024.

By conducting a thorough analysis of inflation rate and year in Canada and Australia, this paper confirms the existence of a non-linear relationship between the two variables, see Figure 5. The data reveals that, over time, the inflation rate exhibits a gradual downward trend. To ensure the accuracy of this observation, the authors employed the Mean Square Error (MSE) as a metric to assess the model's fitting quality and predictive capabilities.

However, this data only focuses on the impact of the year on the inflation rate, without considering other potential influencing factors. Therefore, there may be some uncertainty in the experimental results. To further improve the accuracy of the prediction, future research can consider incorporating more factors that may affect the inflation rate, such as economic growth rate, monetary policy, and international trade conditions. By considering these factors comprehensively, it can help the experiment to establish a more comprehensive and accurate prediction model, providing strong support for countries to formulate more scientific and reasonable economic policies and plans.

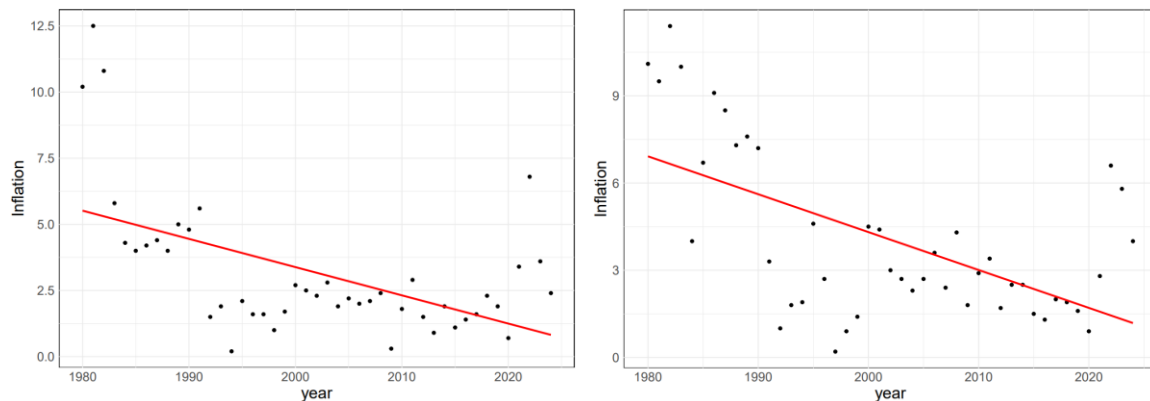


Figure 5. Left and Right: The inflation rates of specific regions of Australia and Canada over the period 1980-2024 and their trend lines.

In addition, the change of inflation rate is a complex and dynamic process, which is affected and restricted by many factors. Therefore, in the process of prediction and analysis, the experiment needs to constantly adjust and improve the prediction model to adapt to the constantly changing economic environment. Through deep research and analysis of the relationship between inflation rate and years, this experiment can provide an important basis for predicting the trend of future inflation rate changes and contribute to the stable growth of the global economy.

Finally, this experiment aims to provide a basis for predicting the future trend of inflation rate and provide a reference for countries to formulate future and price forecasts, in order to promote the stable growth of the global economy. In today's increasingly close global economy, the trend of inflation rate

has a profound impact on the economic development, price stability, and consumer purchasing power of countries. Therefore, accurately predicting the trend of inflation rate is crucial for governments and enterprises in various countries.

4. Conclusion

Regression analysis is an important analytical method in statistics, which aims to explore the relationship between variables and predict future trends. Among them, linear regression is the most used form, and its theoretical basis mainly relies on the least square method, which describes the linear relationship between two variables by fitting a straight line. Linear regression can predict the trend of model data change to a certain extent, and provide strong data support for decision makers. However, many phenomena in the real world often involve multiple factors, which requires the use of multiple linear regression. By introducing more independent variables to predict the change of dependent variables, multiple linear regression can more accurately describe the relationship between data. For example, when analyzing GDP, one can consider several factors such as total import and export volume, total energy consumption, total population, and total retail sales of consumer goods, and construct a multiple linear regression model to understand the impact of these variables more fully on GDP. In addition to linear regression, nonlinear regression is also an important part of regression analysis. Nonlinear regression can predict models of arbitrary relationships between variables and therefore can provide more accurate predictive data. For example, when analyzing the relationship between the global inflation rate and the year, the nonlinear regression equation can help the model find that there is a nonlinear relationship between the global inflation rate and the year with a downward trend, so as to help the decision maker make more accurate calculations. When performing regression analysis, choosing the right method is crucial. Different data types and problem backgrounds of models require different regression equations to ensure the accuracy and reliability of data analysis. In addition, when the data model is validated, it is necessary to check for anomalies and possible effects on the predicted results.

Although representative data and rigorous calculation are selected as far as possible in this paper, the experimental results may be biased from the actual results because other factors that may affect the prediction results are not considered. For example, in the experimental analysis of linear regression equation, in addition to the impact of the unemployment rate on the economic growth rate, the impact of urban development rate, import and export trade tax rate, technology iteration and other factors were not considered. So, the experiment needs to collect more data to make more accurate predictions. In the future, with the development of data science, the application of regression analysis will be more extensive, and more new methods and advanced technologies will be applied to regression analysis. The author of this paper hopes that more scholars can make more accurate and realistic interpretation of regression models and forecast data. This will become the basis for more people to make decisions.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

- [1] Jardin, M., & Stephan, G. (2011). How Okun's law is non-linear in Europe: a semi-parametric approach. Rennes, University of Rennes.
- [2] Yahia, A. K. (2018). Estimation of Okun Coefficient for Algeria. *International Journal of Youth Economy*, 2(1), 1-16.
- [3] Adenomon, M. O., & Tela, M. N. (2017). Application of Okun's law to developing economies: a case study of Nigeria. *Journal of Natural and Applied Sciences*, 5(2), 12-20.
- [4] McCarthy D W, Probst R C, Low F J. (1985). Infrared detection of a close cool companion to Van Biesbroeck. *Astrophysical Journal*, 290, L9-L13.
- [5] Guo W. (2022). Gravitational wave detection of black hole rendezvous. *Progress in Astronomy*, 40(3), 382-393.

- [6] Du F. (2023). Are primordial black holes related to dark matter. Beijing: Science and Technology Daily.
- [7] Yang, Ke, Tian, Feng-ping, Lin, Hong. (2013). Research on International Co-movement in Global Inflation: A Study Based on Bayesian Dynamic Latent Factor Model. *International trade issues*, 6, 145-156.
- [8] Pang Zhen, Wang Kai. (2018). An empirical analysis of the nonlinear effect of inflation on China's economic growth. *Statistics and decision*, 10, 123-126.
- [9] Liu, Tie-Ying, Lee, Chien-Chiang. (2021). Global convergence of inflation rates. *North American journal of economics and finance*, 58, 101501.
- [10] Ciccarelli, M., Mojon, B. (2010). Global Inflation. *Review of Economics and Statistics*, 92, 524-535.