**Trends in Social Sciences and Humanities Research**
**ISSN: 2959-9938**
**DOI: 10.61784/tsshr3018**

# ANALYSIS OF FACTORS INFLUENCING HUNAN PROVINCE'S GDP TOTAL

QiBin Zhu
*School of Mathematics and Statistics, Guangxi Normal University, Guilin 541006, Guangxi, China.*
*Corresponding Email: 1138398110@qq.com*

**Abstract:** This article employs multiple linear regression, ridge regression, and LASSO regression methods to analyze the total GDP of Hunan Province from 2002 to 2021, focusing on eight influencing factors including individual employment, total value of goods imports and exports by foreign-invested enterprises, and local fiscal expenditure. The results indicate a significant positive correlation between the total value of goods imports and exports by foreign-invested enterprises, local fiscal expenditure, and Hunan's GDP. Conversely, factors such as research and development (R&D) activities of industrial enterprises above designated size show a significant negative correlation with GDP. The experimental analysis results suggest positive implications for the healthy and stable growth of Hunan Province's GDP.
**Keywords:** Multiple linear regression; Ridge Regression; LASSO regression; Factors influencing GDP; Multicollinearity

## 1 INTRODUCTION

GDP, also known as Gross Domestic Product, refers to the final output of production activities by all resident units in a country or region during a specified period. GDP is a core indicator of national economic accounting and a crucial measure of a country or region's economic condition and development level. It reflects the scale of economic development, assesses overall economic strength, and evaluates the pace of economic development of a country or region, demonstrating its comprehensive national power[1].

Additionally, GDP is used for economic structure analysis, such as industrial, demand, and regional structure analysis, providing essential foundations for macroeconomic decision-making. GDP growth is vital for all regions and countries, as it meets the needs of economic and social development[2]. During GDP growth, each region enhances its status to varying degrees by expanding its economic size locally. This is because regions contribute more to their development as their economic strength increases. When a region's GDP reaches a certain level, its influence on its resident population within the region grows significantly[3]. This influence also extends to the entire regional population. GDP, when combined with related indicators, helps calculate other significant metrics of importance[4].

Since the beginning of the 21st century, especially since China's accession to the World Trade Organization, China's GDP has experienced rapid development[5]. By 2010, it surpassed Japan's GDP to become the world's second largest, trailing only the United States[6]. Since 2010, particularly following the 18th National Congress of the Communist Party of China, high-quality development has become a defining characteristic of Hunan Province. The economic aggregate has consistently surged forward. Hunan Province's Gross Regional Product (GRP) exceeded 2 trillion yuan in 2012 and surpassed 4 trillion yuan in 2020, marking a rapid ascent across three trillion-level thresholds in just eight years. It is projected to reach another milestone by surpassing 5 trillion yuan for the entire year[7].

Hunan Province's per capita GRP has exceeded $10,000, doubling compared to 2012. Given the impact of the COVID-19 pandemic and the current complex international situation, it is essential to study the factors influencing Hunan Province's GDP to sustain healthy development, avoid economic crises, steadily enhance residents' income, improve their quality of life, and provide relevant recommendations for future development[8].

This article utilizes data related to Hunan Province's GDP from 2002 to 2021, employing multiple linear regression, ridge regression, and LASSO regression methods combined with relevant literature to quantitatively and qualitatively analyze the factors influencing GDP, aiming to provide recommendations for future development.

## 2 PRELIMINARY KNOWLEDGE

### 2.1 Variable Selection and Explanation

To comprehensively consider the factors influencing GDP and based on the actual situation in Hunan Province, this paper selects Hunan Province's GDP as the dependent variable. The independent variables chosen are individual employment, total value of goods imports and exports by foreign-invested enterprises, local fiscal expenditure, per capita consumer expenditure of residents, total water supply, total retail sales of social consumer goods, electricity generation, and research and development (R&D) activities of industrial enterprises above designated size. To facilitate subsequent research, alphabetic symbols are used to represent these nine variables as detailed in the table 1 below:

**Table 1** Variable Description

| φ | Total GDP |
|---|---|
| $x_1$ | Individual employment |
| $x_2$ | Total value of goods imports and exports by foreign-invested enterprises |
| $x_3$ | Local fiscal expenditure |
| $x_4$ | Per capita consumer expenditure of residents |
| $x_5$ | Total water supply |
| $x_6$ | Total retail sales of social consumer goods |
| $x_7$ | Electricity generation |
| $x_8$ | Research and development (R&D) activities of industrial enterprises above designated size |

The units for the variables in the table are respectively: 100 million yuan, persons, billion US dollars, 100 million yuan, yuan, billion cubic meters, 100 million yuan, billion kilowatt-hours, and 100 million yuan.

### 2.2 Data Source

The purpose of this article is to study the factors influencing the total GDP of Hunan Province and to fit a regression model to explain the linear relationship between explanatory variables and the dependent variable. The data used in this study were sourced from the official websites of the National Bureau of Statistics (http://www.stats.gov.cn/) and the Hunan Provincial Bureau of Statistics (http://www.tjj.hunan.gov.cn/), covering nine economic indicators from the years 2002 to 2021, spanning a period of 20 years. The downloaded data were processed to convert them into a standard data format. Due to the age of some indicators, early data points were missing. To ensure the integrity of the analysis, this article employed simple non-random methods to impute missing values, using techniques such as mean, median, and mode.

## 3 MULTIPLE LINEAR REGRESSION

### 3.1 Model Establishment

In many real-life problems, there are often multiple factors influencing the dependent variable. When there exists a linear relationship between the dependent variable and several independent variables, this modeling problem is referred to as multiple linear regression.

The basic model of multiple linear regression is as follows:

$\varphi = \beta_0 + \beta_i x_i + \varepsilon$, and $1 \le i \le 9$.

In the equation, $\beta_i$ represents the regression parameters of each factor in the model.Obtained through the method of least squares or maximum likelihood estimation;$\beta_0$is the regression intercept, $x_i$is the independent variable, also known as the explanatory variable or predictor variable; ε represents the error term, which, similar to the simple linear regression model, εhas a mean of 0 and a variance of $\sigma^2$.

The regression equation satisfies the following basic assumptions:

**A1.** The explanatory variable $x_i$ is a constant variable.Furthermore, it is not a random variable, and the independent variables in the design matrixXare mutually independent, meaning $XX^T$ is nonsingular with a nonzero determinant.The number of samples n should be greater than the number of independent variables k, and X is a full-rank matrix.

**A2.** The random error term of the regression equation has the characteristics of a mean of 0, homoscedasticity, and independence.

**A3.** The random errors must follow a normal distribution.

### 3.2 Results and Analysis

#### *3.2.1 Multiple linear regression modeling*

Due to the lack of uniformity in data units and their large magnitudes, standardizing the independent variables allows for more accurate comparison of their effects on the dependent variable. The scale function in R language is used for data standardization. Based on the method of least squares, the linear regression equation is established using the lm function. From Table 2, the multiple linear regression equation is obtained as follows:

$$\varphi = 7.98 \times 10^3 - 1.418 \times 10^{-4} x_1 + 0.1055 x_2 + 0.9085 x_3 + 1.406 x_4 \quad (1)$$

$$+ 5.518 \times 10^{-4} x_5 - 26.25 x_6 + 4.844 x_7 - 5.192 x_8 \quad (2)$$

**Table 2** Linear Regression Parameter Estimates

| Variables | Parameter Estimates | Standard Error | T value | Pr>\|t\| |
|---|---|---|---|---|
| Intercept | 7.980e+03 | 1.025e+04 | 0.778 | 0.45291 |
| $x_1$ | -1.418e-04 | 1.878e-04 | -0.755 | 0.46605 |
| $x_2$ | 1.055e-01 | 5.298e-01 | 0.199 | 0.84576 |
| $x_3$ | 9.085e-01 | 6.433e-01 | 1.412 | 0.18553 |

| | | | | |
|---|---|---|---|---|
| $x_4$ | 1.406e+00 | 7.863e-01 | 1.788 | 0.10127 |
| $x_5$ | 5.518e-04 | 5.154e-04 | 1.071 | 0.30722 |
| $x_6$ | -2.625e+01 | 3.045e+01 | -0.862 | 0.40714 |
| $x_7$ | 4.844e+00 | 1.197e+00 | 4.045 | 0.00193 |
| $x_8$ | -5.192e+00 | 5.995e+00 | -0.866 | 0.40492 |

The fitted coefficient $R^2$ reached 0.9995, adjusted to 0.9991, indicating that the explanatory variables effectively explain the incidence of hypertension.This statement that explains the incidence of hypertension very well. However, at a significance level of 0.05, from Table 2, it is observed that all independent variables except $x_7$ are not significant. Additionally, individual employment (persons), total retail sales of social consumer goods (100 million yuan), and R&D activities (100 million yuan) show a negative correlation with total GDP, which is contrary to common sense. Therefore, it is necessary to test the model for heteroscedasticity and autocorrelation of the error term. Check whether there is a linear relationship between variables and whether there is multicollinearity among explanatory variables.

### *3.2.2 Test for linearity of error terms*

We used the plot function on the results obtained from the lm function, and then used the crPlots function to test for linearity. The evaluation of model fit based on Figure 1 is shown below.

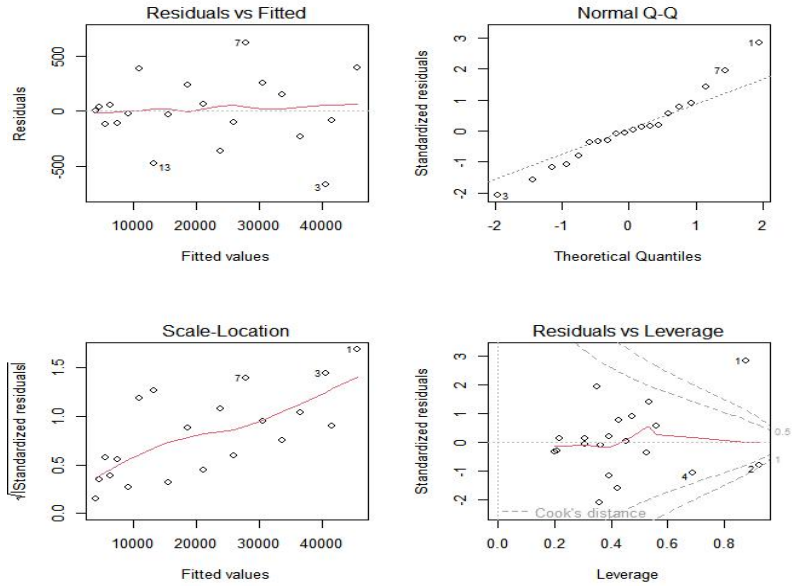


**Figure 1** Linear Relationship Diagnostic Plot

### *3.2.3 Test for multicollinearity*

The heteroscedasticity test and autocorrelation test passed, indicating no issues with the error terms. Considering the multicollinearity between variables, the variance inflation factor (VIF) was computed using the vif function from the car package. VIF measures the extent of variance inflation among explanatory variables. Generally, a VIF > 10 indicates severe multicollinearity. The diagnostic results are shown in Table 3:

**Table 3** Results of Multicollinearity Test

| | vif | Vif > 10 |
|---|---|---|
| $x_1$ | 38.029722 | TRUE |
| $x_2$ | 1292.832916 | TRUE |
| $x_3$ | 402.769319 | TRUE |
| $x_4$ | 2314.114228 | TRUE |
| $x_5$ | 184.884440 | TRUE |
| $x_6$ | 5.056438 | FALSE |
| $x_7$ | 40.460161 | TRUE |
| $x_8$ | 82.761979 | TRUE |

According to the data in the table, except for variable $x_6$, the variance inflation factor (VIF) values of the other independent variables are very high, clearly exceeding VIF > 10. This indicates significant multicollinearity among the explanatory variables that cannot be ignored.

## 4 RIDGE REGRESSION AND LASSO REGRESSION

### 4.1 Ridge Regression Modeling

From the earlier scatter plot matrix, it is evident that there is linear correlation among multiple sets of independent variables. Additionally, the VIF also indicates a significant multicollinearity issue among the independent variables. Therefore, the credibility of the multiple linear regression model built from these data is not high. The analysis of the model coefficients further confirms this point. Ridge regression is essentially an improved least squares estimation method that sacrifices some unbiasedness of least squares estimation to obtain regression coefficients that are more practical and reliable, albeit at the cost of losing some information and reducing precision.
Using the linearRidge function from the ridge package in R for ridge regression, simultaneously selecting ridge regression parameters. Utilizing the lm.ridge function from the MASS package to set parameter ranges, the ridge trace plot is obtained as shown in Figure 2, and the parameter estimates are shown in Table 4.

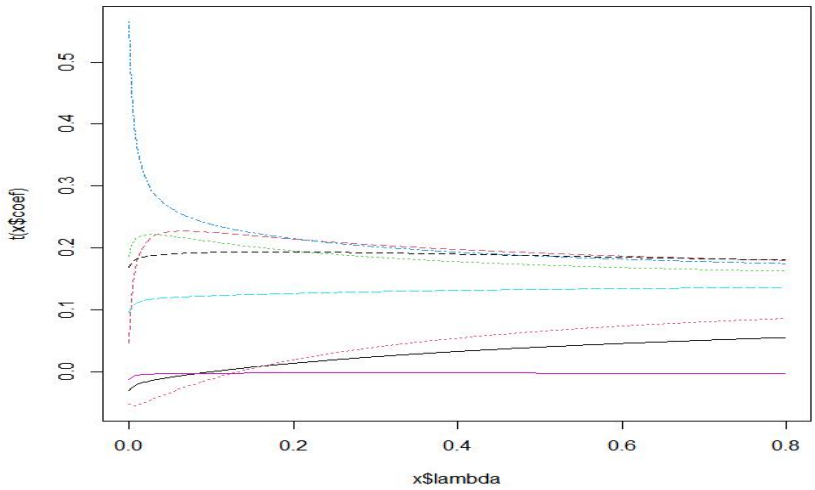


**Figure 2** Ridge Trace Plot

Next, the regression equation obtained will be tested by randomly splitting the data into two parts: 70% of the data will be selected as the training set to train the neural network model, while the remaining 30% will serve as the test data to validate the model. The root mean square error (RMSE) will be used as the evaluation criterion. The RMSE for the training set and test set, as well as for the training set and test set after ridge regression, are shown in the table below:

**Table 4** RMSE of Training Set and Test Set

| DATA TYPE | RMSE |
|---|---|
| traindata | $4.003 \times 10^{-18}$ |
| testdata | 0.147 |
| Traindata(after ridge regression) | $4.644 \times 10^{-17}$ |
| Testdata(after ridge regression) | 0.034 |

Based on the above data, the RMSE for the training samples and the test samples obtained from ridge regression are essentially consistent.

### 4.2 Lasso Regression

LASSO, first proposed by Robert Tibshirani in 1996, stands for Least Absolute Shrinkage and Selection Operator. It is a method based on the principle of shrinkage estimation. By constructing a penalty function, LASSO aims to obtain a more refined model that compresses certain regression coefficients, enforcing the sum of their absolute values to be less than a specific value. Additionally, it sets some regression coefficients to zero, thereby retaining the benefits of subset shrinkage. LASSO is particularly effective for biased estimation in data with complex collinearity.
Similar to ridge regression, LASSO transforms a constrained optimization problem into an unconstrained penalty function optimization problem by adding a penalty term. However, unlike ridge regression, LASSO does not yield an analytical solution. Nevertheless, its regression results assist in appropriate feature selection, making it advantageous compared to ridge regression.
Figure 3 illustrates the results of coefficient changes with parameter variations. The horizontal axis represents the ratio of model coefficients, the vertical axis represents the corresponding explanatory variables, dashed lines represent variables, and vertical lines denote penalty values.

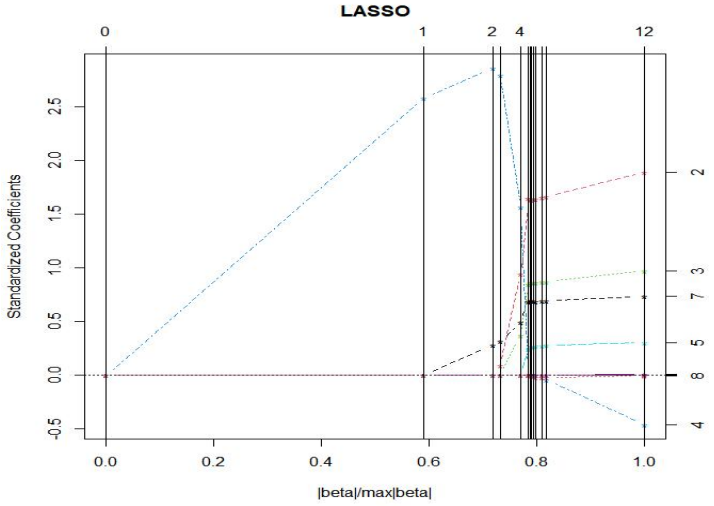


**Figure 3** Lasso Regression Plot

Use cross-validation to select T value,Get the left picture in Figure 4.At this point we know that the recommended T value is around 0.8. At this time, it can be seen from the above figure: when the value of T is 0.84848, a model containing 7 independent variables is generated. The right figure in Figure 4 shows the estimated coefficients. We can clearly see that the 7 independent variables used to predict the dependent variable.
The model obtained using LASSO regression is as follows:

$$\varphi = 0.50x_2 + 0.26x_3 - 0.03x_4 + 0.008x_5 + 0.20x_7 - 0.01x_8 \quad (3)$$

Clearly, LASSO regression has identified $x_2$、$x_3$、$x_4$、$x_5$、$x_7$、$x_8$ to predict the dependent variable. These six variables are respectively the total value of goods imports and exports by foreign-invested enterprises, local fiscal expenditure, per capita consumer expenditure of residents, total water supply, electricity generation, and R&D activities. Among them, per capita consumer expenditure and R&D activities have a negative impact on the GDP of Hunan Province, but their coefficients are close to zero in absolute value, which aligns with the actual situation.

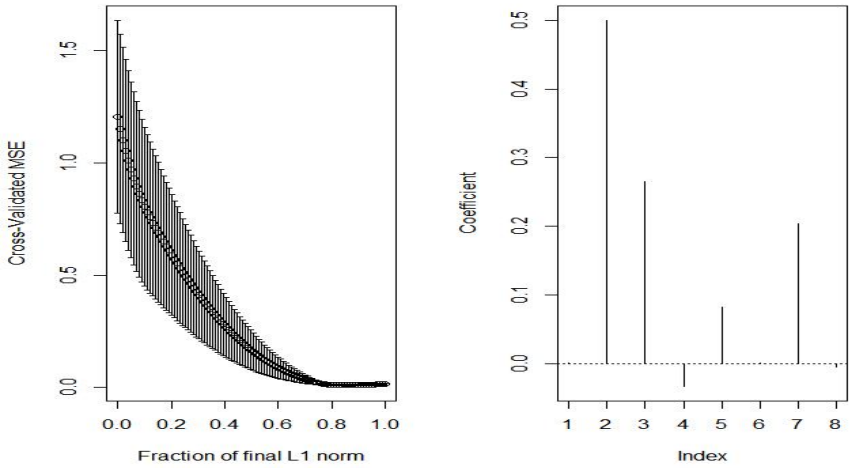


**Figure 4** Cross-Validation Performance Plot

## 5 CONCLUSION AND OUTLOOK

### 5.1 Problems with this Article

#### *5.1.1 Error term test subjective*

In the heteroscedasticity test of the error term, only the method of observing the normal distribution chart is used to make judgments, and the image is relatively subjective and not very convincing. The model may have some effects of heteroskedasticity without improvement. Although the p value satisfies the condition of non-significant, it is initially believed that the error terms are independent. However, the DW value is 1.824. Although it is close to 2, it still shows that there is a weak, that is, part of the negative correlation. There is no moderate improvement here.

#### *5.1.2 Exception points are not processed*

On the one hand, outliers may be due to different sources of data collection, resulting in the data obtained in a certain year not being collected under the same standard. On the other hand, it would be too hasty to directly eliminate outliers. This article has not found a more suitable method to deal with these outliers.

#### *5.1.3 Missing value imputation may be suboptimal*

This paper uses the mean and median to replace missing data on prevalence. The advantage is that the number of samples is not reduced and random errors are not introduced. The disadvantage is that the variance of the variables calculated for non-random data is reduced. If the number to be supplemented is large, the standard deviation will be underestimated, leading to errors in the judgment of variable correlation. Need to be used with caution.

### 5.2 Conclusion

The article analyzes the influencing factors of Hunan Province's GDP through multiple linear regression, ridge regression and LASSO regression methods. By comprehensively comparing the results of the three analysis methods, we obtain the total import and export of goods by foreign-invested enterprises, local fiscal expenditures, and per capita consumption of residents. Expenditure, total water supply, and power generation all have a non-negligible impact on the total GDP of Hunan Province, and the total import and export of goods by foreign-invested enterprises, local fiscal expenditures, per capita consumption expenditure of residents, and power generation are all positively related to the GDP of Hunan Province, but R&D is negatively related to Hunan Province's GDP.

Therefore, in order to promote the healthy and sustainable development of GDP in Hunan Province, the following suggestions can be put forward:

(1) The total import and export of goods by foreign-invested enterprises has a significant positive correlation with the GDP of Hunan Province. Therefore, it is necessary to vigorously attract foreign enterprises to invest in Hunan and increase the total import and export volume.

(2) There is a significant positive correlation between local fiscal expenditure and Hunan Province's GDP. Local fiscal expenditures have a significant impact on promoting local employment, economic development, and infrastructure development.Therefore, we must find ways to increase local fiscal revenue so that we can more effectively increase GDP development.

(3) R&D can have a positive impact on the high-tech industry in Hunan Province. Educational experiments are investments that must be invested in costs. They are the key to truly solving the bottleneck problem Therefore, even if R&D will have a negative impact on total GDP, we cannot stop investing in it.

## COMPETING INTERESTS

The author have no relevant financial or non-financial interests to disclose.

## REFERENCES

[1] Li Bingxiao, Zhang Shiwei, Zheng Shuyu, Zhao Zhifan. Research on hydropower prediction based on a combined model of multivariate linear regression and ARIMA. Science and Technology Innovation, 2022(33): 71-74.

[2] Liang Haonan. Empirical study on factors affecting GDP in Anhui Province - Based on multiple regression analysis . Times Finance, 2021(24): 73-75.

[3] Ma Liyun. Ridge regression analysis of life insurance demand factors in my country. Modern Commerce and Industry, 2019 (05): 117-118.

[4] Shu Shuhua. Analysis of factors affecting household paper consumption in my country based on ridge regression . China Paper, 2022, 43(14): 48-51.

[5] Zhang Lei, Wu Hao. Multiple linear regression analysis of factors influencing employment numbers. Fujian Computer, 2022, 38(10): 12-16.

[6] Li Bingxiao, Zhang Shiwei, Zheng Shuyu, Zhao Zhifan. Research on hydropower prediction based on a combined model of multivariate linear regression and ARIMA. Science and Technology Innovation, 2022(33): 71-74

[7] Wang Peng, Cheng Wenshi. Research on the intensive land use and driving factors of cultivated land in Jiuquan City based on ridge regression model. Land and Natural Resources Research, 2022(04): 1-5.

[8] Zhang Qingxiu, Li Hongmei. Analysis of influencing factors of consumption demand in Hebei Province based on ridge regression. China Market, 2022(23): 23-27.