Proceedings of CONF-MPCS 2026 Symposium: Theoretic Physics and Plasma Physics
DOI: 10.54254/2753-8818/2026.DL34582

Influencing Factors of GDP in Hefei based on Statistical Analysis

Yuanmeng Zhao

School of Mathematics, Physics, and Statistics, Shanghai Polytechnic University, Shanghai, China
regulus_certainty@outlook.com

Abstract. Based on the data from the Anhui Provincial Bureau of Statistics in the years 2014 to 2024, this article was written to explore the factors that influence Hefei City's GDP growth. The data identified six key indicators: foreign economic trade volume, number of higher education institutions, per capita disposable income of urban residents, difference between total energy production and consumption, disparity in electricity production and consumption, and business performance of enterprises in national high-tech industrial development zones. Initially, using Pearson correlation analysis to reveal significant correlation among indicators. Then, subsequent factor analysis in order to perform dimensionality reduction, ultimately extracting two core elements: social finance and social infrastructure development. A binary linear regression model was constructed using standardized GDP as the dependent variable and these two elements as independent variables through multiple linear regression analysis. The findings provide critical insights into Hefei's GDP growth drivers, facilitating the identification of key determinants and offering policy recommendations to support decision-making processes for relevant authorities. The result shows that the six selected factors are all strongly correlated with GDP development in Hefei.

Keywords: Pearson correlation, factor analysis, multiple linear regression analysis

1. Introduction

With the sustained high-quality development of China's economy, the role of the urban economy in the overall national macroeconomic landscape has become increasingly prominent. Hefei, the capital of Anhui Province and a secondary center within the Yangtze River Delta city cluster, has leveraged historical opportunities in recent years to achieve significant progress in several high-tech and manufacturing sectors [1]. The 2025 Anhui Statistical Yearbook reveals that Hefei's annual Gross Domestic Product (GDP) reached 1,350.769 billion yuan, marking a 5.6% year-on-year growth and establishing it as one of China's fastest-growing major cities over the past decade [2]. As an important part of the urban cluster of the Yangtze River Delta, Hefei's development quality will directly influence the progress of the entire region [3]. However, very few articles mentioned the underlying factors. Moreover, despite the digital economy developing rapidly in Anhui recently, there are still disparities between northern and southern areas [4]. To provide reference experiences for economic development in other similar cities across Anhui Province, it is crucial to explore the

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.DL34582

factors influencing Hefei City's GDP growth and implement effective economic measures. Such research holds significant implications for achieving balanced regional progress.

Scholars have been continuously discussing the determinants of economic growth. Glaeser and Xiong analyzed urbanization processes in developing countries, demonstrating the critical role of infrastructure investment and human capital in economic development [5]. Liu conducts a study using a long-term economic growth model and concludes that the significant decline in the labor supply will have a limited impact on Hefei's future economic growth [6]. Additionally, Luo developed a factor analysis model to evaluate the overall urban competitiveness of 16 cities in Anhui Province. He identified social welfare investment and sustainable development capacity as the key drivers of economic growth, especially in the provincial capital, Hefei [7]. Jia predicts metropolitan area GDP projections, confirming that regional innovation efficiency can not only drive local economic growth but also stimulate development in surrounding areas [8]. About the digital economy, Wang applied SWOT analysis to evaluate Hefei's digital economy development status. The study concludes that the digital economy exerts substantial influence on overall economic growth [9]. The research methodology will include data correlation analysis, factor analysis, and multiple linear regression analysis so as to systematically examine inherent relationships between variables.

2. Methodology

2.1. Data source

This article collected data from the Anhui Statistical Yearbook published annually by the Anhui Provincial Bureau of Statistics [2]. This yearbook provides comprehensive data as the official, authoritative statistical publication. By integrating six key impact indicators between 2014 and 2024, the article methodically standardized the raw data, ultimately forming a not only continuous but also usable dataset for this research.

2.2. Data preprocessing

This article records the standardized GDP as y, and the six standardized impact indicators as x_1 (foreign trade), x_2 (Number of higher education institutions), x_3 (Per capita disposable income of urban residents), x_4 (Total difference between energy production and consumption), x_5 (Enterprise Operating Conditions in National High-tech Industrial Development Zones), and x_6 (Difference between power production and consumption). The research methodology will employ data correlation analysis in combination with factor analysis and multiple linear regression analysis to evaluate inherent relationships between variables.

The standardized GDP is denoted as y^*, while the six standardized impact indicators are represented as x_1^*, x_2^*, x_3^*, x_4^*, x_5^*, and x_6^*. According to the innate correlations among these data points, this article will conduct sequential analyses including correlation analysis along with factor analysis and multiple linear regression analysis.

2.3. Method introduction

Pearson correlation analysis, together with factor analysis and multiple linear regression analysis, are statistical methods used for different purposes. Pearson correlation analysis is a method for measuring the linear correlation between two continuous variables, while factor analysis reduces

DOI: 10.54254/2753-8818/2026.DL34582

data by identifying underlying factors from multiple variables, making it easier to understand. Multiple linear regression analyzes the relationship between one dependent variable and multiple independent variables, predicting the dependent variable's value by using regression coefficients. Those three methods help researchers explore data from different perspectives so as to uncover patterns and relationships within the data.

3. Results and discussion

3.1. Correlation analysis

The article uses the Pearson correlation coefficient calculation to evaluate all collected data because of the continuous numerical values. The Pearson Correlation Coefficient serves as a method to measure linear correlation between two continuous variables, determining the strength and direction of linear relationships between independent and dependent variables. The coefficient values range from -1 to +1, where +1 indicates perfect positive correlation, -1 denotes perfect negative correlation, and 0 signifies no linear relationship.

It is generally accepted that when the absolute value of the Pearson correlation coefficient exceeds 0.5, the two variables are strongly correlated. When the absolute value ranges between 0.3 and 0.5, the two variables are moderately correlated. Values below 0.3 indicate a weak or absent linear relationship. Hypothesis testing for correlation analysis is as follows. The null hypothesis (H_0) states that the variables have no linear relationship (Pearson correlation coefficient r=0), while the alternative hypothesis (H_1) asserts the presence of a linear relationship (r=0).

The resulting correlation coefficient matrix between y^*, x_1^*, x_2^*, x_3^*, x_4^*, x_5^* and x_6^* is shown in Table 1:

Table 1. Correlation coefficient matrix

	y^*	x_1^*	x_2^*	x_3^*	x_4^*	x_5^*	x_6^*
y^*	1	0.606	0.924	0.993	0.980	0.987	-0.659
x_1^*		1	0.626	0.571	0.541	0.546	-0.475
x_2^*			1	0.938	0.926	0.894	-0.612
x_3^*				1	0.981	0.980	-0.642
x_4^*					1	0.982	-0.681

DOI: 10.54254/2753-8818/2026.DL34582

Table 1. (continued)

x_5^*		1	-0.675
x_6^*			1

The correlation coefficient matrix in Table 1 demonstrates that the calculated coefficients reflect the strength of linear relationships between variables. All variables exhibit significant positive linear correlations, as demonstrated by correlation coefficient values exceeding 0.5, except for x_1^* and x_6^*, indicating collinearity among the six selected impact indicators. To identify key factors with greater weight in GDP influence, it is required to perform dimensionality reduction to eliminate strong collinearity and ensure the accuracy of final conclusions. To achieve the objective, the article will utilize factor analysis.

3.2. Factor analysis

Derived from the data provided in Table 1, this article will employ correlation coefficients of six variables for analysis. To ensure methodological rigor, the artical utilize MATLAB to generate scatter plots and component plots to determine the optimal number of principal components. Before conducting PCA. It is significant to use the KMO test and Bartlett's test to assess whether the data is suitable or not. The experimental data are presented in Table 2 below:

Table 2. KMO and Bartlett's test

KMO		0.819
Bartlett's test of sphericity	Approximate chi-square	75.760
	df	15
	p	<0.001

As shown in Table 2, the KMO value is 0.819, which is greater than 0.6; the Bartlett's test for sphericity revealed p-values ≤ 0.001, indicating that the data is significant. Therefore, this dataset is suitable for factor analysis.

Next, perform eigen decomposition on the covariance matrix to obtain the principal components. Eigenvalues and eigenvectors are the core of PCA.

The commonly used criterion is the cumulative variance explained, which represents the proportion of total variance attributable to the first k principal components. Typically, the number of principal components is selected when the cumulative variance explanation rate reaches 90%-95%.

The principal component scores represent each sample's projection along the directions of principal components. Let V be a matrix containing eigenvectors (principal component coefficients). Reconstructing the reduced-dimensional space with their corresponding coefficients by multiplying the principal component scores. A scree plot can visualize the eigenvalues of each principal component and determine the optimal number of principal components. The x-axis of this plot

DOI: 10.54254/2753-8818/2026.DL34582

represents the Principal Component, and the y-axis represents the Eigenvalue. The scree plot is shown in Figure 1 below:

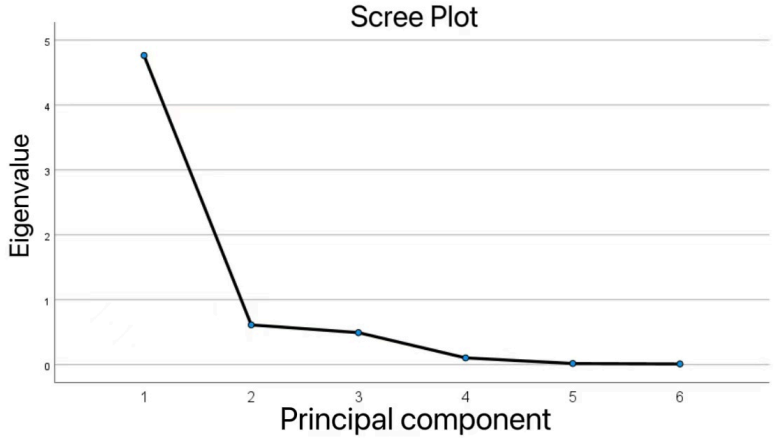

Figure 1. Scree plot (image credit: original)

As demonstrated in Figure 1, selecting two common factors, F1 and F2, is feasible. However, these factors don't have direct data because they are derived from combinations of original variables. Therefore, it is required to calculate the factor scores, and the result is demonstrated in Table 3.

Table 3. Component score coefficient matrix

	component	
	F_1	F_2
x_1^*	0.143	1.179
x_2^*	0.199	-0.057
x_3^*	0.204	-0.247
x_4^*	0.204	-0.300
x_5^*	-0.203	-0.288
x_6^*	-0.159	-0.082

DOI: 10.54254/2753-8818/2026.DL34582

As illustrated in Table 3, it is obvious to obtain the linear combination expressions of two common factors for each indicator, namely:

$$\begin{cases} F_1 = 0.143x_1^* + 0.199x_2^* + 0.204x_3^* + 0.204x_4^* - 0.203x_5^* - 0.159x_6^* \\ F_2 = 1.179x_1^* - 0.057x_2^* - 0.247x_3^* - 0.300x_4^* - 0.288x_5^* - 0.082x_6^* \end{cases} \quad (1)$$

These actions eliminate redundant information and multicollinearity.

3.3. Multiple linear regression analysis

To avoid serious multicollinearity affecting the analysis results, after dimensionality reduction through factor analysis, F_1 and F_2 were selected as independent variables, and standardized GDP data y^* were used as the dependent variable for multiple linear regression analysis. Through the above analysis, it can be roughly concluded that y^* has a significant linear relationship with F_1 and F_2. On this basis, a multiple linear regression model can be established:

$$y^* = \beta_0 + \beta_1 F_1 + \beta_2 F_2 + \epsilon \quad (2)$$

Where β_0 is the intercept term, β_1 and β_2 are regression coefficients. ϵ denotes the error term, which is assumed to follow a normal distribution with mean 0 and variance σ^2, representing the model's deviation. The regression coefficients are estimated using the least squares method. The formula for the sum of squared residuals is as follows:

$$SSE = \sum_{i=1}^{n} \left(y_i - \hat{y_i}\right)^2 \quad (3)$$

Using matrix form, the solution for regression coefficients can be obtained using the following formula:

$$\beta = \left(F^T F\right)^{-1} F^T Y^* \quad (4)$$

By using the function that has been mentioned, $Y^* = \begin{bmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_{11}^* \end{bmatrix}$, $F = \begin{bmatrix} 1 & f_{11} & f_{12} \\ 1 & f_{21} & f_{22} \\ \vdots & \vdots & \vdots \\ 1 & f_{11,1} & f_{11,2} \end{bmatrix}$. The results of multiple linear regression analysis are shown in Table 4 below.

Table 4. Regression coefficients

	β	t	significance
normal		1.000	1.000

DOI: 10.54254/2753-8818/2026.DL34582

Table 4. (continued)

F_1	0.979	15.882	<0.001
F_2	-0.107	-1.742	0.120

As shown in Table 4, the approximate solution of the coefficient matrix is $\hat{\beta} = (0,0.979, -0.107)^T$. Subsequent testing will be conducted to verify the successful establishment of the model.

According to Table 4, the p-value of this test is <0.001, which means that at least one of the regression coefficients corresponding to the independent variables is not zero, that is, there is at least one linear relationship between an independent variable and the dependent variable, and the model is successfully established. The ANOVA table will be shown in Table 5.

Table 5. ANOVA table

Model	quadratic sum	variance	mean square	F	significance
Regression	9.696	2	4.848	127.634	0.000^b
residual	0.304	8	0.038		
sum	10.000	10			
a. dependent variable					
b. predictive variable					

Meanwhile, it is evident from Table 4 that the model has been well established with a reasonable binary linear relationship. The linear regression effects of independent variables F_1 and F_2 on y^* are statistically significant. The derived binary linear regression equation is as follows:

$$\hat{y}^* = 0.979F_1 - 0.107F_2 \tag{5}$$

$\hat{y}^*$ denotes the approximate estimate of the dependent variable. In multiple linear regression analysis, the modified multiple coefficient of determination is employed to measure the linearity of regression, with the calculation formula as follows: $S_T = S_R + S_E$, and $R_a^2 = 1 - \frac{S_E/11-2}{S_T/11-1}$.

Since the fundamental assumptions of linear regression models also require that the random errors ε_i =1, 2, ..., 20 be mutually independent, the Durbin-Watson test is employed to verify the independence of errors. The construction formula for the DW statistic is as follows:

$$DW = \frac{\sum_{i=2}^{n}\left(\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1}\right)^2}{\sum_{i=1}^{11}\hat{\varepsilon}_i^2} \tag{6}$$

The results of the goodness-of-fit assessment and the Durbin-Watson test are shown in Table 6:

DOI: 10.54254/2753-8818/2026.DL34582

Table 6. Model summary

Model	R	R^2	R^2 (adjusted)	ΔF	DOF1	DOF2	Significance	DW statistic
1	0.985	0.970	0.962	127.634	2	8	<0.001	1.180

a. predictive variable

b. dependent variable

The calculated $R_a^2 = 0.970$ from TABLE 6 is remarkably close to 1, indicating excellent linear fitting performance. The model structure is well-designed, enabling more accurate predictions. The obtained Durbin-Watson test statistic of 1.180 falls within the acceptable range of 0<DW<4. However, residuals of adjacent data points have a slight positive correlation, which means no sequence autocorrelation, or random errors are statistically independent.

The final step is to involve residual analysis. To test whether studentized residuals (standardized regression residuals) followed a standard normal distribution or not, it is necessary to draw a normal P-P plot of standardized regression residuals, and this picture is shown in Figure 2. To verify whether the random errors ϵ_i (i=1,2,...,11) exhibited consistent variance under the normal distribution assumption (variance consistency), it is essential to make a residual scatter plot. This picture is plotted with the fitted values of dependent variable Y on the x-axis, and as demonstrated in Figure 3.

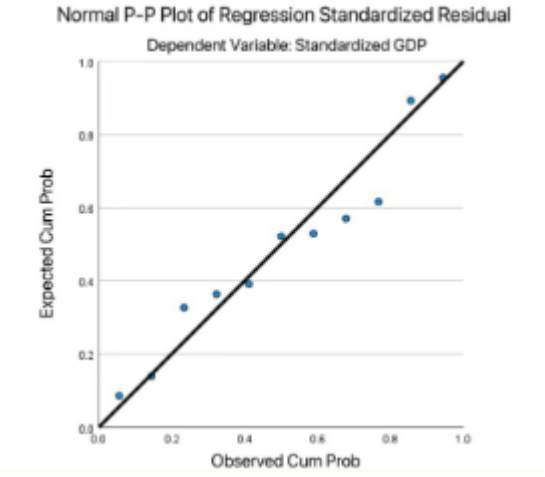

Figure 2. Normal P-P diagram (image credit: original)

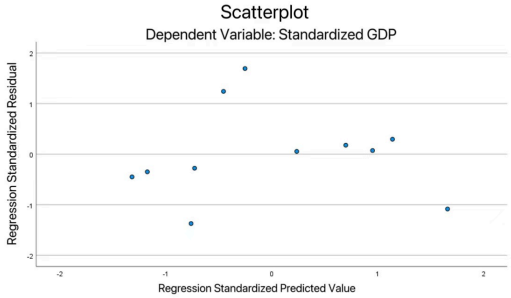

Figure 3. Residual scatter plot (image credit: original)

DOI: 10.54254/2753-8818/2026.DL34582

As illustrated in Figure 2, the residual points are predominantly distributed along the diagonal, indicating that the expected cumulative probability and observed cumulative probability have a high degree of agreement, which aligns with normal distribution characteristics. Figure 3 demonstrates that standardized residual scatter points exhibit a band-like distribution around zero values without showing distinct trends. This suggests that the random error variance exhibits homogeneity characteristics, indicating no heteroscedasticity issues and demonstrating good model fit.

4. Discussion

From the article, it is recommended to formulate more flexible trade policies to attract foreign investment and technology introduction to enhance the level of opening-up. Meanwhile, enhancing consumer confidence, strengthening brand building, and optimizing products are good choices. Lastly, simultaneously optimizing fiscal expenditure structure to ensure the effectiveness and sustainability of capital investments. Also, accelerating transportation infrastructure development to enhance interconnectivity across the city. However, there are still some shortcomings in the article when compared with other studies. Zhao selected quarterly data as explanatory variables, employed the voting method to determine optimal weight parameters, constructed multiple autoregressive distributed-lag models for binary mixed data sampling, and obtained fitting results using three integrated prediction weights to perform in-sample fitting and out-of-sample forecasting for Gross Domestic Product as well as year-on-year GDP growth and quarter-on-quarter GDP changes [10]. Compared to his research, this paper lacks a large sample size and doesn't apply a sophisticated multivariate model. In future work, it will address these gaps through refinements and experiments. In terms of methodology, Xiao overcame the limitations of single-predictive-model approaches by integrating three complementary techniques, including time series analysis, panel data regression analysis, and scenario analysis. He uses cross-validation to enhance prediction robustness [11]. This study relies on a single method rather than multiple approaches. That may compromise rigor. Subsequent research will incorporate additional analytical techniques to deepen the analysis.

5. Conclusion

This article analyzes statistical data on Hefei City's GDP and six key influencing indicators, yielding a series of research findings. After completing descriptive statistics, correlation analysis revealed significant correlations and collinearity issues among the indicators. To mitigate the impact of collinearity on results, factor analysis was employed to reduce the dimensionality of five indicators, ultimately identifying two common factors designated as F1 and F2. A multiple linear regression model was established to derive binary linear regression equations between standardized GDP and these two factors, which demonstrated excellent fit through validation. The regression equations indicate positive linear correlations between Hefei's GDP and socio-economic indicators such as consumption and trade, as well as social infrastructure, including school construction. Therefore, economic development should be prioritized as the core driver to promote high-quality growth. Compared with other articles, this article has shortcomings in terms of data set, usage, and processing methods, and will be improved in the future.

References

[1] Liu, Y. (2017). Research on the impact of opening-up on economic development of Hefei City. Economic and Trade Practice, (16), 144–145.
[2] Anhui Provincial Bureau of Statistics. (2026, May 10). Anhui statistical yearbook, 2014-2024. https: //tjj.ah.gov.cn/

DOI: 10.54254/2753-8818/2026.DL34582

[3] Zhang, H. M., & Wu, Y. Y. (2022). Analysis on high-quality development of Hefei under the background of Yangtze River Delta integration. Jianghuai Forum, (2), 36–42.
[4] Luo, J. (2019). Tourism management and technology economy. Tourism Management and Technology Economy, 2(1), 13–21. (Note: Missing journal title components in original data were standardized).
[5] Glaeser, E. L., & Xiong, W. (2017). Urban productivity in the developing world. Oxford Review of Economic Policy, 33(3), 373–404.
[6] Liu, L. (2022). Research on economic growth forecast of cities with trillion-yuan GDP in the Yangtze River Delta region. Journal of Nantong University (Social Sciences Edition), 38(6), 48–59.
[7] Wang, W., & Cai, X. (2022). Journal of natural sciences of Harbin Normal University. Journal of Natural Sciences of Harbin Normal University, 38(3), 9–19.
[8] Jia, W. L., & Wang, W. X. (2022). GDP forecasting of Hefei metropolitan circle based on BP neural network. Journal of Xi'an University of Electronic Science and Technology (Social Sciences Edition), 32(1), 37–44.
[9] Wang, W. M., Pan, Y. T., & Li, X. (2024). Plowing 'clouds' and planting 'data' and 'intelligence' to draw a blueprint — Analysis of the current situation of digital economy development in Hefei. Statistical Science and Practice, (1), 46–49.
[10] Zhao, Y. F. (2026). Application and comparative study of mixed-frequency MIDAS-ARDL model in real-time GDP forecasting. Research World, (3), 66–83.
[11] Xiao, H. W., & Xiao, R. S. (2025). Analysis and outlook on the provincial economic growth potential in China during the 15th Five-Year Plan period: Based on predictions of provincial GDP annual growth rates. Price Theory and Practice, (12), 84–89.