Dean&Francis
ISSN 2959-6157

From Data to Prediction: A Study on China's GDP Forecasting

Yiheng Yun

College of Science, China University of Petroleum, Qingdao, Shandong Province, 266000, China

2309010125@s.upc.edu.cn

Abstract:

This study utilizes the Least Absolute Shrinkage and Selection Operator (LASSO) regression model, implemented through the R programming language, to refine GDP forecasting methods in China. The analysis centers on the function of the three main monetary indicators—M0, M1, and M2—using time-series data from reliable sources, including the People's Bank of China and the National Bureau of Statistics. These variables have been identified as significant drivers of GDP fluctuations. By leveraging LASSO's ability to select relevant variables and control for overfitting, the model achieves a streamlined and accurate approach to real-time economic forecasting. The results highlight the importance of money supply factors in predicting GDP and emphasize the LASSO model's efficiency in enhancing traditional forecasting techniques. Furthermore, this study offers valuable insights for policymakers and business strategists, who may use these findings to guide economic planning. Integrating modern statistical methods with classical economic models, this research also paves the way for future exploration into global economic forecasting.

Keywords: China GDP Forecasting; LASSO Regression; Economic Prediction Models; Time Series Analysis

1. Introduction

China, as the world's second-largest economy, has garnered significant global attention for its Gross Domestic Product (GDP) growth. In recent years, China's economic expansion has faced numerous uncertainties due to structural transformation, international trade tensions, and the impact of the COVID-19 pandemic. Accurate GDP forecasts serve a wide range of stakeholders, enabling policymakers to develop adaptive economic policies, businesses to plan strategically, and academics to explore economic trends.

A natural resources company that partnered with KPMG to enhance its forecasting capabilities. By integrating advanced GDP prediction models into its planning, budgeting, and performance management processes, the company could anticipate market shifts and economic conditions more accurately. This led to improved decision-making, better risk management, and optimized investment strategies, ultimately increasing the firm's operational efficiency and financial performance.

Data-driven economic forecasting is becoming increasingly important, yet many existing models still

ISSN 2959-6157

struggle to balance complexity and overfitting. In particular, traditional models often rely on a limited number of variables or suffer from excessive noise when incorporating large data sets. Furthermore, the use of contemporary machine learning methods for GDP forecasting in the context of big data, such as the Least Absolute Shrinkage and Selection Operator (LASSO) regression, has not received enough attention. However, Yi pointed out that the LASSO model is particularly suitable for GDP forecasting in a big data environment because it can effectively handle high-dimensional data by setting irrelevant variables to zero and reducing overfitting [1].

According to Anuarbekkyzy, it is necessary to develop new theoretical and methodological provisions to confirm the impact of monetary processes in the economy on GDP in different national economic conditions, which will make it possible to determine the channels and extent of the impact of monetary processes on the real sectors of the economy [2]. In addition, Zuo mentioned that the financial market is an important factor affecting a country's GDP growth, and it indirectly affects the economy by influencing channels such as investment, consumption, and credit [3]. At the same time, this view is highly consistent with what Zhu pointed out [4].

The primary objective of this research is to improve the accuracy of GDP forecasting for China by applying the LASSO regression model. This study aims to enhance real-time GDP forecasting accuracy by integrating macroeconomic analysis with modern data science techniques, providing actionable insights for policymakers and businesses.

2. Model Selection and Parameter Tuning

The study uses cross-validation to optimize the regularization parameter lambda (λ) to adjust the penalty intensity in LASSO regression. The process of adjusting lambda is particularly critical because it requires a trade-off between bias and variance to avoid overfitting of the model and ensure the accuracy of its predictions. The model uses 10-fold cross-validation to train the data set and divides it into training and validation sets to reduce the risk of overfitting and improve the model's adaptability to unknown data.

Considering the characteristics of time series data in GDP forecasting, the study tests multiple lambda values to find the best value that minimizes the forecast error. The model is also adjusted for the common data autocorrelation problem in economic time series forecasting.

The LASSO model stands out in this analysis because it can simultaneously achieve variable selection and regularization, improve the accuracy of predictions, and shrink the coefficients of some variables to zero to avoid overfitting. The model introduces cash in circulation (M0), narrow money (M1), and broad money (M2) as money supply variables, representing different levels of money supply. This approach allowed the study to focus more precisely on key predictive factors while significantly reducing the complexity of the model.

2.1 Research Design and Methods

The core principle of the LASSO regression is to impose an L1 penalty on the regression coefficients, which forces some coefficients to be exactly zero. This results in a sparse model that selects the most relevant variables, reducing the risk of overfitting while retaining predictive power. The LASSO method, therefore, balances the trade-off between bias and variance, which is particularly useful when working with large datasets containing many potential predictors.

A large range of models can be fitted using Lasso (l1-) penalties. These models may now be used to big data sets thanks to newly created computational techniques that take advantage of sparsity for computational and statistical advantages. Numerous disciplines, including statistics, computer science, engineering, and mathematics, are conducting intriguing research on the lasso [5].

2.2 Data Collection Methods

The websites of the People's Bank of China and the National Bureau of Statistics of China provided the data for this investigation. These official institutions provide highly reliable and authoritative economic indicators, ensuring the accuracy and integrity of the dataset used in the analysis. The data cover a range of key variables, such as monetary supply (M0, M1, M2), fiscal expenditure, and other macroeconomic indicators essential for forecasting GDP. By relying on these trusted sources, the study ensures a solid foundation for the subsequent analysis and model development.

3. Result and Analysis

In this analysis, cross-validation is used to select the optimal value of the regularization parameter (lambda), which controls the degree of penalization applied to the model's coefficients. The model is trained on the independent variables M0, M1, and M2, with GDP as the dependent variable.

3.1 Key Steps

1. Data Setup

The independent variables (M0, M1, M2) representing different monetary supply measures are combined into a matrix (X), while GDP serves as the dependent variable

(y).
2. Cross-Validation for Lambda Selection
Cross-validation is conducted to find the best lambda (λ), which minimizes the model's prediction error. The lambda value that yields the lowest cross-validation error is chosen as the optimal regularization parameter.
3. Model Fitting
After identifying the optimal lambda, the LASSO model is fitted using this value to minimize overfitting and improve the model's generalization.
4. GDP Prediction
The trained model is then used to generate predicted GDP values based on the input data. A comparison between actual and predicted GDP values is visualized in a plot, where the actual GDP is represented by a solid blue line and the predicted values by a dashed red line.
5. Model Coefficients
The coefficients from the LASSO model indicate the relationship between each monetary supply variable (M0, M1, M2) and GDP. The LASSO model automatically selects the most relevant predictors by shrinking irrelevant coefficients to zero.
With the optimal lambda identified, the LASSO model is trained on the input data (M0, M1, M2) to predict GDP values. The fitted model is then used to generate predictions for the actual GDP data, which are compared to the true values to assess the model's performance.

In addition to comparing actual and predicted GDP values, the trends in M0, M1, M2, and GDP are analyzed and visualized. To ensure that these variables are comparable on the same scale, they are standardized (scaled) based on their respective means and standard deviations. The standardized data for M0, M1, M2, and GDP are then plotted in a single graph, showing the trends over time. This visualization helps to illustrate the relationship between monetary supply and GDP growth.
To predict GDP for the full year of 2024, the monetary supply data for the first half of the year is used to forecast the second half. A linear regression model is applied to the available data for M0, M1, and M2 from the first six months of 2024, generating predictions for the second half of the year.
The predicted values for M0, M1, and M2 for the second half of 2024 are combined with the first half's data to create a complete dataset for the entire year. This updated data matrix is then used as input to the LASSO model to forecast the GDP for 2024.
Then it gets the output: "The predicted GDP in 2024 is 1308380.3"

3.2 Result

Table 1 shows Actual and Predicted GDP Based on LASSO Regression Using Monetary Supply Variables

Table 1. Actual vs. Predicted GDP Based on LASSO Regression Using Monetary Supply Variables (Unit: 100million yuan)

GDP	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Predicted	640436.7	693528.8	782667.9	861506.9	913128.9	961477.7	1034123.6	1101316.7	1184746.3	1271825.4
Actual	643563.1	688858.2	746395.1	832035.9	919281.1	986515.2	1013567.0	1149237.0	1204724.0	1260582.1

Figure 1 shows Actual and Predicted GDP Based on LASSO Regression Using Monetary Supply Variables

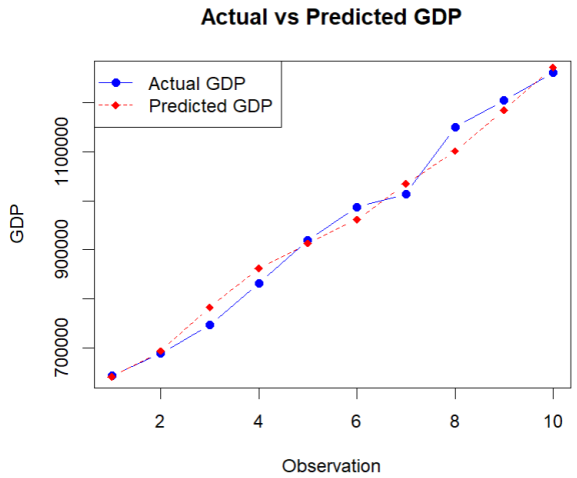

Fig. 1 Actual vs. predicted GDP based on LASSO regression using monetary supply variables (Photo/Picture credit: Original).

A comparison of the actual GDP figures and those forecasted by the LASSO regression model is shown in Table 1 and Figure 1. Figure 1 shows the actual GDP numbers as represented by the blue line and the expected GDP based on the monetary supply variables as represented by the red dashed line.
As shown in Fig. 1, the LASSO model provides a reasonably close prediction of GDP based on the available monetary data. Notably, the model assigns non-zero coefficients only to the variables M1 and M2, while M0 is shrunk to zero. This indicates that the supply of M1 and M2 plays a more significant role in explaining variations in GDP growth, which aligns with economic theory, as higher levels of monetary aggregates (M1, M2) are often

ISSN 2959-6157

more closely related to economic activities than narrower measures like M0.

Units:

· Observation: Years, from 2014 to 2023.

· GDP: 100 million yuan.

Fig. 2 below illustrates the trends of the monetary supply variables (M0, M1, M2) and GDP over the period from 2014 to 2023. The horizontal axis represents the observation years, while the vertical axis represents the scaled values of the respective variables, where all data points have been standardized by subtracting the mean and dividing by the standard deviation.

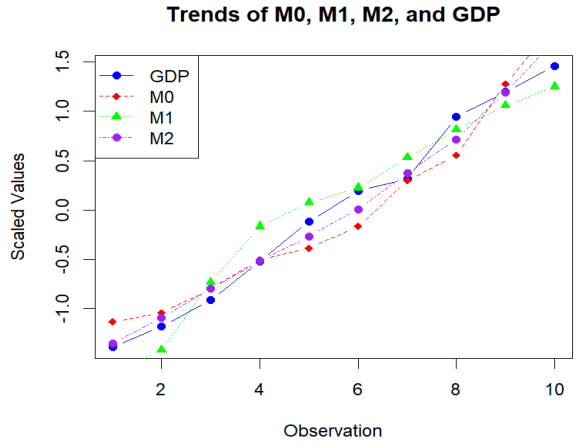

Fig. 2 Trends in monetary supply (M0, M1, M2) and GDP (Photo/Picture credit: Original).

Each monetary supply variable is shown to follow a similar upward trend as GDP, indicating a strong positive correlation over time. This standardized scaling allows for a clearer comparison of the growth trajectories across all variables, demonstrating how closely aligned changes in the monetary supply are with GDP growth during the period under study.

Units:

· Observation: Years, from 2014 to 2023.

· Scaled Values: Standardized units (no dimensions).

To examine the linear relationship between the monetary supply variables (M0, M1, M2) and GDP, two main analyses were conducted: Pearson correlation and scatterplot visualization. The first step involved calculating Pearson correlation coefficients to quantify the strength and direction of the relationships between each monetary variable (M0, M1, M2) and GDP. Stronger correlations are indicated by values around 1 or -1, whereas weaker or no correlation is suggested by values near 0. These coefficients give a numerical representation of the linear relationship.

For the second analysis, scatterplots were generated to visualize the relationships between the monetary supply variables and GDP. These plots allow for a clearer understanding of the trends and patterns in the data. Additionally, regression lines were added to the scatterplots to highlight potential linear relationships between each variable and GDP. The use of these lines helps in identifying the direction and strength of any observed linear trends.

A quantitative indicator of the linear relationship between the variables is provided by the Pearson correlation coefficients, which are shown in Table 2. A strong positive linear link is indicated by a positive correlation coefficient around 1, whereas a strong negative association is suggested by a value near -1. Coefficients near 0 imply no significant linear correlation.

Table 2. Pearson correlation matrix for M0, M1, M2, and GDP

	M0	M1	M2	GDP
M0	1.0000000	0.9178530	0.9910764	0.9643113
M1	0.9178530	1.0000000	0.9572503	0.9748293
M2	0.9910764	0.9572503	1.0000000	0.9886106
GDP	0.9643113	0.9748293	0.9886106	1.0000000

In addition, the scatterplots (Figures 3, 4, and 5) offer visual evidence of the relationships between M0, M1, M2, and GDP. In each plot, the blue dots represent the observed data points, while the red lines show the fitted linear regression models. These figures help to determine whether linearity exists between the variables and GDP, and the slopes of the regression lines provide insight into the direction and strength of these relationships.

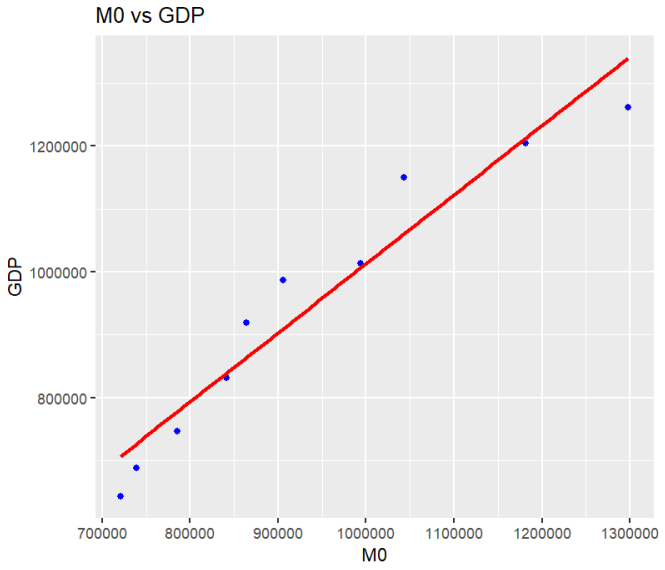

Fig. 3 Scatterplot of M0 vs. GDP (Photo/Picture credit: Original).

Fig. 3 illustrates the relationship between M0 and GDP, with the blue points representing the actual data and the red line depicting the linear regression fit.

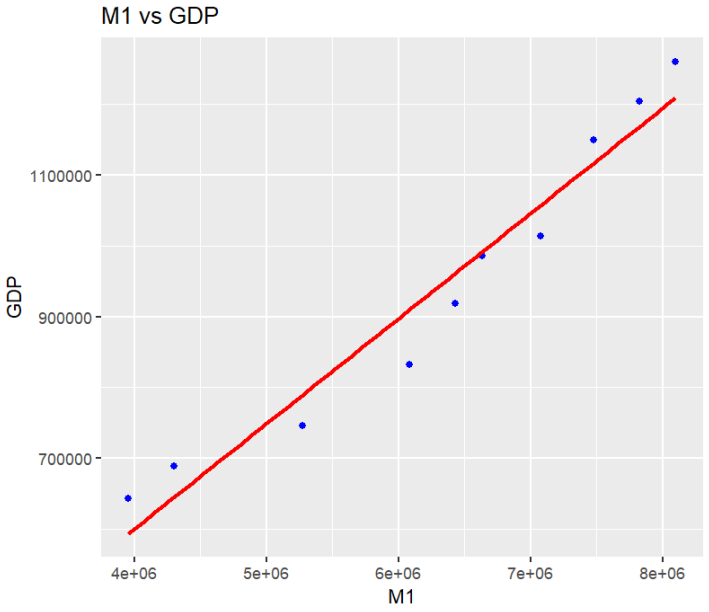

Fig. 4 Scatterplot of M1 vs. GDP (Photo/Picture credit: Original).

Fig. 4 illustrates the relationship between M1 and GDP, with the blue points representing the actual data and the red line depicting the linear regression fit.

ISSN 2959-6157

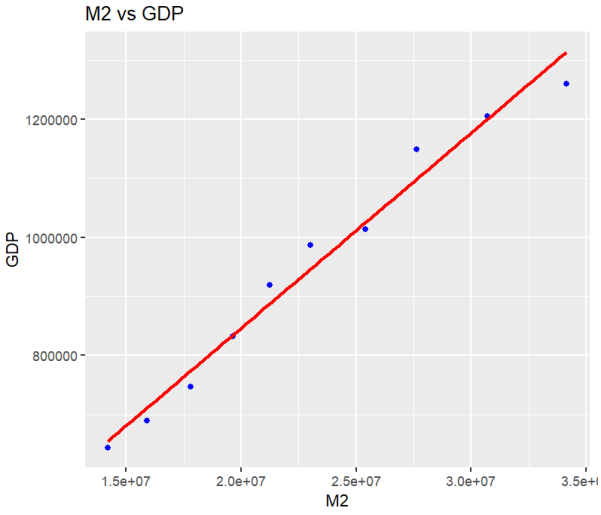

Fig. 5 Scatterplot of M2 vs. GDP (Photo/Picture credit: Original).

Fig. 5 illustrates the relationship between M2 and GDP, with the blue points representing the actual data and the red line depicting the linear regression fit.

LASSO's balance between accuracy and interpretability makes it the most suitable model for GDP forecasting, particularly when the goal is to identify key economic drivers.

4. Discussion

The findings of this study underscore the significance of macroeconomic variables in real-time GDP forecasting and highlight several avenues for further exploration and improvement. This discussion delves into the implications of the results, suggests avenues for future research, and acknowledges the limitations encountered.

4.1 Implications of the Results

The identified macroeconomic variables—government fiscal expenditure, money supply, retail sales of consumer goods, and import value—consistently demonstrated significant predictive power in real-time GDP forecasting. These variables reflect critical economic activities and policy dynamics that influence GDP fluctuations [6].

Government fiscal expenditure emerges as a pivotal driver, influencing economic growth through its direct impact on public investments and infrastructure development [7]. Moreover, the robust association of industrial value-added with GDP underscores the importance of manufacturing output in economic performance [8].

4.2 Limitations

Despite its strengths, this study has several limitations that merit consideration. LASSO regression, while effective in variable selection, assumes linear relationships between predictors and GDP, potentially overlooking non-linear effects [9]. The study's reliance on historical data limits its ability to anticipate unforeseen events or structural shifts in the economy.

Furthermore, the scope of macroeconomic variables included in the analysis, while comprehensive, may not encompass all factors influencing GDP dynamics. Future studies could benefit from exploring additional contextual factors, such as demographic changes or environmental policies, to provide a more holistic view of economic forecasting [10].

4.3 Recommendations for Future Research

Future research should consider integrating additional macroeconomic indicators to enhance forecasting accuracy. For instance, incorporating data on international trade dynamics and financial market indicators could provide deeper insights into external influences on domestic economic conditions. Exploring advanced modeling techniques, such as ensemble methods combining LASSO with other regression approaches or machine learning algorithms, could further refine predictive models [11].

5. Conclusion

This study applied LASSO regression and linear correlation analysis to investigate the relationship between key monetary supply variables (M0, M1, M2) and GDP. The LASSO model identified M1 and M2 as significant contributors to GDP prediction, while M0 had minimal influence, with its coefficient reduced to zero. These results emphasize the greater impact of broader monetary aggregates (M1, M2) on economic output, consistent with established economic theories.

Further analysis, using Pearson correlation and scatter-

plots with regression lines, confirmed the existence of linear relationships between M1, M2, and GDP. The Pearson correlation coefficients revealed a stronger positive correlation between M1, M2, and GDP, indicating that increases in these monetary aggregates are associated with GDP growth. In contrast, M0 demonstrated a weaker relationship with GDP, aligning with the LASSO regression findings.

The scatterplots reinforced these conclusions, displaying clear positive trends between M1, M2, and GDP, as evidenced by the slopes of the regression lines. This analysis highlights the critical role that broader measures of money supply, particularly M1 and M2, play in explaining fluctuations in GDP, offering valuable insights for future economic forecasting models.

The study validates the use of factor compression models for real-time GDP forecasting in China, emphasizing the importance of macroeconomic indicators such as consumption, and trade. While LASSO regression proved effective, limitations include its potential oversimplification by shrinking coefficients to zero, possibly missing weak but relevant correlations. Future research could address these limitations by integrating additional variables or using advanced modeling techniques like elastic net regression or non-linear models.

References

[1] Yanping Y, Dejin H, Xi W. Real time GDP forecast based on macro big data. Economics (Quarterly), 2024, 03: 843-860.

[2] Anuarbekkyzy G, Zhanibekova G, Imramziyeva M, Zholdasbayeva T, Yerkin B, Kenzhin Z. Systematic approach to analyzing the impact of monetary processes in the economy on GDP. Eastern-European Journal of Enterprise Technologies, 2024, 13: 79-90.

[3] Chenrui Z. Predicting China's GDP growth rate: Analysis based on R language and machine learning. Modern Management, 2024, 04: 830-844.

[4] Zhu Q. Analysis of factors influencing Hunan Province's GDP total. Trends in Social Sciences and Humanities Research, 2024, 4.

[5] Tibshirani R. Regression shrinkage and selection via the lasso: A retrospective. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2011, 73(3): 267-288.

[6] Perera P, Fernando A. Impact of energy prices and macroeconomic variables on GDP prediction UK: Machine learning approach. Journal of Business and Management Studies, 2024, 5: 113-124.

[7] Afonso A, Rodrigues E. Is public investment in construction and in R&D, growth enhancing? A PVAR approach. Applied Economics, 2024, 24: 2875-2899.

[8] Babubudjnauth A. An empirical analysis of the impacts of real exchange rate on GDP, manufacturing output and services sector in Mauritius. International Journal of Finance & Economics, 2020, 2: 1657-1669.

[9] Chan Lau J. Lasso regressions and forecasting models in applied stress testing. IMF Working Papers, 2017, 108: 1-1.

[10] Yan G, Zhenfeng S, Xiao H, Bowen C. GDP forecasting model for China's provinces using nighttime light remote sensing data. Remote Sensing, 2022, 15: 3671-3671.

[11] Hao W, Levinson D. The ensemble approach to forecasting: A review and synthesis. Transportation Research Part C. 2021.