Development Economics of China
2025, VOL. 9, NO. 9, 31-35
DOI: 10.47297/wspdecWSP2515-797307.20250909

Application and Research Review of Linear Regression in Stock Forecasting

Longbin Hu

Xiamen University Affiliated Science and Technology High School Xiang'an Campus,Xiamen, Fujian, 361006, China

ABSTRACT
Stock forecasting, as a core issue in the financial field, has long attracted extensive attention from both academia and practitioners. As a classical statistical method, linear regression has a long history and wide application in stock forecasting research due to its model simplicity and strong interpretability. However, with increasing financial market complexity and the rise of nonlinear models like machine learning, the applicability, effectiveness, and integration pathways of linear regression models with emerging technologies demand systematic review and evaluation. This paper aims to comprehensively examine the application context, primary paradigms, advantages, and limitations of linear regression models in stock forecasting through a systematic literature review. First, it reviews the theoretical foundations of linear regression and its evolutionary applications in financial forecasting. Second, it summarizes typical applications and improvement strategies of linear regression models in domestic and international research, including their integration with factor models, time series analysis, and machine learning models. Third, it conducts in-depth case studies to examine their performance across different market environments. Finally, it summarizes existing research gaps and outlines future directions, such as enhancing model interpretability, testing cross-market cycle robustness, and deep integration with deep learning frameworks. This paper aims to provide researchers and practitioners in related fields with a clear theoretical framework and practical reference.

KEYWORDS
Linear regression; Stock prediction; Quantitative investment; Factor models; Literature review

1 Introduction

1.1 Research Background and Significance

Stock markets serve as vital barometers of national economies, with price fluctuations influenced by multiple factors including macroeconomic conditions, corporate fundamentals, and market sentiment, making prediction exceptionally challenging. As a fundamental tool in econometrics and statistics, linear regression was introduced to finance as early as the mid-20th century to explore linear relationships between asset prices and various factors. Despite the rapid advancement of artificial intelligence technologies in recent years, linear regression retains its significant role in quantitative investment and academic research due to the transparency of its parameter estimation, the interpretability of its results, and its irreplaceable function in factor exposure analysis. Systematically reviewing the application, evolution, and limitations of linear regression in stock forecasting holds important theoretical and practical significance for understanding the developmental logic of quantitative investment methods and optimizing model selection strategies.

1.2 Research Objectives and Scope

The primary objective of this study is to conduct a systematic review of linear regression applications in stock price forecasting. Specific research components include:

(1) tracing the foundational theories and methodological evolution of linear regression models in this domain; (2) summarizing the current state, major achievements, and representative case studies of relevant domestic and international research; (3) Analyzing the strengths and limitations of linear regression models across different market scenarios; (4) Exploring complementary and integrated pathways with other forecasting models; (5) Identifying potential future research directions based on existing gaps.

2 Literature Review

2.1 Foundations of Linear Regression Models

The core of linear regression lies in establishing a linear relationship model between a dependent variable (e.g., stock returns) and one or more independent variables (e.g., valuation metrics, momentum factors). Its mathematical expression is concise, and its economic implications are clear, providing a fundamental framework for risk factor pricing and return forecasting.

*** Corresponding Author:** Longbin Hu, 15985866668@163.com

2.2 Current State of Domestic Research

Although domestic research began later, it has developed rapidly. Early studies primarily focused on testing and refining classic foreign models (such as the Fama-French three-factor model) for China's A-share market, constructing factor models better aligned with local market characteristics. In recent years, research trends have shifted toward integrating linear regression with advanced algorithms. For instance, some scholars combine linear regression with LSTM neural networks, leveraging LSTM to capture nonlinear time-series features and then applying linear regression for factor weighting, thereby enhancing overall model performance. Research by scholars like Professor Fan Jianqing in high-dimensional linear models and robust statistics has also provided crucial insights for handling financial big data and enhancing prediction robustness. Additionally, industry experts such as Liang Wenfeng have applied deep learning concepts like Attention mechanisms to market analysis. While their approach does not rely entirely on linear regression, it offers valuable inspiration for constructing hybrid forecasting models.

2.3 International Research Landscape

International research has a longer history, evolving from early attempts to fit stock prices using simple linear relationships to current applications of various complex regularized regression methods (e.g., Lasso, Ridge, Elastic Net) to address high-dimensional data and multicollinearity issues. Research focus has also shifted from pure prediction accuracy to model robustness, interpretability, and integration with market microstructure theory. For instance, the Efficient Market Hypothesis (Fama, 1970) establishes theoretical prerequisites for linear regression applications, while stochastic process theories like Markov chains are combined with linear regression to describe market state transitions and enhance predictive adaptability.

2.4 Review of Current Research Status

While existing research has yielded substantial results, the following gaps remain: (1) A systematic definition of the "methodological spectrum" formed by linear regression and its extended models (e.g., quantile regression, weighted least squares) in stock forecasting is lacking; (2) Performance evaluation criteria vary across different market cycles (bull, bear, sideways) and data frequencies (daily, weekly, monthly), leading to context-dependent conclusions; (3) Insufficient exploration of the "complementary" rather than "substitutive" relationship between linear regression and machine learning models, with a lack of effective integration frameworks.

3 Application Analysis of Linear Regression in Stock Forecasting

3.1 Primary Model Types and Application Scenarios

This section systematically outlines applications ranging from Ordinary Least Squares (OLS) to various extended models, including:

3.1.1 Multiple Linear Regression: The Core Application In Multi-Factor Models

Multivariate linear regression (MLR) is a statistical analysis method used to explore linear relationships between two or more independent variables and a dependent variable. When a linear regression model includes two or more independent variables that exhibit linear dependence with the dependent variable, it is termed multivariate linear regression analysis. The multivariate linear regression estimate is defined as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

Here, y is the dependent variable, representing the value we seek to predict. $x_1, x_2, \cdots, x_p$ are the independent variables used to predict p features. β_0 is the intercept term, representing the predicted value of the dependent variable when all independent variables are 0. It represents the baseline level of the model. $\beta_1, \beta_2, \cdots, \beta_p$ are the regression coefficients, which are the core parameters of the model. β_p indicates the average change in y caused by a one-unit increase in x_p, holding all other independent variables constant. It measures the "net effect" of each independent variable on the dependent variable. ε denotes the random error term, representing the portion of variation in y that cannot be linearly explained by the p independent variables. We typically assume $\varepsilon \sim N(0, \sigma^2)$.

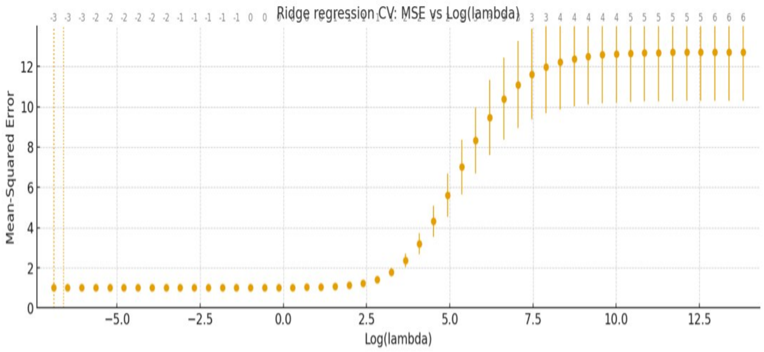

Figure 1

3.1.2 Regularized Regression (Ridge Regression, Lasso, Elastic net)

(1) Ridge regression: To address multicollinearity, Hoerl & Kennard introduced ridge regression in 1970. The ridge regression estimate is defined as

$$\beta\_{Ridge} = \operatorname{argmin}_{\beta} \{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \}$$

where $\lambda \geqslant 0$ is the penalty parameter; larger values of λ result in greater coefficient shrinkage. Specifically, when $\lambda=0$, ridge regression degenerates into LS regression. The penalty term imposes no penalty on the constant term β_0; adding a constant to each response variable does not affect the regression coefficients.

The solution for ridge regression is further derived as:

$$\beta\_{Ridge} = (X^T X + \lambda I)^{-1} X^T Y$$

As seen from the above equation, the ridge regression solution is derived from the LS regression solution by adding a positive penalty parameter λ. Therefore, when some column vectors of the matrix $X^T X + \lambda I$ are approximately linearly correlated, the singularity of this matrix is lower than that of $X^T X$, thereby reducing the variance of the estimates and improving estimation accuracy. However, ridge regression also has limitations. Its regression results include all predictor variables, as ridge regression reduces the magnitude of all features without setting coefficients to zero. Consequently, it does not perform variable selection, which may affect model accuracy.

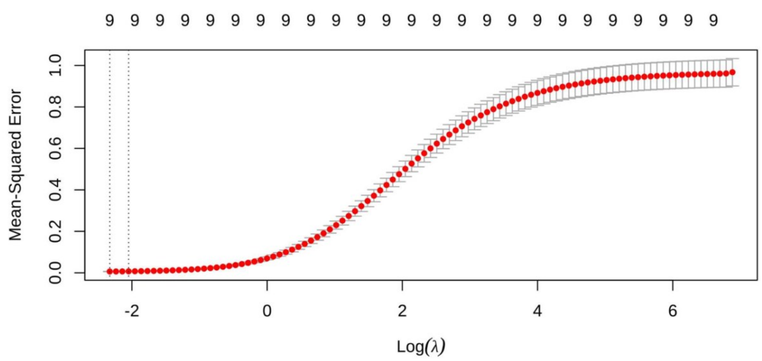

Figure 2 Cross-validation results for ridge regression

(2) Lasso regression: To address the lack of variable selection in Ridge regression, Tibshirani introduced Lasso regression in 1996. The Lasso estimate is defined as

$$\beta\_{Lasso} = \operatorname{argmin}_{\beta} \{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \}$$

Compared to the quadratic penalty term $\lambda \sum_{j=1}^{p} \beta_j^2$ in ridge regression, the linear penalty term $\lambda \sum_{j=1}^{p} |\beta_j|$ in Lasso not only shrinks non-zero predictor coefficients β_j toward zero but also selects valuable predictors (those with large $|\beta_j|$ values). This is because $\lambda \sum_{j=1}^{p} |\beta_j|$ induces less shrinkage on β_j coefficients than $\lambda \sum_{j=1}^{p} \beta_j^2$, enabling Lasso to select more precise models. Consequently, Lasso regression automatically selects important features while excluding irrelevant or insignificant ones, thereby simplifying the model and enhancing interpretability.

Although Lasso regression significantly reduces prediction variance compared to LS regression, achieving coefficient shrinkage and variable selection, it also has limitations. For example:

A. In the Lasso regression solution path, for an $n \times p$ design matrix, at most $\min\{n, p\}$ variables can be selected. When $p > n$, at most n predictor variables can be selected.

B. For typical $n>p$ scenarios, if strong multicollinearity exists among predictor variables, Lasso's predictive performance is constrained by ridge regression.

The optimal regularization parameter λ was selected via 10-fold cross-validation (see Figure 2). As λ increases, model error gradually rises. The minimum error λ was identified as λmin=0.004151, ultimately screening out four significant feature variables: Highest Price (X2), Lowest Price (X3), Trading Volume (X4), and Price Change Rate (X5).

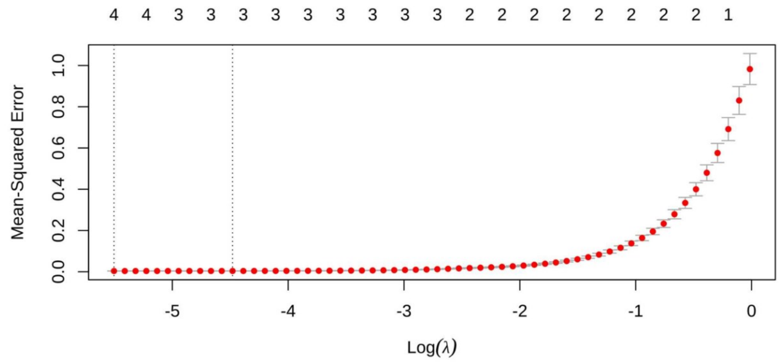

Figure 3 Cross-validation results for lasso regression

(3) Elastic net regression: Addressing limitations of Lasso regression, Zou & Hastie introduced Elastic Net regression in 2005. The regression coefficient expression is

$$\beta_{EN} = \arg\min_\beta \{ \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda[\alpha \sum_{j=1}^p |\beta_j| + (1-\alpha) \sum_{j=1}^p \beta_j^2] \}$$

It follows that the Elastic Net regression penalty term $\lambda[\alpha \sum_{j=1}^p |\beta_j| + (1-\alpha) \sum_{j=1}^p \beta_j^2]$ is precisely a convex linear combination of the ridge regression penalty term and the Lasso regression penalty term. When $\alpha=0$, elastic net regression becomes ridge regression; when $\alpha=1$, it becomes Lasso regression. Thus, elastic net regression combines the advantages of both Lasso and ridge regression.

As shown in Figure 3, the optimal α value selected via 10-fold cross-validation is 0.951. As λ increases, the model error gradually grows. The λ value corresponding to the minimum error can be identified in the figure. Calculations reveal that four significant feature variables are ultimately selected: Highest Price (X2), Lowest Price (X3), Trading Volume (X4), and Price Change (X5).

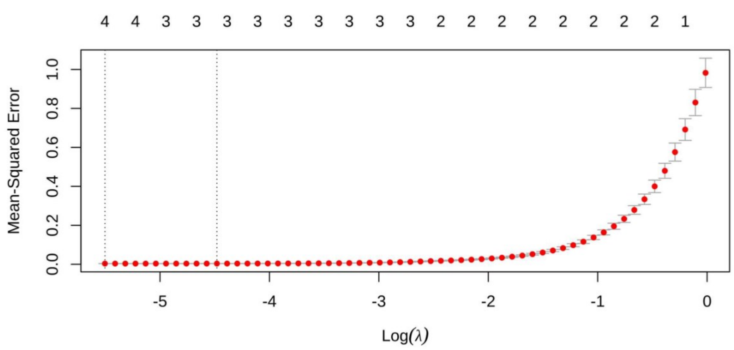

Figure 4 Cross-validation results for elastic net regression

3.1.3 Quantile Regression

Used to examine factor influences at different quantiles (e.g., bear and bull markets).

3.1.4 Robust Regression

Employed to mitigate the impact of outliers on model estimates.

3.2 Performance Across Market Conditions

Compares the predictive stability and accuracy of linear regression in bull, bear, and sideways markets. Generally, linear regression performs well in environments dominated by linear relationships or with weaker market efficiency. Its limitations become more apparent in settings with pronounced nonlinear characteristics and high market volatility.

3.3 Comparison and Integration with other Forecasting Models

Analyze comparative studies between linear regression and traditional time series models (ARIMA), machine learning models (SVM, Random Forest), and deep learning models (LSTM, Transformer). Focus on exploring hybrid modeling approaches like "linear model + nonlinear correction," such as using linear models to capture primary factor effects while employing nonlinear models to capture complex patterns in residuals.

4 Future research Directions

Based on the preceding analysis, potential future research directions are proposed:

(1) Balancing interpretability and complex models: Investigate how linear regression's interpretability can provide insights into "black-box" models (e.g., deep learning).

(2) Cross-Cycle and Cross-Market Robustness Testing: Systematically evaluate the generalization capabilities of different linear regression models across economic cycles and national markets.

(3) Deep Integration with Cutting-Edge AI Technologies: Explore new paradigms combining linear regression with attention mechanisms, graph neural networks, and other approaches to better capture market structures and information propagation networks.

(4) Applications with high-frequency and alternative data: Investigate challenges and opportunities when applying linear regression to high-frequency trading data and alternative data like news sentiment.

5 Conclusion

This paper systematically reviews the application and research progress of linear regression in stock forecasting. Research indicates that despite the growing power of nonlinear models, linear regression remains an indispensable tool in financial forecasting due to its theoretical simplicity, interpretable results, and foundational role in factor analysis. Its future value lies not in replacing complex models but in forming effective complementarity with them. Future research should focus more on the model's application boundaries, robustness, and its new role within hybrid model frameworks.

References

[1] Zhu C., & Xie X. (2024). Stock price prediction based on elastic net regression. E-Commerce Review, 13(4), 5359-5374.

[2] Ping, Yalu. Analysis of the Impact of Multivariate Linear Regression on Construction Prices[J]. Statistics and Applications, 2020, 9(1): 19-25.

[3] Wu, Rengkang. Predictive Analysis of Securities Indices Based on Ridge Regression: Taking the Shanghai Composite Index as an Example [J]. Business Globalization, 2016, 4(2): 47-55.

[4] Zeng Jin, Zhang Yuhao, Du Qiancheng. Research on Tracking the SSE 50 Index Based on Multiple Linear Regression [J]. Operations Research and Fuzzy Science, 2022, 12(4): 1356-1364.

[5] Grace Xing Hu, Can Chen, Yuan Shao, Jiang Wang. Fama–French in China: Size and Value Factors in Chinese Stock Returns. January 23, 2018.

[6] J.Margaret Sangeetha, K. Joy Alfia. Financial stock market forecast using evaluated linear regression based machine learning technique.

[7] A. Ariyo, A. O. Adewumi and C. K. Ayo, "Stock Price Prediction Using the ARIMA Model," 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, Cambridge, UK, 2014, pp. 106-112.

[8] Aldridge, Irene. The AI Revolution: From Linear Regression to ChatGPT and Beyond and How It All Connects to Finance.

[9] R. Karim, M. K. Alam and M. R. Hossain, "Stock Market Analysis Using Linear Regression and Decision Tree Regression," 2021 1st International Conference on Emerging Smart Technologies and Applications (eSmarTA), Sana'a, Yemen, 2021, pp. 1-6.

[10] C. C. Emioma and S. O. Edeki. Stock price prediction using machine learning on least-squares linear regression basis. 2021 J. Phys.: Conf. Ser. 1734 012058.

[11] Fei Wang, Wanling Chen, Bahjat Fakieh, Basel Jamal Ali. Stock price analysis based on the research of multiple linear regression macroeconomic variables. Applied Mathematics and Nonlinear Sciences, ISSN-e 2444-8656, Vol. 7, No. 1, 2022, pp. 267-274.