Theory and Application of Linear Regression Model in Financial Market Forecasting

Hongze Huang

University of California Irvine, Irvine, 92612, USA

Abstract. Linear regression model in the theoretical basis and practical application in the financial markets was discussed by this article, the article first introduces basic principle of linear regression models and assumptions, such as least square method, multiple linear regression was also covers the core concept, And then explores the model used to predict the strengths and weaknesses of the financial markets and focuses on how to choose suitable for the independent variable and dependent variable, and how to deal with the particularity of financial time series data, the article also detailed in this paper, the model of diagnosis and evaluation methods, such as the multicollinearity test, heteroscedasticity test, autocorrelation test, such as the application part, In the application section, this paper shows the specific application of linear regression model in stock return prediction, bond yield curve analysis, exchange rate prediction and other fields through actual case studies, and discusses how to combine other advanced statistical methods with machine learning algorithms to improve the predictive ability of linear regression model. Finally, This article summarizes the linear regression model in the financial market forecast, development trend and the future research direction of financial practitioners and researchers to provide valuable theoretical guidance and practical reference.

Keywords: Linear regression; Financial market forecasting; Model diagnosis; Time series analysis; Machine learning.

1. Introduction

In linear regression model is the basis of statistics and econometric tools and plays an important role in the financial market forecast of big background, because of the complexity of the financial market rising and uncertainty to predict market movements and asset price changes has become the great challenge of investors and policy makers, Fortunately, linear regression model has become one of the most important methods for financial market prediction because of its simplicity, ease of implementation and interpretation.

In recent years, the financial market has developed rapidly and the trend of diversification is obvious. According to the data of the World Bank, the total market value of the global stock market in 2016 was 65.7 trillion US dollars, and it will grow to 93.7 trillion US dollars in 2020 with an average annual growth rate of 9.3%. During the same period, the size of the global bond market also increased from USD 92.2 trillion in 2016 to USD 123.5 trillion in 2020 with an average annual growth rate of 7.6%. Such a rapid growth trend of the financial market makes the forecast of the financial market more and more important.

Linear regression model has a wide range of applications in financial market prediction, including stock return prediction, bond yield curve analysis, exchange rate prediction, commodity futures price prediction and many other fields. However, financial market data have special characteristics such as high frequency, non-stationary and non-linear, which makes the application of linear regression model face challenges. So further to explore the theoretical basis of the linear regression model and explore the practical application of it in the financial market, has important theoretical and practical significance.

The basic principle of the linear regression model, assumptions, parameters estimation method, and evaluation index by first introduced in this paper, which lay a theoretical foundation for subsequent application analysis, and then the article will explore the linear regression model in the stock market return, bond yield curve prediction, foreign exchange rate forecasting and commodity futures prices forecast financial market segments such as how specific application, This paper with

theory and actual case analysis as financial practitioners and researchers to provide theoretical guidance and practical reference value, and also for the further improvement and development of financial market forecasting method.

2. Theoretical basis of linear regression model

2.1. Basic principles of linear regression model

Linear regression model is a statistical method used to analyze the linear relationship between the dependent variable and one or more independent variables. Its basic principle is based on the assumption that there is a linear relationship between the dependent variable Y and the independent variable X, and the relationship is described by estimating the model parameters. Simple linear regression is the simplest linear regression model, and its mathematical expression is as follows:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

In this formula, the dependent variable is represented by Y, the independent variable is represented by X, the intercept term is β_0, the slope (regression coefficient) is β_1, and the random error term is ε. For the multiple linear regression model, the general form can be expressed as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Here, X_1, X_2,... , X_k denotes k independent variables, β_1, β_2,... , β_k is the corresponding regression coefficient.

The core idea of linear regression model is to estimate the model parameters by minimizing the sum of squared errors between the observed and predicted values. This method is called least square method (OLS), which can obtain unbiased and minimum variance parameter estimates[5].

2.2. Assumptions of the linear regression model

For a linear regression model to be valid and reliable, the following key assumptions must be met.

1. Linearity: the existence of a linear relationship between the independent variable and the dependent variable is the basic premise of the linear regression model. If the actual relationship is nonlinear, the nonlinear regression model or variable transformation may be considered.

2. Independence: the observations are independent of each other and there is no autocorrelation. This assumption is particularly important in time series data, after all, financial market data often have time dependence.

3. Homoscedasticity: the variance of the error term is constant and does not change with the change of the independent variables. If this assumption is violated, it is called heteroscedasticity and may reduce the efficiency of parameter estimation.

4. Normality: The normal distribution is the distribution that the error term follows. This assumption is particularly critical for parameter estimation and statistical inference in the case of small samples, while the central limit theorem can ensure the validity of the estimation even if the assumption of normality is violated in the case of large samples.

5. No multicollinearity: There is no perfect linear correlation or high correlation among independent variables, because the presence of multicollinearity will make the parameter estimation unstable and increase the standard error.

6. Correct model setting: all important independent variables are included in the model and irrelevant variables are not included. If the model setting is wrong, biased estimates may occur and wrong conclusions may be drawn.

Financial market forecast, many assumptions are often affected by challenges, like the financial time series data often have heteroscedasticity and the correlation, so need more complex, like the ARCH and GARCH model to deal with, so ah, predicted with the linear regression model to do the financial markets, have to be careful inspection these assumptions, And if necessary, appropriate correction methods should be adopted.

2.3. Parameter estimation method of linear regression model

In the construction of linear regression models, the least squares method (OLS) is the most commonly used in the key step of parameter estimation. However, there are also many other important estimation methods, such as maximum likelihood estimation (MLE) and generalized moment estimation (GMM). The following will focus on the basic principles of least squares method.

Least squares method to make observations and forecast error variance between minimum as core idea, is a simple linear regression model Y = beta zero + beta ı X + epsilon, its by looking for zero beta and the estimate of beta ı let to minimize the sum of squared residuals to achieve their goals, specifically is to minimize the following function.

$$Q = \sum(Y_i - \beta_0 - \beta_1 X_i)^2$$

The estimated values of β_0 and β_1 can be obtained by taking partial derivatives of β_0 and β_1 in Q and setting them equal to zero, and the multiple linear regression model can be expressed in matrix form and solved for parameter estimation.

The parameter estimation to the data in the financial market particularity into consideration, like the time series data may have to the generalized least squares (GLS) to tackle the problem of autocorrelation and heteroscedastic data can use weighted least squares (WLS) or heteroscedasticity robust standard error, in addition, in the face of the high-dimensional data or multicollinearity problem, In addition, regularization methods such as ridge regression and LASSO can improve the stability of parameter estimation and forecasting performance.

2.4. Evaluation index of linear regression model

The key to ensuring the effectiveness and reliability of linear regression models lies in the evaluation of their performance. The following evaluation indicators are commonly used:

Table 2-1 Main evaluation indicators of linear regression models

Indicators	Description	Calculation Method	Range of values	Optimal value
R^2 (coefficient of determination)	Measure how well the model explains variation in the dependent variable	1 - SSE/SST	$0 \leq R^2 \leq 1$	Close to 1
Adjust R^2	Consider the R-squared modified version of the number of independent variables	1 - (1-R^2)(n-1)/(n-k-1)	Can be negative	Close to 1
RMSE (root mean square error)	Standard deviation of forecast error	$\sqrt{(\Sigma(Y_i - \hat{Y}_i)^2 / n)}$	≥ 0	Close to 0
MAE (Mean absolute error)	The average absolute value of the forecast error	Σ	$Y_i - \hat{Y}_i$	/ n
AIC (Akaike Information Criterion)	Balance model complexity and goodness-of-fit	2k - 2ln(L)	No limit	Smaller value

Where, the sum of squared residuals is denoted by SSE, the sum of squared total deviations is denoted by SST, the sample size is n, the number of independent variables is k, and the value of the likelihood function is L[6].

The goodness-of-fit, predictive ability and complexity of the model will be evaluated by these indicators from different dimensions. In addition to the above indicators, evaluation indicators such as Sharpe ratio and information ratio, which are special to the financial field, are often used in financial market prediction, and appropriate evaluation indicators must be selected according to specific prediction objectives and application scenarios. Such as stock yield prediction may be more important how accurate predict whether direction rather than the absolute value accurately.

In addition, in the extremely dynamic and uncertain environment of financial markets, cross-validation and out-of-sample prediction are very important methods to evaluate model performance,

and the segmentation of data sets into training sets and testing sets can better evaluate the generalization ability and prediction performance of models.

3. The linear regression model in the application of the financial market forecast

3.1. Stock market return forecast

Linear regression model is widely used in stock market return prediction. Here is a simplified multi-factor linear regression model that can be used to forecast stock market return.

Table 3-1 Examples of stock market return prediction models

Dependent variable	Independent variables	Model form
Stock market returns	Market risk premium, size factor, value factor, momentum factor	$R = \alpha + \beta_1 MKT + \beta_2 SMB + \beta_3 HML + \beta_4 MOM + \varepsilon$

In this model, the stock market return is represented by R, the market risk premium is MKT, the size factor is SMB, the value factor is HML, and the momentum factor is MOM. This multi-factor model is based on the Fama-French three-factor model and Carhart four-factor model, which can capture the impact of different risk factors on the stock return.

In practical application, researchers may add other factors such as macroeconomic indicators and industry characteristics according to specific circumstances. For example, A study on China's A-share market shows that the predictive ability of the model will be greatly improved by adding factors such as price-earnings ratio (P/E) and price-to-book ratio (P/B) from 2016 to 2020. R squared up to 0.72 from 0.65.

However, linear regression models also encounter some challenges when forecasting stock market returns, because the nonlinear and non-stationary characteristics of the stock market may bias model predictions, and the rapid changes in the financial market may weaken the predictive power of historical data, so the model parameters should be updated regularly when applying the linear regression model. At the same time, other technical analysis or machine learning methods can improve the prediction accuracy.

3.2. Bond yield curve forecasting

In bond yield curve forecasting, linear regression model plays an important role, and the yield curve used to describe the different period, the relationship between bond yields in the fixed-income market analysis and forecast is an important tool, and Nelson - Siegel model the wide application of the yield curve modeling methods can describe as a linear regression model.

$$y(\tau) = \beta_0 + \beta_1 \left[\frac{1 - e^{-\lambda\tau}}{\lambda\tau}\right] + \beta_2 \left[\frac{1 - e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\right] + \varepsilon$$

In this model, the time limit for tau bond yields by y (tau) and beta zero, beta 1, beta 2 to estimate the three parameters, and a shape parameter lambda, with different period of bond yields can estimate the data regression analysis these parameters and then build a complete yield curve.

In recent years, the rapid development of the bond market has made yield curve prediction more and more important. According to the data of the Bank for International Settlements (BIS), the size of the global bond market was 92.2 trillion US dollars in 2016, and it will grow to 123.5 trillion US dollars in 2020 with an average annual growth rate of 7.6%. Accurately predicting the yield curve to the investment decision-making, risk management and monetary policy is of great significance.

However, bond yield curve forecasting presents some challenges because a variety of factors such as the economic cycle, inflation expectations and monetary policy may affect the shape of the yield curve, which can be difficult to fully capture in linear models, and traditional linear models may not perform well in very low or negative interest rate environments. Therefore, in practice, researchers

often have to combine other nonlinear methods or machine learning algorithms to improve the prediction effect.

3.3. Foreign exchange rate prediction

It is well known that the linear regression model is widely used in foreign exchange rate forecasting, and the linear regression model based on the interest rate parity theory is a commonly used exchange rate forecasting model.

$$\Delta S(t+k) = \alpha + \beta\big(i(t) - i^*(t)\big) + \varepsilon(t+k)$$

Here, the exchange rate change from t to t+k is represented by ΔS(t+k), the domestic and foreign interest rates are represented by i(t) and i*(t) respectively, and the parameters α and β need to be estimated.

Interest rate parity theory is the basis of the model and the model hypothesis rates vary as foreign currency exchange difference for the future, but the empirical study shows that the simple linear model in short-term prediction often predict ability is poor, so the researchers generally will be added in the model as inflation rate difference, the difference of economic growth, the other variables, such as balance of trade to improve the prediction accuracy.

Foreign exchange market volatility has increased in recent years, the bank for international settlements (BIS) data showed that average daily trading volume in 2016, global foreign exchange market is $5.1 trillion and to increase to $6.6 trillion, 2019, the growth trend to predict currency becomes more important and more challenging.

In practice, there are several major challenges in foreign exchange rate forecasting that pose to linear regression models.

1. Nonlinear relationships: The relationship between exchange rates and their influencing factors may be nonlinear, so simple linear models may not capture this complexity.

2. Structural changes: There may be structural changes in the foreign exchange market, such as policy adjustments or financial crises, which may cause model parameters to become unstable.

3. High-frequency data: The foreign exchange market operates 24 hours a day, which makes the processing and modeling of high-frequency data face new challenges.

Therefore, when researchers use linear regression models to forecast foreign exchange rates, they often need to combine techniques such as time series analysis and machine learning algorithms to improve the accuracy and stability of the forecast.

3.4. Commodity Futures Price prediction

Linear regression model has an important application in commodity futures price prediction, and its typical price prediction model may include the following variables:

$$F(t,T) = \alpha + \beta_1 S(t) + \beta_2 I(t) + \beta_3 R(t) + \beta_4 V(t) + \varepsilon$$

Among them, the futures price of term t at time T is represented by F(t, T), the spot price is S(t), the inventory level is I(t), the interest rate is R(t), and the market volatility index is V(t)[2].

Futures pricing theory is the basis of the model and the main factors affecting futures prices are taken into account. However, different commodity prices may be affected by different factors. For example, crude oil futures have to take into account geopolitical risk, OPEC production decision and other factors, while agricultural futures have to take into account weather conditions, harvest expectations and other factors.

In recent years, the trading volume of global commodity futures market has been increasing. According to the data of the World Association of Futures Exchanges (FIA), the trading volume of global commodity futures and options contracts in 2016 was 4.6 billion, but in 2020, it has reached 7.3 billion, with an average annual growth rate of 12.2%. This growth trend reflects the increasing importance of commodity futures price forecasting.

Researchers should pay attention to the following points when linear regression model is used to predict commodity futures price:

1. Seasonality: Many commodity prices have obvious seasonal characteristics, so the model should be properly handled.

2. Extreme events: Commodity prices can fluctuate dramatically due to extreme events such as natural disasters and policy changes, and this effect may not be captured by linear models.

3. Market sentiment: Short-term prices may be significantly affected by speculation and market sentiment, which are difficult to quantify directly and incorporate into linear models.

4. Cross-market correlations: There may be complex correlations between different commodity markets, so more aspects and interactions between markets need to be considered.

In practice, to enhance the predictive power of linear regression models, researchers often combine methods such as time series analysis and machine learning. Moreover, studies have shown that combining ARIMA models with linear regression models can greatly improve the accuracy of commodity futures price prediction, and if the prediction is to be effective all the time, it is also important to regularly update the model parameters and re-evaluate the model structure.

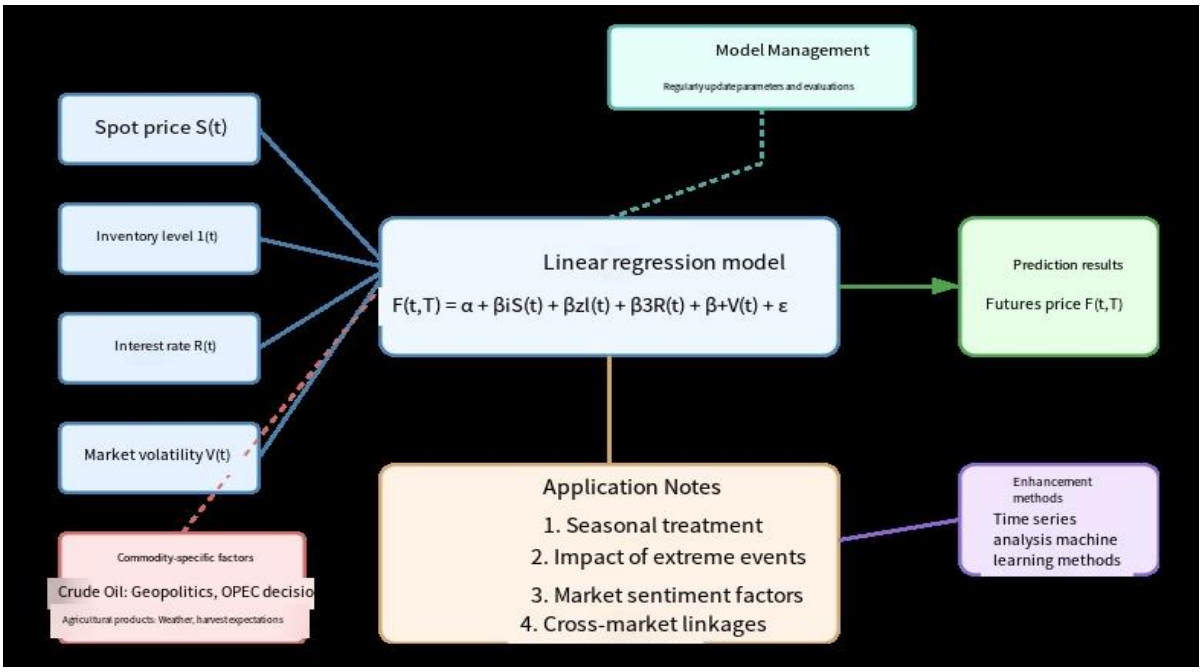

Figure 1 Linear regression prediction model for commodity futures prices

4. Optimization and improvement of linear regression model in financial market prediction

4.1. Multiple linear regression model

There are many extensions of linear regression model, one of which is multiple linear regression model, which can use multiple independent variables to predict the dependent variable, and is widely used in financial market prediction to analyze the impact of multiple economic factors on financial asset prices or yields. Its basic form can be expressed as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

The basic form can be expressed as follows: Y such as stock return is used as the dependent variable, and X_1, X_2,... , X_k are k independent variables, the intercept term is β_0 and the regression coefficients of the respective variables are β_1, β_2,... , β_k, and the error term is ε[4].

In the financial market prediction, the multiple linear regression model has the advantage of considering multiple influencing factors at the same time to make the prediction more comprehensive, just like the prediction of stock return, the company's financial indicators, macroeconomic indicators, market sentiment indicators and many other factors can be taken into account, so as to capture the interaction between different economic variables, so that the prediction is more accurate.

However, when using multiple linear regression models, we should be careful of the problem of multicollinearity, that is, high correlation between independent variables may make the estimation of regression coefficients unstable and affect the predictive ability of the model. To solve this problem, variable selection methods (such as stepwise regression or principal component analysis) can come in useful. Because they can pick out the combination of variables with the strongest predictive power.

4.2. Ridge regression and LASSO regression

Ridge regression and LASSO regression, two regularization regression methods, are often used to solve the problems of multicollinearity and overfitting in multiple linear regression, and have important applications in financial market forecasting, especially in dealing with high-dimensional data and improving the generalization ability of models.

Ridge regression relies on adding L2 regularization term to the objective function to control the complexity of the model, and it imposes a penalty on the regression coefficient to reduce the absolute value of the coefficient to reduce the over-fitting of the model to the training data. One of its advantages is that it can deal with the highly correlated issues among independent variables to improve the stability of the model.

LASSO (LeastAbsoluteShrinkageandSelectionOperator) regression using L1 regularization item, can not only reduce coefficient value, and can be crushed the variable coefficient of some less important to zero to achieve the effect of variable selection, especially useful when financial market forecast, Because it can independently identify the most predictive variables, it simplifies the model and improves the explainability.

In practice, researchers can pick the right regularization method according to the specific problem. For example, when predicting stock returns, LASSO regression can help to screen out the most critical one if there are many potential predictors, and Ridge regression may be more suitable for building macroeconomic forecasting models to deal with complex correlations among economic indicators.

4.3. Time series regression models

When linear regression models are used to deal with time series data, they become time series regression models. This is a special case that plays an important role in financial market forecasting, and such models take into account the time dependence of the data, so as to capture the trends, seasonality and cyclicality that often exist in financial markets.

Autoregressive integrated moving average (ARIMA) model is one of the most commonly used time series regression models. It can effectively process non-stationary time series data by combining the three parts of autoregressive (AR), difference (I) and moving average (MA). In the forecasting of financial market, it is often used to predict time series data[8] such as stock prices, exchange rates and interest rates.

Another important time series regression model is vector autoregressive (VAR) model, which can process multiple interrelated time series variables at the same time, and has excellent performance in analyzing the dynamic relationship between macroeconomic variables and predicting financial market variables. For example, the interaction between stock market, bond market and foreign exchange market can be analyzed by VAR model.

In addition, in the financial market forecasting, conditional heteroskedastic models such as GARCH (generalized autoregressive conditional heteroskedastic) model are very important tools, especially in the financial asset return volatility forecasting is very suitable, and can grasp the phenomenon of volatility aggregation in the financial market to improve the effectiveness of risk management.

4.4. Integration of machine learning methods

The rapid development of big data and artificial intelligence technology has made it a major trend to integrate machine learning methods with traditional linear regression models in the field of

financial market prediction. This fusion can not only improve the accuracy of prediction, but also cope with more complex nonlinear relationships and handle large-scale data sets.

Support vector machine (SVM) regression, a commonly used machine learning method, deals with nonlinear relationships by mapping data into high-dimensional space. It can be used to deal with complex prediction tasks such as stock prices and option pricing in financial market prediction, and is particularly good at dealing with nonlinear and high-dimensional data.

In the field of financial market prediction, decision trees and random forest algorithms are also widely used because they can automatically identify important features, deal with missing values, and capture complex interactions among variables. For example, random forest can predict stock returns, which integrates the results of multiple decision trees to improve the stability and accuracy of prediction.

Long short term memory network (LSTM), a deep learning method, is excellent in processing time series data and can capture long-term dependencies, which is suitable for complex financial tasks such as predicting stock price movements and assessing credit risk.

Ensemble learning methods, such as GradientBoosting and XGBoost, which combine multiple weak learners to build powerful prediction models, are excellent in financial market prediction, especially when dealing with high-dimensional features and nonlinear relationships.

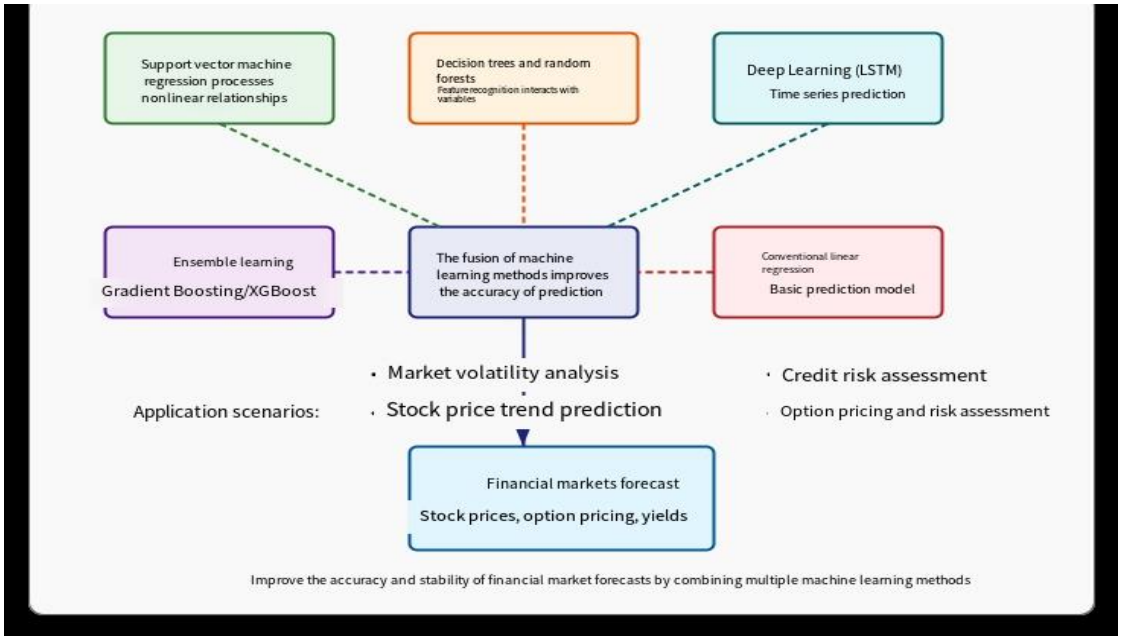

Figure 2 Integrated application of machine learning methods in financial market prediction

5. Conclusions

The theoretical basis, application methods and optimization and improvement strategies of linear regression model in financial market prediction are deeply explored in this study. After analyzing multiple linear regression, ridge regression, LASSO regression, time series regression model and machine learning methods, the following important conclusions are drawn:

Linear regression model is the basis of the financial market forecasting tools, has a theory is simple and strong explanation strength, and multivariate linear regression model to multiple influencing factors into account, along with all bring financial markets predict comprehensive analysis framework, but in actual use must pay attention to handle the multicollinearity problems such as, Otherwise, the stability and predictive ability of the model can not be guaranteed.

Second, the traditional linear regression model fitting and variable selection problem by ridge regression and LASSO regression regularization method to solve effectively, and deal with high-dimensional data and improve the model generalization ability performance outstanding, these methods to forecast provides a more robust financial markets, more precise tools.

Moreover, time series regression models such as ARIMA and VAR models take the time dependence of data into account to better grasp the dynamic characteristics of financial markets, and have unique advantages in predicting time series data such as stock prices, exchange rates and interest rates.

Finally, the integration of traditional linear regression models and machine learning methods opens up a new direction for financial market forecasting. Advanced algorithms such as support vector machine, random forest and deep learning can deal with more complex nonlinear relationships and large-scale data sets, thus improving the accuracy and applicability of forecasting.

References

[1] Lu Xiaojun; Cheng Changjie; Application of Combination Model Based on Multiple Linear Regression and ARIMA in Hot Rolling Spot Price Prediction [J]. Information and Computer (Theory Edition),2022(05):11-13.

[2] Tan Yaochen; Forecasting and Exploring the Application Degree of CSI 300 Stock Index Options in the Capital Market -- Based on S& Multiple linear regression model of P500 stock index options [J]. Finance and Economics,2020(13):129-131.

[3] Li Xiaoning; Application of Multiple linear regression and time series Model in stock prediction [J]. Science and Technology Entrepreneurship Monthly,2019(02):157-159.

[4] Xin Dong-sheng; Wang Meifang; Ma Ying-hua; Liu Hui; Application of Linear Regression Analysis Forecasting Model in Footwear Market Demand Forecasting [J]. Western Leather,2017(19):20-21.

[5] Sun Hao; Song Pingping; Economic forecasting ability of term structure of interest rates: a quantitative analysis [J]. Shanghai Finance, 2017(06): 11+19-24.]

[6] Cheng L Juan; . Based on the part of the functional improvement of linear regression model [J]. Journal of statistics and decision, 2017 (11): 72-74.

[7] Zhong L Yan; Gao Shulan; Application of multiple linear regression model in the analysis and prediction of housing price trend [J]. Science and Technology Entrepreneurship Monthly,2017(09):100-102.

[8] Zhang Hua; Zhang Desheng; Huang Shijuan; Chang Zhenhai; Partial Linear Autoregressive Prediction Model Based on Wavelet and Its Application in Shanghai and Shenzhen Stock Markets [J]. Journal of Yanbian University (Natural Science Edition),2009(01):22-26.

[9] Liu Kun. Research on Stock Market Public Opinion Analysis and Its application based on Machine Learning [D]. University of Electronic Science and Technology of China,2023.