Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/59/20231134

Exploring the Impact of Different Linear Regression Approaches on Stock Forecasting

Yiding Yang[1,a,*]

[1]*School of Information, Renmin University of China, Beijing, China*
a. yangyiding636@ruc.edu.cn
**corresponding author*

Abstract: Since 2020, epidemics, the rise of tech stocks, stimulus measures and monetary easing have all had a huge impact on stock prices. With this in mind, this paper summarizes the historical background of stock price research, which has evolved at different stages from technical analysis to fundamental analysis to quantitative analysis and behavioral finance. At the same time, through a comparative study of multiple linear regression and univariate linear regression on stock price data from three Chinese securities firms, this paper finds that different stocks may require different prediction models to obtain more accurate prediction results. The significance of this study is that it provides investors with information about stock price fluctuations and trends, which helps to formulate more informed investment strategies and risk management. In addition, it helps economists to study the impact of stock market on macroeconomics. Overall, these results shed light on guiding further exploration of stock price prediction based on various models.

Keywords: univariate linear regression, multiple linear regression, stock prediction

1. Introduction

The COVID-19 outbreak in early 2020 had a huge impact on the global stock market. Many markets experienced dramatic volatility with rapid declines and rallies. Government embargo measures and market panic led to massive stock market declines. China's Shanghai Composite Index suffered a maximum drop of 18%, with thousands of stocks falling [1, 2]. These show the negative impact of the epidemic on stock prices. However, in further analyzing the fixed sectors, the study found that the returns of different sectors were not all negatively affected. For example, sectors such as materials and energy received positive shocks, while sectors such as transportation and auto parts manufacturing received negative shocks [3].

As a result of the previous years' epidemics, people have become more reliant on remote working, distance learning, online shopping and digital entertainment. Technology companies have been favored by investors as they provide the tools and platforms to support these activities. Cloud service providers, network infrastructure companies and collaboration tool providers have benefited as a result. Second, consumers are turning more to online shopping as physical storefronts are restricted. E-commerce platforms, logistics companies, and tech companies related to digital payments and supply chains have all seen growth opportunities. This has led to a rapid rise in tech stock prices. Coupled with the U.S.-China trade friction, highly-followed tech stocks experienced a greater decline in returns than low-followed stocks [4].

DOI: 10.54254/2754-1169/59/20231134

Many countries implemented stimulus measures to mitigate the impact of the epidemic on the economy, which also provided some support to the stock market. At the same time, many central banks implemented monetary easing and lowered interest rates to stimulate economic growth. The very beginning was the rise of technical analysis. Technical analysis was an important part of early stock price research. in the late 1800s and early 1900s, investors began to use charts, trend lines, and patterns to analyze stock price movements in order to predict future price direction. Then came the development of fundamental analysis. Over time, investors gradually began to focus on the fundamental data of a company, such as earnings, revenue, and price-to-earnings ratio. Fundamental analysis emphasizes the intrinsic value of a company as a way to assess the fair price of a stock. With the development of computer technology and mathematical methods, quantitative analysis has gradually been applied in the study of stock prices. Econometric methods are used to analyze the relationship between stock prices and other factors, as well as to predict market trends. Behavioral finance is then used to study the psychological and emotional factors of investors in the decision-making process. It reveals that investors are often influenced by emotions, which leads to irrational behavior in the market. In recent years, the emergence of big data and artificial intelligence has opened up new opportunities for stock price research. These techniques can help analyze large amounts of market data and discover patterns and trends, thus improving forecasting capabilities.

ARIMA (Autoregressive Integrated Moving Average Model) Model bases forecasts on past price data, taking into account the autocorrelation and moving average nature of time series data [5]. The GARCH(Generalized Autoregressive Conditional Heteroskedasticity) model focuses on predicting the volatility of stock prices. It takes into account the relationship between different time periods of stock price volatility and can be used to predict the degree of price volatility. The results of Herben Arashi's study show that the overall forecasting effect of the conditional heteroskedastic ARCH model is better than that of the autoregressive differential moving average ARIMA model [6]. Linear regression can be used to establish the relationship between stock prices and some fundamental factors (e.g., revenue, earnings, etc.) to predict future prices. SVR (Support Vector Machine Regression) can be used to classify and regress predictions by learning the relationship between stock prices and other related factors [7]. Deng Jiali, Zhao Fengqun, and Wang Xiaoman introduced the Artificial Ecosystem Optimization Algorithm (AEO) to select parameters in the SVR algorithm, which improved the prediction accuracy of the model [8]. Researchers use past data such as price and volume to train LSTM (Long Short-Term Memory Network) models and use them to predict future stock prices [9]. The application of LSTM neural network can significantly reduce the number of parameters to be learned in the network. A research team consisting of Yan et al. clarified the superior advantages of using LSTM temporal recurrent neural networks in terms of accuracy and data resource consumption when dealing with stock series data [10].

2. Data and Methodology

The data used in this study are the 300 stock price situations of three securities companies in the SSE 50 in 2022, namely 600030-SH CITIC Securities, 600109-SH Sinolink Securities and 600837-SH Haitong Securities. For each stock, 80% of the 300 data were selected as the training set, and the remaining 20% were selected as the test set. The names and meanings of various attributes of the original data are given in Table. 1. In this study, the closing price is set as the final y that needs to be predicted, and the most suitable subset of features is first selected using the analysis method of Pearson correlation coefficient. The results of the correlation analysis of the three companies are presented in Table. 2. From the magnitude of the correlation coefficient, it can be seen that only the opening price, the highest price and the lowest price are most correlated with the closing price, so these three variables are set to x1, x2 and x3 in turn.

DOI: 10.54254/2754-1169/59/20231134

Table 1: Data description.

Feature	Data structure	Descriptions
Codes	String	Code of the stock
Open	Double	Opening price
High	Double	Highest price
Low	Double	Lowest price
Close	Double	Closing price
UDamount	Double	Up and down amount
UDrate(%)	Double	Up and down rate
Volume	Int	Trading volume
TransAmount	int	Trading amounts
Amplitude(%)	Double	Amplitude of the volatility
Turnover(%)	Double	Turnover rate

Table 2: Correlation analysis.

Feature	CITICS	Sinolink	Haitong
High	0.9986	0.9996	0.9996
Low	0.9958	0.9993	0.9991
Open	0.9948	0.9991	0.9989
TransAmount	-0.045	0.0346	0.1247
UDrate	-0.0883	0.0261	0.0786
Amplitude	-0.0895	0.05	-0.0583
Turnover	-0.1007	0.0707	0.0927
Volume	-0.1451	-0.1428	-0.0185
UDamount	-0.1492	-0.0532	0.1677

3. Results and Discussion

The highest price of the day, which is the most relevant for all three companies, is chosen as the unique independent variable x. The model is $y=b_0+b_1x$. For CITIC Securities, the univariate regression equation is calculated as: $y=0.9771x-0.0117$. For Sinolink Securities, the one-dimensional regression equation is calculated as: $y=0.9835x-0.0016$. For Haitong Securities, the one-dimensional regression equation is calculated as: $y=0.9884x-0.0166$ As for Multiple Linear Regression Model, the model can be expressed as $y=b_0+b_1x_1+ b_2x_2+b_3x_3$. For CITIC Securities, the multiple regression equation is calculated as $y=-0.332x_1+ 0.7432x_2+0.58x_3+0.0203$. For Sinolink, the multiple regression equation is calculated as $y=0.6591x_1+ 0.3362x_2+0.012x_3-0.0028$. For Haitong Securities, the multiple regression equation is calculated as $y=-0.1769x_1+ 0.795x_2+0.3776x_3+0.0035$. The comparison between the predicted data of these three companies and the actual data is shown in the Fig. 1, where the blue line represents the actual value and the orange line represents the predicted value.

DOI: 10.54254/2754-1169/59/20231134

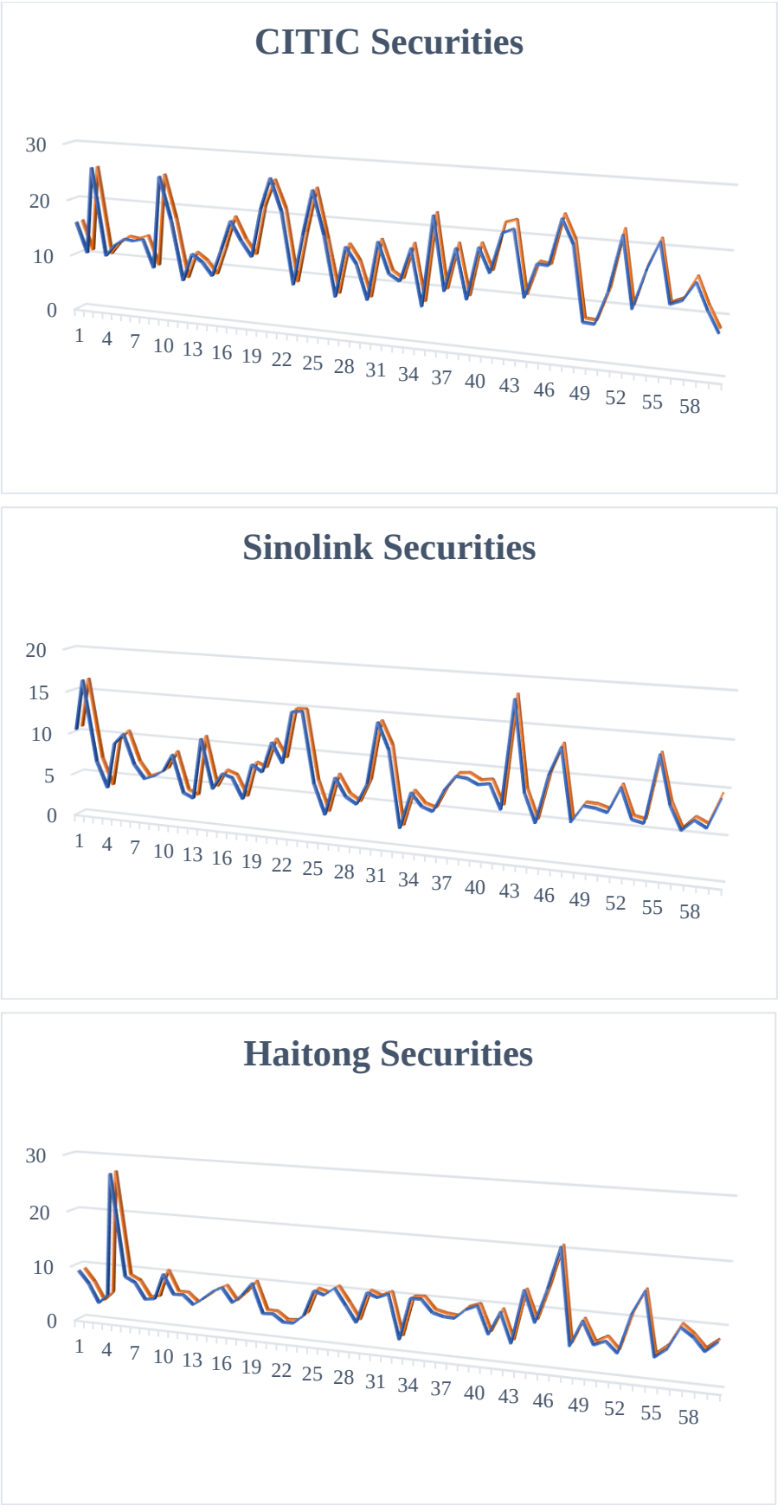

Figure 1: Price prediction results.

DOI: 10.54254/2754-1169/59/20231134

In this study, four metrics, mean absolute error, root mean squared error, relative absolute error, root relative squared error, were chosen to measure the results of the simulation, and the formulas for each metric are as follows:

$$\text{MAE} = \frac{\sum_{i=1}^{n}|p_i - a_i|}{n} \tag{1}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}} \tag{2}$$

$$\text{RAE} = \frac{\sum_{i=1}^{n}|p_i - a_i|}{\sum_{i=1}^{n}|\bar{a} - a_i|} \tag{3}$$

$$\text{RRSE} = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{\sum_{i=1}^{n}(\bar{a} - a_i)^2}} \tag{4}$$

Here a is the actual target while p is predicted target. From the numerical magnitude of each error in the Table 3, it can be seen that for both stocks CITIC and Haitong, the prediction error of multiple regression is significantly smaller than the error of univariate linear regression, while for Sinolink, the error of univariate regression equations is smaller and the simulation is better, and the conclusions of the three stocks are not exactly the same.

Table 3: Evaluation Metrics.

	Method	MAE	RMSE	RAE	RRSE
CITICS	Univariate	0.1615	0.2223	0.0336	0.0391
	Multiple	0.1011	0.1459	0.0210	0.0256
Sinolink	Univariate	0.0471	0.0601	0.0219	0.0234
	Multiple	0.0471	0.0645	0.0220	0.0251
Haitong	Univariate	0.0699	0.0924	0.0274	0.0285
	Multiple	0.0584	0.0850	0.0230	0.0262

4. Limitations and Prospects

In this experiment, the most significant issue encountered is multicollinearity. This leads to several potential problems.

Firstly, multicollinearity results in inaccurate parameter estimation. When there is a high degree of correlation between the independent variables, the model will have difficulty in distinguishing the independent effects of each independent variable, which leads to unstable parameter estimation. This makes it difficult to explain the contribution of each independent variable to the dependent variable in the model.

Secondly, multicollinearity makes it difficult for the model to accurately explain the variability in the relationship between the independent variables and the dependent variable, as a portion of the variability may be explained by more than one highly correlated independent variable, making it difficult to understand the actual impact of the respective variables.

Lastly, in the presence of multicollinearity, the model may perform erratically in predicting new data because it relies on the relationships between the independent variables in the training data.

To address the issue of multicollinearity in subsequent research, several recommended approaches can be considered.

DOI: 10.54254/2754-1169/59/20231134

One effective strategy is feature selection, where highly correlated independent variables are identified and removed from the analysis.

Another method is principal Component Analysis (PCA), which can be employed to transform the original independent variables into a set of uncorrelated principal components. This approach helps alleviate the issue of multicollinearity by reducing the correlation among the variables.

Increasing the sample size is also a viable option. A larger sample size reduces the effect of multicollinearity on parameter estimation.

Additionally, regularization methods such as ridge regression and LASSO can be used to introduce penalty terms in the model to help control for problems caused by multicollinearity.

5. Conclusion

To sum up, this study has performed multiple linear regression and univariate linear regression for each of the three stocks and the results show different predictive effects. Specifically, for the stocks CITICS and Haitong, the multiple linear regression models show better predictions with smaller prediction errors that are closer to the actual observations. This suggests that in the case of these stocks, the multiple regression model is able to more accurately capture the complex relationship between multiple independent variables and the dependent variable, providing more reliable predictions. However, for the stock Sinolink, the univariate linear regression model presents better predictions with smaller errors. This may be due to the fact that Sinolink's data characteristics make the linear relationship between a single independent variable and the dependent variable more significant, thus allowing the univariate regression model to show an advantage in prediction. In summary, the results of this study suggest that the predictive effectiveness of multiple linear regression and univariate linear regression models may differ in different stock contexts. Investors and analysts should make decisions based on stock-specific data characteristics and relationships when selecting forecasting methods to ensure more accurate and effective forecasting results.

References

[1] *Duan, Y. (2020) The Impact of the New Crown Pneumonia Epidemic on China's Stock Market: An Empirical Analysis Based on the Pharmaceutical Industry. China Business Journal, 18, 28-30.*

[2] *Yin, Z. (2020) China's Stock Market Under the Influence of Epidemic. China Economic Review, 1, 44-47.*

[3] *Jing, T. (2022) Impact of the New Coronavirus Epidemic on China's Stock Market Analysis- Based on Stock Market Returns of Different Industries. Northwest University of Agriculture and Forestry.*

[4] *Guo, X. (2020) Study on the Impact Of China-US Trade Friction on the Return of Technology Stocks-A Perspective from Investors' Limited Attention. University of International Business and Economics.*

[5] *Zhang, L. (2023) Research on the Prediction and Heterogeneity of Chinese Stock Prices Based on Artificial Intelligence. Journal of Jinan (Philosophy and Social Science Edition), 3, 123-132.*

[6] *He, B. (2008) Optimal Choice Model for Stock Price Forecasting. Statistics and Decision Making, 6.*

[7] *Xiao, J.H., Zhu, X.H., and Huang, C.X., et al. (2018) A New Approachfor Stock Price Analysis and Prediction Based on SSA and SVM. International Journal of Technology &Decision Making, 1-17.*

[8] *Deng, J., Zhao F., and Wang X. (2022) MTICA-AEO-SVR Stock Price Prediction Model. Computer Engineering and Applications, 8, 257-263.*

[9] *Su, D. (2022) Stock Price Prediction Based on Lasso Method and LSTM (Master's thesis, North China Electric Power University.*

[10] *Peng Y., Liu Y.H. and Zhang R.F. (2019). Modeling and Analysis of Stock Price Prediction Based on LSTM. Computer Engineering and Applications, 11, 209-212.*