

A Stock Market Prediction Model Based on Linear Regression and Matrix Method

Muchen Liu

Shandong Experimental Middle School, Jinan, Shandong, 250001, China
tian_lin666@outlook.com

Abstract: Stock market predictions, being one of the most important areas in financial studies, have for long been sought after by many researchers in these areas, with their applications being of utmost importance in studies on theory. Here, in an endeavor to bring about an efficient yet interpretative model for stock market predictions, there has been an impressive amalgamation of approaches based on linear regression models, matrix theory, and error correction mechanisms. Here, in particular, while model parameters have been identified with utmost precision based on least squares, there have been improvements with regard to computational speeds based on matrix calculation, with improved market versatility for resistance to dynamic markets based on error correction mechanisms. On testing applications, there has been an analysis based on historical Google stock market performances. It has been determined in these studies, based on their applications, that there has indeed been remarkable stability in performance levels in identifying stock performance based on its predictive nature, while particularly emphasizing their high flexibility levels in real-time applications based on its localization levels in its sliding window-based method. Though there have been limitations with regard to response levels based on short-term market changes, there have indeed been many advantageous levels in computational simplicity, interpretability, and versatility based on its applications in predictive conclusions for financial dealers. These studies have indeed paved new ways in achieving efficient fusion in financial predictive studies, with appropriate applications in their development in computational finance based on innovations brought about in finance theory based on modern computing innovations.

Keywords: Linear regression model; Stock market prediction; Least squares estimation.

1. Introduction

Of late, there has been too much instability in the international economic environment, in addition to drastic changes in financial markets. These changes in stability, along with changes in financial markets, have clearly manifested themselves in many ways, such as friction in trade, and unexpected occurrences in general public health, which have notably affected the normal flow in financial markets [1]. Now, considering the tariff dispute in the US, for example, such occurrences happen too often, thereby setting up an unexpected environment for financiers to invest. Change in tariffs not only affects those involved in import, export, cost, profit, difference, but also affects many other upstream, downstream businesses due to chain reactions, thereby finally resulting in its impact on stock prices for overall industries [2]. It is really difficult for financiers to invest according to rules in financial markets. Conventional mathematical models for predictions, along with actuarial models, have proved incompetent to fulfill such complex conditions concerning societal environments.

Conventional predictive models rely on some general assumptions, such as model patterns, which are not general in nature in a dynamic market environment in many cases. For example, some models like linear regressions, based on some historical data, assume some relationship between some variables in a linear fashion. But, in real-life scenarios in a market environment, these relationships could be different in different ways [3]. Although some actuarial models have considered some risk factors to some extent, these constraints have become apparent while addressing some unconventional risks in highly uncertain market environments. These models, in addition to not being able to successfully reduce some

investment risks, could increase these risks due to late responses to market changes [4].

With such a background, stock market prediction, being an important research hotspot in finance, has attracted more and more attention. Stock market, being an economic barometer, its stock price changes contain overall expectations for macroeconomic conditions, business performance, and policy elements. Scientific and rational stock price predictions not only provide an important reference for investors to manage risks, grasp opportunities, retain and increase their assets, but also provide an important reference for macro-control policy decision-making. For example, correct stock price predictions can guide policy decision-makers to understand policy acceptability by the market, thus improving policy reasonableness and efficiency, ensuring it for financial market sound and stable growth [5]. On the other hand, for financial institutions, achieving correct stock predictions aids in their asset management, product planning, increasing competitiveness, improving financial profitability, etc.

Presently, the common techniques for stock market predictions primarily include two types: statistic-based methods for time series, and nonlinear methods based on machine learning. On one hand, statistic-based models for time series, such as the Autoregressive Integral Moving Average Model (ARIMA) model, and Generalized Autoregressive Conditional Heteroscedastic model (GARCH), have achieved improved predictive accuracy to some extent with their respective distinct theory systems. On the other hand, the ARIMA model seeks to capture the linear features in time series data through autoregression, differences, or moving averages. Correspondingly, the GARCH model primarily applies to modeling aggregation features in financial data volatility, thereby being more

competent in addressing financial data with heterogeneous variances [6]. Nevertheless, in complex market data analysis, these models face innumerable problems, such as high computational complexity, low interpretability, and sensibility to noisy data. With increasing dimensions and complexity in market data, the computation complexity for ARIMA or GARCH models grows exponentially, resulting in low computational efficiency. Simultaneously, these models have complex parameter estimation, rendering result interpretations difficult for laymen to understand. Moreover, these models are more sensibility to noisy data, for which noisy data could significantly deteriorate predictive accuracy for these models [7].

Though models with nonlinear features, such as those built with neural networks or support vector machines, have the capability to extract nonlinear relationships from data, their accuracy on predictions is improved to some extent. Neural networks, which model the operating mode of nerve cells in the human brain with multiple layers, have the capability to extract data with complex nonlinear relationships through self-learning. Support vector machines classify data or launch predictions with regard to data via finding the optimal classification hyperplane [8]. But there are also some issues, such as complexity in model, long model training time, and interpretability, being involved in these models. For neural networks, their model structure is generally complex, with many parameters, which leads to consuming much computing time on model training and optimizing. At the same time, because these models are "black box" models, their decision-making process together with decision explanation are relatively difficult, especially for applications requiring high interpretability. For support vector machines, there is low computational efficiency for their operation on large-scale data [9].

2. Research Design

Previous studies conducted by Muhammad Farhan et al. used Samsung Electronics for demonstration purposes to further understand the use of linear regression in stock price predictions [10]. It has been observed in the research study that while it is possible to detect stock price patterns with the use of linear regression analysis, there are some inaccuracies in its predictions. But in contrast to other research studies, there is a major difference in the context of carrying out predictions with linear regression analysis with error correction. On the other hand, while carrying out predictions with error correction, there is also effort to explore other dimensions for incorporating artificial intelligence in stock market predictions.

Linear regression model:

$$Y = kX + \epsilon \quad (1)$$

where Y is the target to predict (e.g., closing price), X is an $n \times p$ feature matrix, k is a $p \times 1$ regression parameter vector, and ϵ represents the error term (such as pandemic shocks, tech stock surges, or policy interventions).

least squares estimation:

$$\min(Y - Xk)^T(Y - Xk) \quad (2)$$

Get the analytical solution:

$$k = (X^T X)^{-1} X^T Y \quad (3)$$

The matrix method directly solves for k , demonstrating its

unbiasedness and optimality conditions.

The estimated k pairs are applied to the new sample X_{new} calculate:

$$Y' = k X_{\text{new}} \quad (4)$$

If the forecast value Y is higher than the previous days closing price, it indicates a potential rise; otherwise, it may fall.

3. Data Processing

Concretely, in this research, historical stock information is precisely chosen as the object to be analyzed, including, in particular, such major indicators as closing prices, opening prices, highest prices, lowest prices, and trade volumes. These indicators encapsulate comprehensively all changes in prices, as well as trading activity, on particular stocks for particular moments in time, thereby being extremely informative for constructing efficient predictive models.

For ensuring the integrity and accuracy of data, in this research, there is also a complete and meticulous process for handling the original data. For one, there is an adopted normalization process for alleviating the differences in scales among data, thereby not being affected by such differences in training the model. Secondly, for addressing the possible missing data in the given information, there is also a need to apply appropriate interpolation techniques in filling in these gaps, thereby ensuring the connectivity and integrity of the information. Finally, there is an extraction process for representatives in feature variables via several techniques in conducting feature engineering for comprehending additional information from the data.

The stock market predictive model built with linear regression, matrix calculation, and methods in this research has many outstanding benefits. Looking at its calculation speed, it is impressive to note that the model skillfully uses matrix calculation to solve model parameters in one step via the least square technique. It should be emphasized that, in comparison with other traditional techniques, its calculation speed is greatly improved, there are no complex non-convex problems to be solved, and model parameters could be determined precisely, which is ideal in real-time stock market transactions with extremely high speed demands.

On the aspect concerning the reliability from statistics, under ideal conditions in which the error condition is defined through the normal distribution according to mean square error, there is substantial guarantee for predictive outcomes from least squares estimation, which remains least biased with minimum variance. Compared with other methods concerning stock market predictions, its properties in the proposed model for this research remain significantly more superior, having the capacity to preclude predictive inaccuracies in relation to biased estimations while maintaining predictive accuracy stability. Another significant benefit for such a model is its interpretability. It is possible for the regression coefficient to clarify in a logical manner the marginal contribution level of every characteristic for achieving the goal attribute (stock value or rate of return). According to the analysis on regression coefficients, investors could gain deep insight into which factor exerts crucial influence on stock value changes, thereby developing an investment decision with rational judgment to offset the risk caused by blind decision-making.

Considering typical strong linear relationship features in financial markets like the stock market, for example, the

linear relationship between historical prices, trade amounts, and short-term variations, the linear regression model will be in an optimal condition to utilize its simulation and predictive functions to capture inherent relationships among these features while at the same time preventing irregular predictions with regard to extreme values typical in previous models.

At the same time, scalability is also one of the crucial features for the model in this research. It is easy to incorporate regularization methods, like ridge regression and Lasso regression, into linear regression models. By creating a parameter size constraint process, it not only prevents problems such as model overfitting, which will impact its generalization performance on new data, but also solves problems such as multicollinearity to some extent. Its feasibility is extremely high because there is enough room for improvements in real applications, such as predictions in the stock market.

4. Research Examples

This study selects Google (stock code: GOOGL), a global

technology leader under the Alphabet Group, as the research object. With a market value exceeding 1 trillion US dollars, it holds a significant position in the global technology field. Google's business is widely involved in multiple cutting-edge fields such as search engines, cloud computing, artificial intelligence, and advertising services. Its business territory covers more than 190 countries and regions around the world, and it has a large and loyal user base as well as extensive and in-depth commercial influence. This vast global business ecosystem endows Google's stock with a strong risk-resistance capacity, enabling its share price to exhibit a distinct cyclical pattern when facing extreme market shocks, such as the 2008 financial crisis and other major events, despite fluctuations, rather than sharp and disorderly fluctuations or cliff-like drops. Therefore, Google stocks, as the research object, can provide representative and stable data support for the study of stock market prediction models, which is conducive to in-depth exploration of the performance of the models in different market environments.



Figure 1. The prediction of the changes in Google's stock stock based on the linear regression method

The data is sourced from the Kaggle platform and the historical data of Google's stocks collated by Adil Shamim

Figure 1 shows how the blue points accurately denote the real closing prices of Google's stock from 2004 to 2024. These blue points in Figure 1 depict the real-time market changes in Google's stock prices over these years. The other curve, which is the red line, shows the simulation curve produced by the linear regression model in attempting to discern the likely evolution in Google's stock prices based on its accuracy in real-time market changes in Google's stock prices in Figure 1. This discussion relies on concepts from the research document "Efficient Resampling Method for Non-Smooth Estimation Functions" by authors Deng Linzeng and Dong Yonglin to explore its applications in accurately assessing Google's stock prices.

Even with the not-so-healthy market condition prevailing in 2022, there was an increasing trend in the prediction curve for the linear regression model, which tended to slightly differ from the prevalent market trend. To counter such an issue, there would be an opportunity to modify and optimize such models in real-time through the inclusion of an error term, or in other words, to recreate a more precise model at such a particular point in time. It is to be noted that, in spite of being on an increasing trajectory from 2004 to 2024, representing the overall data for Google's stock performance, such an increase in Google's stock performance would not solely be due to uncertainties in model accuracy. It is to be remembered

that there are numerous complex market forces responsible for affecting stock performance, and there would inevitably be some minor discrepancies between such predictions in the model's output results and actual market trends.

A closer look at the data from 2020 to 2024 shows changes in Google's stock performance with remarkable volatility, which to some extent introduces natural errors (λ) into Google's stock performance predictions, in addition to revealing the imperfections in the present models with regard to predictive accuracy. Especially in scenarios involving drastic market changes due to major events such as the outbreak of pandemics or a swift recovery in tech stock prices, it's apparent that present models are less accurate in their predictive capacity. To overcome such limitations, future studies could be divided into two major areas: Firstly, more macroeconomic variables like growth rates in Gross Domestic Products, or GDP, inflation rates, consumer confidence index, and so on, in addition to market sentiment information derived from social channels such as social media, press opinion, and other such providers, could be included to increase its sensitivity for market changes; Secondly, nonlinear models: LSTM (Long Short-Term Memory Network Models) and XGBoost (Extreme Gradient Boosting Models) could be employed. These models could effectively capture prevailing market complexities in Google's market performance, thereby being more accurate in predictive capacity concerning drastic market changes on Google's stock

performance.

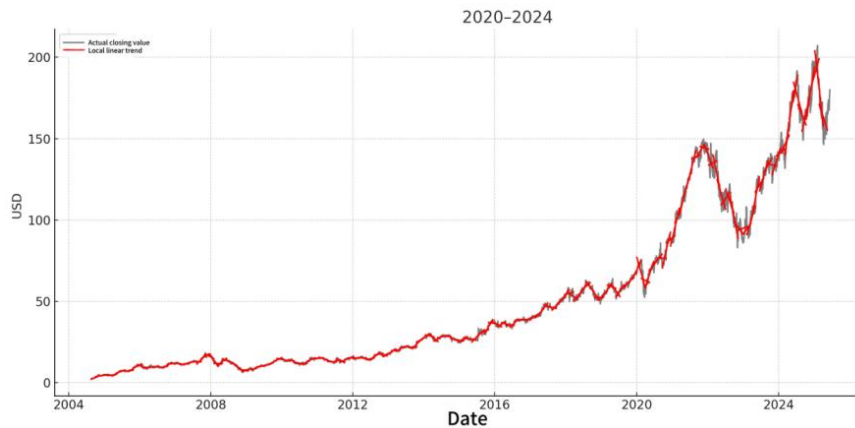


Figure 2. The prediction of Google's stock changes based on the linear regression method

The data is sourced from the Kaggle platform and the historical data of Google's stocks collated by Adil Shamim

As shown in Figure 2, the gray line in the figure clearly depicts the actual stock price trend of Google, while the red line represents the local linear trend fitted within each 60-day window. By pursuing the sliding window linear regression prediction model, it is discovered that there are more benefits for this model in comparison to those discussed above. Initially, with regard to response rates, the sliding window linear regression model is not slow to respond to extreme market situations and value changes, thereby giving investors ample opportunities to make necessary changes in their respective investment strategies once unforeseen events occur. Secondly, with regard to consistency in predictions, after conducting predictions based on different windows, it was observed that predictions based on the model, with every 60 days acting as its point, align with the stock value to some extent. It means that with small windows, predictions based on models are progressively more accurate, with their final fit also being ideal.

According to the above analysis, although there are some disadvantages in applying the least square method to stock market predictions, this model is able to provide a high degree of accuracy in illustrating stock market trends. By applying this model, stock market investors will be given reference information that will enable them to make rational decisions in their respective investments, taking into consideration their market environments.

5. Conclusion

By carrying out in-depth analysis, it has been found in this research work that there has been full confirmation on its feasibility and efficiency regarding the stock market predictions model based on linear regression analysis coupled with matrix methods, related to overall macro-level market trends. By conducting appropriate analysis, it has been concluded that, in comparison to large-scale market data, due to its inherent advantage in least squares computation in matrices, it possesses several strengths in being highly efficient and resulting in stable overall guesses in regard to market trends in itself, thereby being an advantage to provide concrete market references to market players in their effort to formulate overall market trends in their long-term investments.

Nevertheless, from an objective perspective, it is observed from this research work that there are some irregularities in the model regarding dealing with extreme market events such

as outbreaking epidemic markets or sudden changes in policies. It is implied from this point that there are some constraints in applying linear regression models to simulate extreme nonlinear changes in markets, especially while accurately estimating their outcomes on stock markets. To overcome these irregularities, based on research work being carried out in different studies, there will be further branching out in two major aspects in the future. On one hand, there will be some other macroeconomic variables being included in the model to make it more sensitive to external market changes, with the aim of comprehensively estimating dynamic changes taking place in markets. On the other hand, different models with their strengths will be comprehensively utilized to accurately simulate market changes.

Consequently, there have been continued applications for linear regression analysis and matrix methods in stock market predictions. Especially in applications concerning investment decision-making, faster response rates with improved interpretability are needed, with such analysis providing an orderly, understandable, and comprehensible predictive output to financial participants to enable them to understand key aspects in complex market information in their investments. Owing to the increasing evolution in the financial market environment, with improvements in data sciences, it is expected to note improvements in the stock market predictive model analysis for matrices in linear regression to enable further contributions to financial stability.

References

- [1] Sudjono, S. (2024). The Effect of Financial Globalization on Economic Instability: The Case of the Global Financial Crisis. *The Effect of Financial Globalization on Economic Instability: The Case of The Global Financial Crisis*, 1(1), 1-21.
- [2] He, F., Lucey, B., & Wang, Z. (2021). Trade policy uncertainty and its impact on the stock market-evidence from China-US trade conflict. *Finance Research Letters*, 40, 101753.
- [3] Goodwin, P., & Wright, G. (2010). The limits of forecasting methods in anticipating rare events. *Technological forecasting and social change*, 77(3), 355-368.
- [4] Embrechts, P., & Wüthrich, M. V. (2022). Recent challenges in actuarial science. *Annual Review of Statistics and Its Application*, 9, 119-140.
- [5] Haque, M. S., Amin, M. S., Miah, J., Cao, D. M., & Ahmed, A. H. (2023). Boosting stock price prediction with anticipated macro policy changes. *arXiv preprint arXiv:2311.06278*.

- [6] Yahaya, E., & Etuk, H. (2021). Comparative performance of ARIMA and GARCH model in forecasting crude oil price data. *AJPAS*, 15(4): 251-275.
- [7] Ismail, M. T., & Al-Gounmeein, R. S. (2022). Overview of Long Memory for Economic and Financial Time Series Dataset and Related Time Series Models: A Review Study. *IAENG International Journal of Applied Mathematics*, 52(2).
- [8] Pashankar, S. S., Shendage, J. D., & Pawar, J. (2024). Machine Learning Techniques for Stock Price Prediction-A Comparative Analysis Of Linear Regression, Random Forest, And Support Vector Regression. *Journal of Advanced Zoology*, 45.
- [9] Thanaki, J. (2018). *Machine learning solutions: expert techniques to tackle complex machine learning problems using python*. Packt Publishing Ltd., Birmingham.
- [10] Farhan, M., Ahmed, A., Eesaar, H., Chong, K. T., & Tayara, H. (2025). A novel approach to stock price prediction: averaging open and close prices with LSTM. *Digital Finance*, 7(3), 535-551.