Price Prediction in terms of Linear Regression For F, BYDDY and POAHY

Zhaozhilin Yan [1, *]

[1]Paul Merage School of Business, University of California, Irvine, the U.S.

*Corresponding author: zhaozhiy@uci.edu

Abstract. In the environment surrounding the capital markets, investors are keen on predicting the stock market. Many companies and research institutions are following stock price movements every day. As more and more models and algorithms are proposed, the accuracy of forecasting is also greatly improved. This paper mainly focuses on adopting linear regression models to predict the future trends of three stocks, BYDDY from A-shares, POAHY from European stocks, and F from US stocks. By building the linear regression model and calculating the R-squared, the accuracy of the model is above 90% and the MAE between the predicted value and the true value is very small. The correlation analysis shows that the variables Open, High, Low, and Adj Close have a positive correlation with the predicted values. In addition, there is a correlation between the three stocks. In addition to successfully predicting the trends of the three stocks, the areas of stock price fluctuations were also obtained. Forecasting using linear regression models is a proven way to make it easier to understand and calculate. These results shed light on verifying the feasibility of the linear regression model to predict stock prices.

Keywords: Linear Regression; Prediction; Stock Market.

1. Introduction

Stock price prediction is always a difficult topic, not only due to its mobility but also because it is subjected to politics, economics, unexpected events and other micro or macro elements. A lot of literatures have been developed around stock price prediction, and many variables have been found to influence stock markets social media, political events, social topics, etc. Taking social media as an example, when Elon Musk tweeted “Am considering taking Tesla private at $420. Funding secured.”, “Tesla’s stock price rose by 7%” [1]. Not only Twitter, but Facebook’s sentiment analysis also provides a wealth of information on stock market trends for researchers to refer to. In this paper has chosen three stocks, they are Ford in the US, Porsche in Europe, and BYD in the A-share. Ford Motor Company is one of the world's largest automakers. According to Statista, Ford's market share in the U.S. in 2021 is about 12.6 percent, trailing General Motors and Toyota Motor. Porsche is located in Germany and is known worldwide for its sports cars and supercars. BYD is based in southern China. Regarding to Phoenix, its new energy vehicles account for about 17% of the Chinese market. All three stocks are ranked among the top companies in the automotive business.

Many scholars have studied different models to predict stock market results in the past, including machine learning, neural network, etc. Although there is always a debate about whether the stock market can be predicted, most empirical studies have found that stock prices can be forecasted. A model for time series forecasting was proposed in the 1970s. This model is known as Auto Regressive Integrated Moving Average (ARIMA) model proposed by George and Gwilym et al. in the book "Time series Analysis: Forecasting and Control". In addition, Artificial Neural Network (ANN) models for machine learning were proposed by Warren McCulloch and Walter Pitts in the late 1940s, and a decade later Rosenblatt created the first artistic neural network, called a “perceptron” [2]. Although stock market prediction is possible, it is very difficult due to factors such as "the generally accepted, semi-strong form of market efficiency and the high level of noise" [3]. Wang and Lin applied artificial neural networks to stock market price prediction in order to observe the effect of trading volume on prediction [4]. They found that trading volume does not affect prediction performance.

Contemporarily, technology is rapidly developing and there are various models and algorithms to predict stock price trends, e.g., the Long short-term memory (LSTM) model and moving-average model (MA) model. Either model provides people with valuable information about the stock market. This article will use the linear regression model to predict the price of three stocks and determine their accuracy. It includes the correlation analysis between different variables, the observation of R^2 values to analyze the goodness of fit, the calculation of MAE to obtain the error between the predicted and actual values, and the coefficients of various variables. In addition, there is also a correlation analysis of the interaction among the three stocks. The rest part of the paper is organized as follows. The Sec. 2 will discuss methodology. The Sec. 3 will analyze results of model. Eventually, a brief summary is given in Sec. 4.

2. Methodology

2.1 Data

The data for the three stocks F, BYDDY and POAHY are obtained from Yahoo finance. The daily data for five years from July 10, 2017, to July 10, 2022, including the High, Low, Open, Close, Adj Close, and Volume values for each day. In addition, two new columns are created in the table which are the data of percentage change of stock high and low price and percentage change of stock closing price and opening price and set them as HL_PCT and PCT_change. The formula for the HL_PCT variable is $(High - Low)/Adj\ Close \times 100\%$, and the formula for the PCT_change is $(Adj\ Close - Open)/Open \times 100\%$. Both of them and Adj Close and Volume are independent variables, and Adj Close is a dependent variable.

2.2 Models and metrics

The linear regression model is often used for predictive analysis. Its formula is

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon \tag{1}$$

Here, $\beta_0 - \beta_p$ are coefficients of the features. It mainly explores the linear relationship between a dependent variable and one or more independent variables and uses it to predict future outcomes [5]. For other fields, it can also be used to predict the effect of diet on body weight and others. In the research process, exploratory data analysis (EDA) was first used to find out if there are any missing values in the data, and select the number of days to be predicted by time shifting. On account of the missing data due to time shifts, this part of the missing values will become the data in the new "label" column. Afterwards, a heat map will be used to explore the correlation between the values to determine the variables and subsequently eliminate the irrelevant values. When the required values were determined, the model was created and the data was divided into X and Y, the training set and the test set. Concerning the accuracy of the prediction results, the score of this linear regression model was checked. To further evaluate the model, the mean absolute error and R-squared values were examined, and the specific data will be discussed later. MAE is a measure of errors in a set of predictions, and in regression analysis, R-squared measures strength of the relationship between linear model and the dependent variables on a 0 to 1 scale. Here are the two formulae:

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} = 1 - \frac{\Sigma\left(\hat{y} - y^i\right)^2}{\Sigma(\bar{y} - y^i)^2} \tag{2}$$

$$\mathrm{MAE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} |\mathrm{x_i} - \mathrm{x}| \tag{3}$$

3. Results & Discussion

3.1 Correlation analysis

Performing correlation analysis is to find causal relationships. To better understand, an example is an experiment done by economists Emma Harrington and Natalia Emanuel in 2019. In Ref. [6], the authors Dean Karlan and Michael Luca mention that after the two economists investigated the wages in the warehouse of a technology company, the company increased pay for warehouse workers from $16 an hour to $18 an hour. The study found that higher wages not only increased productivity but also significantly reduced turnover. To be specific, a $1 increase reduced the chances an employee would quit by 19% [6].

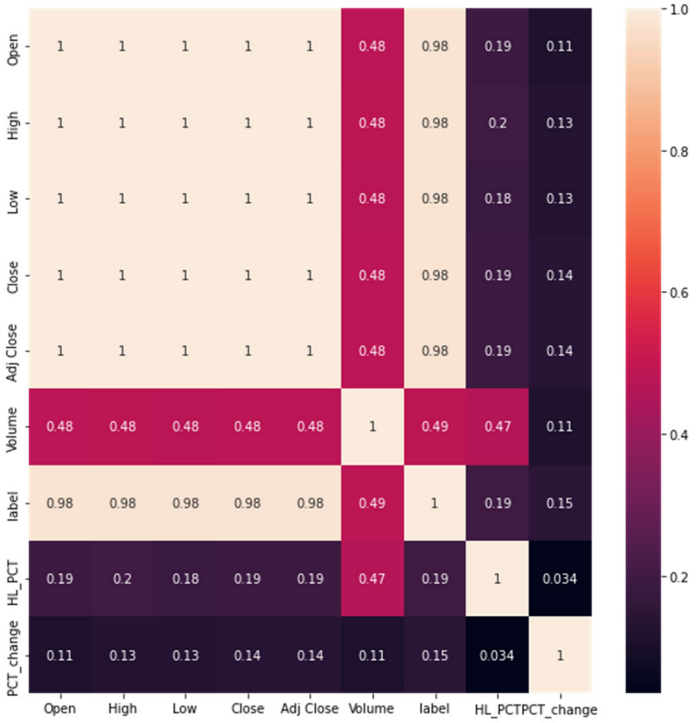


Fig. 1. Correlation analysis of POAHY.

As mentioned before, the new "label" column is created to input future values, the forecast values. As shown in Figure 1, it is mainly to observe the magnitude of the values intersecting with the label column through the heat map. The closer the values are to one, the stronger the correlation and the lighter the color. The correlation between Volume and label is 0.48, which is not high, so it can be judged as irrelevant and discarded in the subsequent process. The closing price is the raw price, and the adjusted closing price considered other factors such as dividends. Since the correlation between Adj Close and label is 0.94, which is greater than the correlation between Close and label (0.92). Therefore, Adj Close is preferred as the independent variable. The rest of the Open, High and Low rows intersect with the label with a value of 0.92, which is close to 1, so it is determined to be positively correlated.

3.2 Regression analysis

The accuracy of the model is important because of the requirement for the accuracy of the predicted values. Cort and Kenji compared the relative ability of RMSE and MAE to describe the model error in “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance”. It was found that RMSE varies with the variability within the distribution of error magnitudes and with the square root of the number of errors, as well as with the average-error magnitude (MAE) [7]. Therefore, the authors argued that MAE can be a more natural and explicit measure of average error and considered dimensioned evaluations and inter-comparisons of average model-performance error, therefore, should be based on MAE [7]. Hence, this study

mainly calculates the MAE and R-squared values to determine the accuracy of the model. The model was built based on the data of stock BYDDY, and the obtained MAE is 2.7. By calculating the data for 2022, the average adjusted closing price is about 75, so the error control of 2.7 indicates that the error of the model is very low. However, since the value of MAE has no limitation and is not taken between 0 and 1, which is not as straightforward as R-squared, the value of R-squared is also calculated.

Table 1. MAE and R-squared of Three Stocks.

	F	POAHY	BYDDY
MAE	0.733	0.391	2.703
R-squared	0.925	0.894	0.966

The closer the value of R-squared is to 1, the smaller the denominator is, and the lower the error rate is. According to the formula, the denominator is the sum of the squares of the difference between the mean and the true value, which is the error generated by the model "predicted value = sample mean" (Baseline Model) [8]. The R-squared value of the model was found to be about 0.97, which indicates that the accuracy of the model is very high and the error between the predicted and true values is very small. Based on the model, the trends and stock price ranges of the three stocks were successfully calculated. The preditced results and the price trends of the three companies are given in Figs. 2-4.

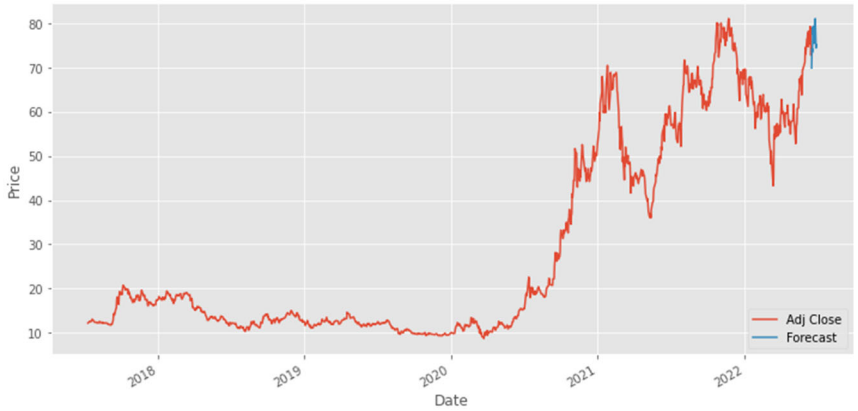


Fig. 2 Stock price of BYDDY.



Fig. 3 Stock price of F.



Fig. 4 Stock price of POAHY

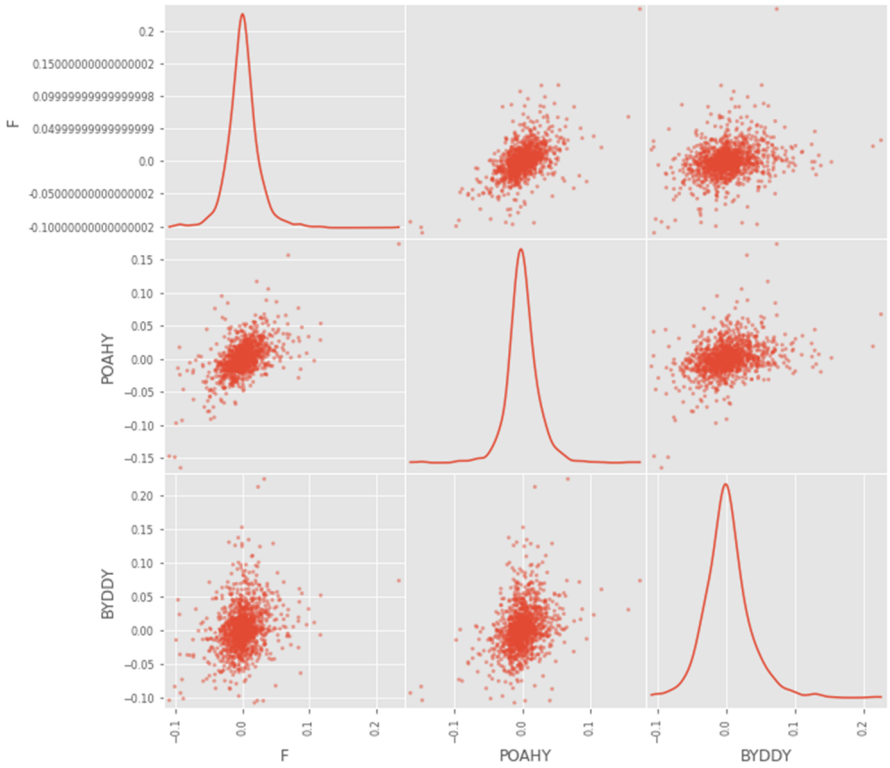


Fig. 5 The scatter diagram for correlation of three stocks

All three stocks belong to the automotive industry and operate in multinational businesses, so in order to know the relationship between the three stocks, this project extracts five years of adjusted closing price data for the three stocks individually and makes them into a new table for correlation analysis. A scatter plot of them is presented in Fig. 5 for better observation. Seen from the results, the distribution among the three stocks is approximately positively correlated. To demonstrate the positive correlation, a heat map is chosen to further visualize the degree of correlation between competing stocks. Besides, the distribution among the three stocks is approximately positively correlated. To demonstrate the positive correlation, a heat map is chosen to further visualize the degree of correlation between competing stocks. The heatmap chart illustrated in Fig. 6 shows that there is a relatively high correlation between some stocks such as stock F and stock POAHY, but some have a low correlation such as stock BYDDY and stock F.

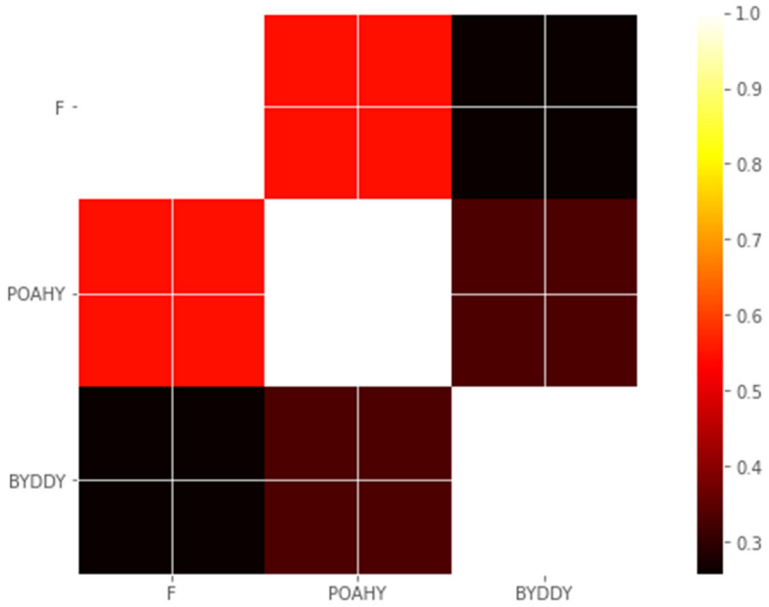


Fig. 6 The heatmap for correlation of three stocks.

3.3 Limitation

Although the study found that the three stocks are correlated, this does not indicate a causal relationship among them. It can only show the trend of the auto industry and not how competing stocks affect each other. This forecast utilizes a linear regression model, but it has some limitations owing to the characteristics of linear regression. For example, it only considers the mean of the dependent variable. Since the mean value cannot give a complete description of a variable, linear regression is not a complete description of relationships among variables [9]. In addition, linear regression models are very sensitive to outliers, so if there is an outlier in the stock it can have a huge impact on this model. In addition, MAE and R-squared can determine the accuracy of the model, but they still have shortcomings. To be specific, MAE lacks uniformity in predicting different things. Although different models can reduce the error of forecasts to some extent, they are more for short-term predictions. Long-term stock forecasting is still a big challenge. Similar to weather forecasting. The “butterfly effect”proposed by Lorenz in 1972 can figuratively describe this dilemma [10]. When there is a small change in any factor that has an impact on the stock market, it can lead to an unpredictable change in the stock price two weeks later. Neural networks, instead, can recognize nonlinear relationships. LSTM can also capture the problems of long-term dependencies, e.g., it can connect previous data to the present task.

4. Summary

In conclusion, this paper investigates the way to predict stock price in terms of linear regression. Based on the linear regression model for stock trend prediction, not only the model was successfully built, but also the trend and stock price range of three stocks were predicted. By analyzing the correlation between the variables, it was found that volume was not related to the predicted value. In the process, the MAE and R-squared were calculated, which show that the accuracy of the model in predicting all three stocks was above 90%. Moreover, through observing the correlation among three stocks, it was found that they also have a positive correlation with each other, although this does not indicate a causal relationship. On the other hand, the linear regression model has some limitations. Specifically, if there are outliers in the historical price of a stock, the regression will be greatly affected and the results will be less accurate. In contrast, the widespread use of linear regression models provides a simpler, more convenient and effective way to apply them to the stock market. Not only does it ensure the accuracy of the prediction, but its algorithm is also easier to understand. In the future, linear regression models can be widely applied in various fields and combined with other state-

of-art scenarios (e.g., machine leaning with neural network). Overall, these results offer a guideline for carrying out prediction in terms of linear regression models.

References

[1] Jordan Bishop, From GameStop To Tesla – How social media is driving the stock market, Maddyness, December 2, 2021, Retreived from: https://www.maddyness.com/uk/2021/12/02/from-gamestop-to-tesla-how-social-media-is-driving-the-stock-market/#:~:text=Social%20media%20has%20a%20huge,effect%20on%20all%20parties%20involved.

[2] Alexx Kay, Artificial Nural Networks", COMPUTERWORLD, Feb 12, 2001, Retreived from: https://www.computerworld.com/article/2591759/artificial-neural-networks.html#:~:text=One%20answer%20is%20to%20use,and%20learned%20to%20recognize%20objects.

[3] Jingyi Shen and M. Omair Shafiq, Short-term stock market price trend prediction using a comprehensive deep learning system, J Big Data 7, 66 (2020).

[4] Xiaohua Wang, P. K. H. Phua and Weidong Lin, Stock market prediction using neural networks: Does trading volume help in short-term prediction?, Proceedings of the International Joint Conference on Neural Networks, vol.4, 2003, pp. 2438-2442

[5] "Linear Regression." IBM, Retreived from: https://www.ibm.com/docs/en/db2-warehouse?topic=procedures-linear-regression.

[6] Dean Karlan and Michael Luca, "How to Use Correlation to Make Predictions", Harvard Business Review, March 6, 2022, Retreived from: https://hbr.org/2022/04/how-to-use-correlation-to-make-predictions

[7] Willmott, Cort J., and Kenji Matsuura. Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. Climate Research, vol. 30, no. 1, 2005, pp. 79–82. JSTOR.

[8] "2.3 The Mother of the Model: Study Notes on the Evaluation Index of Linear Regression." Birost, Retrieved from: https://blog.birost.com/a?ID=01650-72b89284-3447-459c-968b-377d8d2967a2.

[9] Peter Flom, "The Disadvantages of Linear Regression", SCIENCING, March 13, 2018, Retreived from: https://sciencing.com/advantages-using-independent-group-ttest-8647277.html

[10] Alan Chodos, "This Month in Physics History." American Physical Society, Retreived from: https://www.aps.org/publications/apsnews/200301/history.cfm#:~:text=Lorenz%20subsequently%20dubbed%20his%20discovery,range%20weather%20forecasting%20was%20doomed.