Proceedings of the Jepson Undergraduate Conference on International Economics

Volume 4 Article 4

7-2022

Stock Market Analysis Using Linear Regression

Taran Rishi
University of Northern Iowa

Stock Market Analysis Using Linear Regression

Taran Rishi

Department of Economics

University of Northern Iowa

Abstract:

This paper aims to identify the factors affecting the closing price of a stock in the S&P 500. It uses time series data for all companies that constitute the S&P 500 index. Data on closing price, opening price, highest price, lowest price, and volume of each day is used. It turns out that the opening price, highest price, and the lowest price are the most significant variables while predicting the closing price of a stock.

Acknowledgements:

Firstly, I would like to thank my advisor, Dr. Imam Alam, for providing me with valuable information on various concepts of econometrics as well as assisting me in writing this paper. Furthermore, I would like to thank Dr. Shahina Amin for her suggestions towards improving this paper. Lastly, I would also like to thank my family for always supporting me and being there whenever I needed them the most.

Stock Market Analysis Using Linear Regression

Introduction:

Throughout history, it has been observed that the stock market is a great indicator of the economy. A continued decline in stock market often results in a recession. On the other hand, a strong stock market often results in a booming economy. Stock market can also be used to determine a country's financial power. Furthermore, the stock market can be used as a great source of income and generate wealth. But it is crucial to invest in the right stocks at the right time. Mastering this technique requires a high level of knowledge.

The process of determining future stock price for a company is called stock market forecasting. It can be done by using some data mining techniques. Throughout history we have collected so much data. We can use the technology available to us to organize this data and draw conclusions.

Data mining is a method of using data to find meaningful patterns and/or relationship among variables. Stocks are very volatile. There are so many factors affecting the price of a stock such as economic growth, inflation etc. It is impossible to consider all these factors while trading stocks. Thus, we use data mining to figure out which factors are important.

Literature Review:

Logistic Regression (LR)applications are used in banking, corporate finance, investments and other areas. It uses numerous variables to predict the outcome. Various researchers use Multivariate discriminant analysis (MDA) but Öğüt and Aktas (2009) found that data mining methods are a better option while predicting future stock movements realizing the best stocks to invest in for maximum returns. He also found a new method of depicting corporate failure. He encouraged to use empirical analysis.

There have been a lot of studies on stock market analysis. Due to their high volatility, it becomes difficult to predict the direction of movement of various stocks. Traders often capture sudden movements to make big profits as compared to long term investors who look at the performance of a company.

Data and Method:

Data is collected from the website Kaggle.com. It shows the stock prices for different companies in the S&P 500 from 2013 to 2018. The data consists of seven variables with 619,040 observations. The table below describes the different variables and shows their units as well as the direction of their predicted effects on the dependent variable (closing price).

Table 1: Variables

Variable Name	Descriptions	Units	Predicted sign of coefficient
Close	Closing price of stock	$	Dependent variable
Date	Date	Year-Month-Date	None
Open	Opening price of stock on that day	$	+
High	Highest price of stock on that day	$	+
Low	Lowest price of stock on that day	$	-
Volume	Number of shares of stocks traded within that day	Number of shares per day	+/-
Name	Name of the stock	None	None

The first variable is the date. It is recorded in the format Year-Month-Date It tells us the date for which the opening price, closing price, highest price, lowest price, and volume were recorded. The regression is based on time series. The data is collected daily.

The second variable is close. It tells us about the closing price of the stock at a particular date. It is measured in dollars. The closing price is the value of a share of stock at the end of the day, when stock exchange closes. It is important to estimate returns on investment. The closing price is measured in dollars. Close is the dependent variable in the regression. In other words, we will estimate the value of closing price of the stock using opening price, highest price, lowest price, and volume. The closing price is important for day traders and swing traders since they make profits by capturing changes in stock price over a short period of time. The closing price also helps them understand the market and how people feel about the stocks.

The third variable is open. It tells us about the opening price of the stock at a particular day. It is measured in dollars. Opening price is the value of a share of stock when the stock exchange opens for trade. It does not have to be same as the closing price of the previous day. Opening price is good predictor for the direction of movement of the stock on that day. It is one of the independent variables that is used to determine the closing price of that stock.

The fourth variable is high. It tells us about the highest price of a share of stock on a particular day. It is measured in dollars. The highest price is the maximum price for which a share of stock is traded on that day. It helps us determine the general movement of the stock.

The fifth variable is low. It tells us about the lowest price of a share of stock on a particular day. It is measured in dollars., The lowest price is the minimum price for which the stock is traded on that day. It is also useful in determining the general movement of the stock.

The sixth variable is volume. It tells us about the volume of the stock for a particular day. It is measured in number of shares per day. Volume means “the number of shares or contracts

traded in a security or an entire market during a given period of time." It is a crucial concept and proves highly beneficial when trading stocks.

The seventh variable is name. It gives us the name of the stock in S&P 500 for which the data has been collected.

Results:

First a summary of the data was created. The following table was observed. It shows the number of observations, mean standard deviation, maximum, and minimum values for the different variables used in this regression project.

Table 2: Summary of variables

Variable	Observation	Mean	Standard deviation	Min	Max
Date	0				
Close	619040	83.0437	97.3898	1.59	2049
Open	619029	83.0233	97.3787	1.62	2044
High	619032	83.7783	98.2075	1.69	2067.99
low	619032	82.2561	96.5074	1.5	2035.11
Volume	619040	4321823	8693610	0	6.18e+08
Name	0				

Then a linear regression was run using close as the dependent variable and open, high, low, and volume as the independent variables.

Regression equation:

$\text{Close} = \beta_0 + \beta_1 \text{open} + \beta_2 \text{high} + \beta_3 \text{low} + \beta_4 \text{volume} + E$

Where β_0 , β_1 , β_2 , β_3 , β_4 are coefficients of open, high, low, and volume

Table 3: regression table

Variable	Coefficient	Std. Error	t-value	p-value	Sig variable
Open	-0.555523	0.0009759	-569.23	0.000	***
High	7.889902	0.0008396	939.72	0.000	***
Low	0.7667548	0.0007855	976.19	0.000	***
Volume	-4.58×10^{-11}	1.02×10^{-10}	-0.45	0.626	
cons	-0.0058165	0.0012737	-4.57	0.242	***

The coefficient of open is -0.555523. This implies that with 1 dollar increase in the opening price of the stock, the closing price decreases by 0.555523 dollars provided other variables are held constant. The p-value is 0.000. Thus, we can reject the null hypothesis at 1%, 5%, and 10% levels of significance. This implies that the variable open is significant at any level of significance.

The coefficient of high is 0.7889902. This implies that with 1 dollar increase in the highest price of the stock, the closing price increases by 0.7889902 dollars, provided all other variables are held constant. The p-value is 0.000. Thus, we can reject the null hypothesis at 1%, 5%, and 10% levels of significance. This implies that the variable high is significant at any level of significance.

The coefficient of low is 0.7667548. This implies that with 1 dollar increase in the lowest price of the stock, the closing price increases by 0.7667548 dollars, provided all other variables are held constant. The p-value is 0.000. Thus, we can reject the null hypothesis at 1%, 5%, and 10% levels of significance. This implies that the variable low is significant at any level of significance.

The coefficient of volume is -4.58×10^{-11}. This implies that with each extra share of the stock traded that day, the closing price decreases by 4.58×10^{-11} dollars, provided all other variables are held constant. The p-value is 0.626. Thus, we can not reject the null hypothesis at 1%, 5%, and 10% levels of significance. This implies that the variable volume is not significant at these three levels of significance.

The y-intercept is -0.0058165. This implies that the value of closing price of a share of stock is -0.0058165 dollars when open, high, low, and volume are equal to 0. The p-value is 0.242. Thus, we cannot reject the null hypothesis at 1%, 5%, and 10% levels of significance. This implies that the y-intercept is not significant at any of the above levels of significance.

After obtaining these results, the test for multicollinearity was conducted since the results cannot be trusted if the independent variables are correlated to one another. To check for multicollinearity Variance Inflation Factor (VIF) test was conducted.

$VIF=1/(1-R^2)$

If VIF > 10, then at least one of the variables would be collinear to another.

If VIF < 10, then there would be no collinearity present.

Table 4: VIF table

Variables	VIF Score
Open	11794.72
High	8879.15
Low	7504.23
volume	1.03

VIF scores for open, high, and low were greater than 10. This implied that multicollinearity was present. This multicollinearity was solved by removing high and low variables. We kept open because it was most consistent with the theory. Following results were obtained upon running regression for the reduced model.

Linear regression Number of obs = 619,029

R-squared = 0.9997

Root MSE = 1.5585

Table 5: regression table (reduced model) robust

Variables	Coefficient	Standard Error	T value	p-value	95% confidence interval
Open	0.9999	0.0001	6807.46	0.0000	(0.9997,1.0002)
Volume	-1.32×10^{-9}	2.49×10^{-10}	5.29	0.0000	$(-1.81\times10^{-9}, (-8.30\times10^{-10})$
y- intercept	0.0290	0.0118	2.46	0.014	(0.0059,0.5217)

For this reduced model, we obtained the following VIF.

Table 6: VIF table (reduced model)

Variable	VIF
Open	1.02
Volume	1.02

To check for heteroskedasticity white's test was run. In order to perform white's test, the following steps were taken. First, the model was estimated.

$$\text{Close} = \beta_0 + \beta_1 \text{open} + \beta_2 \text{high} + \beta_3 \text{low} + \beta_4 \text{volume} + E$$

Then residuals were calculated as the difference between the observed value and the estimated value of the dependent variable. Then, the residuals were squared. Then a regression was estimated with e^2 as the dependent variable and the independent variables remained the same. As a result, we would obtain the p-value. If p-value is less than the level of significance, then heteroskedasticity is present, otherwise it is not. After performing white's test, we observed a p-value of 0. Thus, we rejected the null hypothesis. This implied that heteroskedasticity was present. Heteroskedasticity was removed by doing the robust test.

Conclusion:

The R^2 value was 1.0000. This indicated that the entire variation in the closing price of a share of stock could be explained by the variation in the opening price, highest price, lowest price, and volume of the stock. Thus open, high, low, and volume are important variables in predicting the closing price. Open, high, and low were statistically significant variables whereas volume was not statistically significant in this model. To remove multicollinearity, we removed high and low variables. The reduced model had R^2 value of 0.9997. This implied that 99.97% of the variation of the closing price of a share of stock could be explained by the variation in the opening price and volume of that stock. Both open and volume were statistically significant variables in this model.

References:

(PDF) analysis of stock market predictor variables using linear regression. ResearchGate. (n.d.). Retrieved December 5, 2021, from https://www.researchgate.net/publication/326253896_Analysis_of_stock_market_predictor_variables_using_linear_regression.

Current volume: Volume 22 (2020). Major Themes in Economics | CBA Journals | University of Northern Iowa. (n.d.). Retrieved December 5, 2021, from https://scholarworks.uni.edu/mtie/?utm_source=scholarworks.uni.edu%2Fmtie%2Fvol17%2Fiss1%2F5&utm_medium=PDF&utm_campaign=PDFCoverPages.

Enke, D., Grauer, M., & Mehdiyev, N. (2011, October 11). *Stock market prediction with multiple regression, fuzzy type-2 clustering and neural networks*. Procedia Computer Science. Retrieved December 5, 2021, from https://www.sciencedirect.com/science/article/pii/S1877050911005035.

Nugent, C. (2018, February 10). *S&P 500 stock data*. Kaggle. Retrieved December 5, 2021, from https://www.kaggle.com/camnugent/sandp500.

Öğüt, H. (2015, February 18). *Detecting stock-price manipulation in an emerging market: The case of Turkey*. Expert Systems with Applications. Retrieved July 11, 2022, from https://www.academia.edu/10897867/Detecting_stock_price_manipulation_in_an_emerging_market_The_case_of_Turkey

Stock price prediction using regression analysis. (n.d.). Retrieved December 5, 2021, from https://www.ijser.org/researchpaper/Stock-Price-Prediction-Using-Regression-Analysis.pdf.