Analysis of the complexity of a dataset on the correlation coefficient in a Linear Regression using Stock Figures

Abstract:
Linear Regression is used as a type of predictive analysis and it is frequently used in finance. It is, in its simplest form, a method to make data into a mathematical equation but, the accuracy of it varies. The correlation coefficient, founded by Sir Francis Galton and modified by Karl Pearson, is a measure which defines the statistical relationship between two variables. In Finance, Linear Regression is used to extrapolate data and predict future outcomes of stocks and the values of companies. In order to test the correlation coefficient, two different data sets were taken in order to see, how the complexity of a dataset will affect the correlation coefficient in Linear Regression. We concluded that the correlation coefficient is affected by the complexity of a dataset.

Keywords: Linear Regression, Correlation Coefficient, Penny Stocks

---

Introduction:
The stock market is highly dynamic and can be difficult to predict without proper analysis. In recent years global events, such as the COVID-19 pandemic, have highlighted the unpredictability of the market, as demonstrated by the sharp drop in the Dow Jones index in March 2020. However, by utilizing linear regression, we can identify patterns and trends in stock prices, which can further improve the accuracy of predictions and enhance our understanding of the market.

The correlation coefficient is a key metric for evaluating the effectiveness of linear regression in stock market analysis. It measures the strength and direction of the relationship between two variables and can indicate whether there is a strong positive or negative correlation between stock prices and opposing factors.

These patterns can provide valuable insights into the behaviour of the stock market and can help investors make more informed decisions.

Linear regression is a powerful tool for stock market analysis that can help to improve the accuracy of predictions and enhance our understanding of the market. By utilizing linear regression, we can identify patterns and trends in stock prices, and use the correlation coefficient to evaluate the effectiveness of our analysis. These techniques can help investors to make more informed decisions and control their navigation of the stock market.

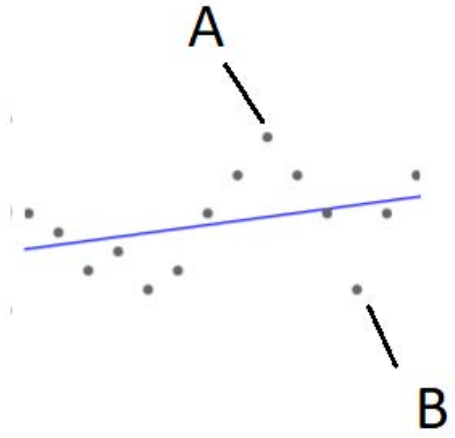


Figure 1

Points A and B are signified as it was necessary to show the way the linear regression was determined. The linear regression helps forecast future trends by going on to show the mean of all the X and Y points and then further extending it in order to extrapolate it. In this scenario, we can go on to conclude that any point below this linear regression will be a good buying point and any point that goes above will be a good selling point. This, overall, is a strong linear regression with quite a strong correlation coefficient (in relation to trading charts)

Literature Review:

Many datasets can be confined into one relation and the way to check the strength of it is by using the correlation coefficient. According to the Journal of Targeting, Measuring and Analysis for Marketing:

"0 indicates no linear relationship.

+1 indicates a perfect positive linear relationship – as one variable increases in its values, the other variable also increases in its values through an exact linear rule.

−1 indicates a perfect negative linear relationship – as one variable increases in its values, the other variable decreases in its values through an exact linear rule.

Values between 0 and 0.3 (0 and −0.3) indicate a weak positive (negative) linear relationship through a shaky linear rule.

Values between 0.3 and 0.7 (0.3 and −0.7) indicate a moderate positive (negative) linear relationship through a fuzzy−firm linear rule.

Values between 0.7 and 1.0 (−0.7 and −1.0) indicate a strong positive (negative) linear relationship through a firm linear rule."

This is extremely important as it can be used for Marketing and even general analysis. The correlation coefficient helps analyze how strong the dataset is and, in finance, the majority of the work is to look for patterns and predict what is to happen with those patterns. The easiest way to analyze a pattern is to make a graph and with the coefficient being as accurate as possible, the analysis and predictions become far easier and less of a hassle. They are used to forecast future trends and to manage the risks within a portfolio (Investopedia). Further proving that it is used for the extrapolation of data in finance.

Methodology:

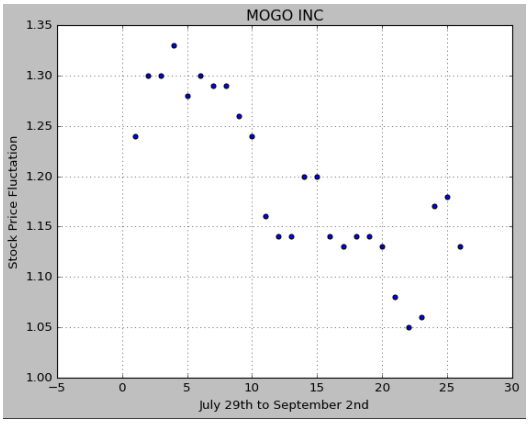
Penny Stocks will have a higher correlation coefficient due to the complexity of the dataset being less than that of a higher-valued stock such as Amazon.

Over the course of a month (July 29th - September 2nd, 2022), the stock prices of Amazon and MOGO Inc were analyzed. The data was collected from Yahoo Finance and we used the opening price of the stocks. The Python library Numpy was used in order to make the arrays of data necessary to calculate our Linear

Regression model. Using the Opening price, we used matplotlibs library to graph it out in order to have a visual example of how the graph would look before plotting the Linear Regression.

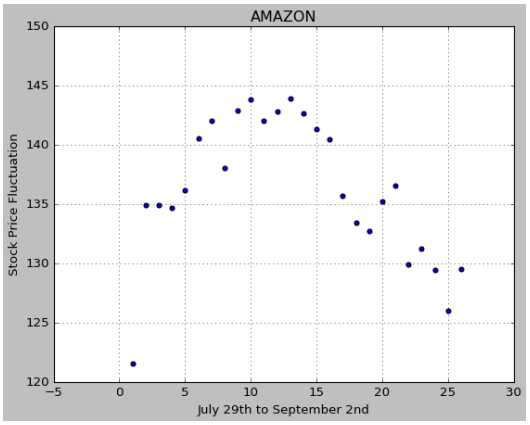
MOGO INC STOCK PRICES *(Figure 2)*

Date	Open	Date	Open
Sept 01, 2022	1.4700	Aug 16, 2022	1.6500
Aug 31, 2022	1.4800	Aug 15, 2022	1.6600
Aug 30, 2022	1.4400	Aug 12, 2022	1.6200
Aug 29, 2022	1.4700	Aug 11, 2022	1.7800
Aug 26, 2022	1.5700	Aug 10, 2022	1.6500
Aug 25, 2022	1.5600	Aug 09, 2022	1.6400
Aug 24, 2022	1.4800	Aug 08, 2022	1.5700
Aug 23, 2022	1.4800	Aug 05, 2022	1.4300
Aug 22, 2022	1.5200	Aug 04, 2022	1.4400
Aug 19, 2022	1.6200	Aug 03, 2022	1.4100
Aug 18, 2022	1.6600	Aug 02, 2022	1.3200
Aug 17, 2022	1.6400	Jul 29, 2022	1.2600

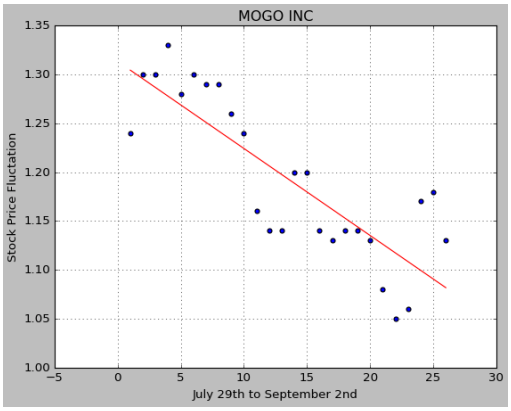


AMAZON STOCK PRICES *(Figure 3)*

Date	Open	Date	Open
Sept 01, 2022	126.00	Aug 16, 2022	143.91
Aug 31, 2022	129.45	Aug 15, 2022	142.80
Aug 30, 2022	131.25	Aug 12, 2022	142.05
Aug 29, 2022	129.90	Aug 11, 2022	143.86
Aug 26, 2022	136.55	Aug 10, 2022	142.90
Aug 25, 2022	135.26	Aug 09, 2022	138.05
Aug 24, 2022	132.75	Aug 08, 2022	142.05
Aug 23, 2022	133.41	Aug 05, 2022	140.10
Aug 22, 2022	135.72	Aug 04, 2022	140.58
Aug 19, 2022	140.47	Aug 03, 2022	136.21
Aug 18, 2022	141.32	Aug 02, 2022	134.72
Aug 17, 2022	142.69	Jul 29, 2022	134.90



Results:



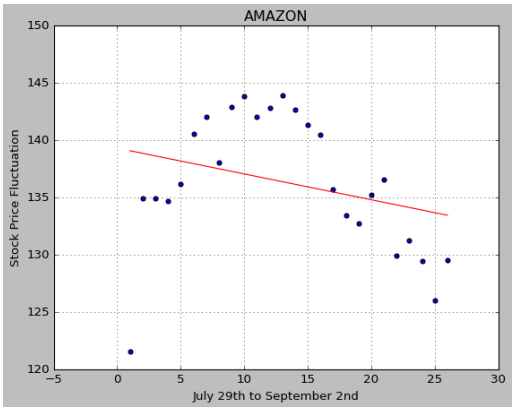


Figure 4

In Figure 4 we can clearly see that the correlation coefficient of MOGO Inc (penny stock) is a lot higher than the one of Amazon further proving that penny stocks have less fluctuation than higher-valued stocks.

Correlation Coefficient:

0.6923765403691774

Intercept:

1.3131692307692306

Slope:

-0.00889573

Figure 5

In Figure 5, just from being able to see it, it shows that the correlation coefficient would be lower due to the Linear Regression not being able to hit most of the points.

Correlation Coefficient:

0.08388745072136605

Intercept:

139.30113846153847

Slope:

-0.22507009

With this data, what actually happened next?





Figure 6
In Mogo Inc, we can see the pattern continues throughout the year. *The correlation coefficient for this dataset stays strong throughout the year at 0.86.* Regressions were actually right in the predictions over the span of 4 months. They both predicted a downward trend and they both were right.

Limits of Linear Regression:
With such great power comes even greater limits. Due to the fact that linear regression essentially just draws out a line of best fit, it is not viable for an untrained financial investor who does not understand marketing patterns. It tells the general pattern over time allowing for people who want to invest for long-term goals to have a good understanding but short-time it is not efficient.

Conclusion:
We were able to come to a final analysis saying that the correlation coefficient fluctuates due to the complexity of a dataset; less fluctuation means a higher correlation coefficient and higher fluctuation means a lower one. Further proving that penny stocks tend to be more stable and easily analyzable in comparison to higher-valued stocks over a shorter time. A higher correlation coefficient allows for a more accurate analysis and, obviously, with a more accurate analysis comes to a better investment opportunity.

Karabulut, Y. (2011). Can Facebook predict stock market activity? Available at http://bus.miami.edu/umbfc/_common/files/papers/Karabulut.pdf

M.Ray, Harswarup Sharma "A Reference book on Mathematical Statistics" 4TH Edition, 2006.

Data Sources from http://yahoofinance.com

Figure 7
For Amazon, the pattern also managed to stay stable throughout the year. *The correlation coefficient for this dataset is 0.88. (higher than Mogo Inc)*

Bibliography:
Ratner, B. The correlation coefficient: Its values range between +1/−1, or do they?. Journal of Targeting, Measuring and Analysis for Marketing 17, 139–142 (2009). https://doi.org/10.1057/jt.2009.5.

Hayes, Adam. "Correlation: What It Means in Finance and the Formula for Calculating It." Investopedia, Investopedia, 19 Dec. 2022, https://www.investopedia.com/terms/c/correlation.asp

Statistics, Solutions. "What Is Linear Regression?" What Is Linear Regression?, 10 Aug. 2021,

Montgomery, Douglas C., et al. "1.1 Regression and Model Building." Introduction to Linear Regression Analysis, Wiley, Chichester, England, 2013, pp. 2–6.

Bulkowski, Thomas N. Trading Classic Chart Patterns. pp 97-426 John Wiley, 2002

Ullrich, Roland. Trading Psychology. John Wiley & Sons, Inc., 2022.

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020).

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011

"Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp.

Liu, Qianqiu and Rhee, S. Ghon and Zhang, Liang and Zhang, Liang, On the Trading Profitability of Penny Stocks (August 26, 2011). 24th Australasian Finance and Banking Conference 2011 Paper, Available at SSRN: https://ssrn.com/abstract=1917300 or http://dx.doi.org/10.2139/ssrn.1917300

Singh, N. (2020, July 27). Advantages and disadvantages of linear regression. OpenGenus IQ: Computing Expertise & Legacy. Retrieved January 26, 2023, from https://iq.opengenus.org/advantages-and-disadvantages-of-linear-regression/