Risk Management with Stock Price Forecast Using Regression Modeling

Anna He

Aragon High School, USA

ABSTRACT

For short-term traders who trade with large amounts of capital within a short time frame, having the analytical tool to predict stock price is crucial for the risk management and planning of entry and exit. While retail traders rely on the visualizations of technical indicators to anticipate general price direction and trend lines to estimate areas of concentrated demand or supply, multiple regression modeling uses known quantifiable indices whose functions are based on historical prices of an equity to yield a concise price prediction. This study aims to build an accessible and accurate model for retail traders without advanced software. In order to do that, we must optimize the quality of forecast while minimizing the number of independent variables. Through multiple linear regression and forward stepwise regression, we isolated 3 key variables – the daily closing price, Nasdaq 10-day moving average, and the Dow intraday average – which, when combined again in multiple regression, would produce a model with a multiple-R of 0.9995 and a low standard error of 2.44556268.

Introduction

In the field of corporate finance, regression models find extensive application for predictive modeling. In order to create a regression equation, the independent variables and dependent variables must be identified in order to find the optimal slope, or coefficient associated with each variable. There are 64 sets of features in this experiment—some have strong correlation with the dependent variable-- stock closing price of the following trading day, and others do not: The objective of this paper is to identify the key variables which correlate the most highly with a security's daily closing price one trading day into the future, and to ultimately model a forecast equation using the shortlisted variables. On StatTool, we first performed multiple regression with the 64 independent features. Then, all variables with less than 99% statistical significance, or p-value of higher than 0.01, were omitted. Subsequently, forward stepwise regression was applied to the remaining variables. Quality-of-forecast tests involved withholding specific variables from regression, and the standard error dataset was examined to correlation. Lastly, forward selection was used to create a multivariable function with coefficients assigned to the five most significant variables. It is important to note that all variables in this experiment are derived based on the available inputs collected at the end of each trading day. As a result, the forecasted values are most applicable for short-term day-to-day risk evaluation.

Literature Review

The idea of a "regression to the mean" was first conceptualized by Sir Francis Galton [6] during an assessment of men's height, where he noted that the offspring of exceptionally tall and short men typically exhibited heights closer to the population average. Galton's exploration of statistical phenomena extended throughout his career as a biometrician. Through the collaboration of other mathematicians, the concept of regression slope evolved into mathematical formulas which to this day, is utilized in a wide range of academic disciplines, including the price forecasting in the equity market. In the 21st century, research endeavors have utilized regression techniques alongside algorithmic models such as decision trees to forecast market prices over extended periods.

A study conducted by Lucas Nunno [3] in 2014 delved into algorithmic modeling, revealing that linear regression outperforms polynomial regression in long-range forecasting. The shortcomings of polynomial regression forecasting occur in how employing multiple variables in regression leads to an overfit function. However, Nunno also observed that using optimal variable placement to eliminate variables with low correlation yields a line of best fit with enhanced forecast precisions which, at the same time, prevents overfitting of the forecasted function. His conclusion pointed to a combination of forward stepwise regression and multiple regression to derive an effective equation that balances complexity with predictive accuracy.

Data Collection

The historical data from Apple, the company which holds the greatest weight in both the Dow Jones (Dow), Nasdaq, and Strong & Poor (S&P) 500 index, will be used as the parameters for the regression model examined in this paper. Since the establishment of electronic trading platforms, the magnitude of hourly price fluctuation of equities has increased; and according to behavioral economist Richard Thaler [4], the increased volatility can be largely attributed to the overconfidence of investors. To minimize these distractions for regression modeling, this paper uses 3732 days of prices collected daily from October 2005 to August 2020. The dataset, retrieved from Kaggle.com, consists of 60 independent variables– a variety of matrices ranging from moving averages to momentum indicators. The dependent variable is the stock closing price of the following day (d+1), which is the price the model aims to yield. The projected price is not solely based on the historical data of Apple's stock itself. Rather, the regression function includes indices from the Invesco QQQ and Dow Jones exchange-traded funds (ETF).

Method/Model

The price movement of an equity is driven by supply and demand, which refers to the selling and buying interest by the market. The concept of stock prices retracing to a trend line or linear slope arises from the inherent need for price consolidation and correction as an equity's price cannot sustain continuous ascent. In an uptrend, for instance, short term traders take profits as the price of an equity increases, creating selling pressure. When supply exceeds demand, sellers would have to ask at lower prices where the buyers place their risk averse bids. The downbidding creates a downward movement towards a certain price level or a linear trend line, otherwise known as the support. Once the stock retraces to its linear trend line which most traders visualize through charting softwares, the retracement prompts renewed buying interest, propelling the price once more on its upward trajectory. These temporary reversals are evident for simple linear regression (y=m*x+b) to be a reliable forecasting tool.

However, in the simple linear regression y=m*x+b–where y is the dependent variable and x is the independent variable– the function has a constant rate of change as it is based on only one single degree variable. Thus, this method presents limitations: for the past 10 years of United States stock market history, the tangent slope, or the instantaneous change of rate of price, has been increasing. Similarly, the nominal value of the $AAPL stock price also saw an increasing rate of growth. Therefore, the application of multiple regression would be suitable for this forecasting as the y-value is based on when the derivative with respect to time is proportional to the everchanging independent variables. The null hypothesis states that the independent variables have no correlation with the dependent variable.

First, multiple regression is used to reject the null hypothesis. In order to obtain an accurate line of best fit, independent variables with p-value higher than 0.01, which denote less than 99% significance, are first omitted. The choice to use 0.01 instead of 0.05 as the threshold stems from the model's primary objective: to maximize accuracy for daily price targets using no more than 5 variables. Then, in order to distinguish the variables that contribute the highest statistical similarity with the target function, variables with a p-value of <0.001, which indicates that the difference between each variable and the predicted close will be considered statistically significant, are plugged into a stepwise regression. Through forward selection, the resulting model will include only those variables that contribute significantly to explaining the variation in the dependent variable. The finding is the closing price, 50-day moving average, and the indicator values of other indices that are closely correlated with the stock. The model also includes two variables that represent the broad market movement – the $QQQ 10-day moving average and the Dow Jones intraday average [(t-1)-(t-5)], since it was observed that $AAPL moves in a magnitude proportional to that of the DOW Jones and Nasdaq. The coefficient is calculated by the formula as follows: $\beta(Y, X) = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}$ Lastly, the standard error, which measures the potential deviation from the predicted function with respect to sample size, helps assess the uncertainty of the variable output. The lower the standard error implies a lower margin of error for day traders.

The closing price (t) of a stock is measured in dollars as the last price at which a stock security is traded when the exchange closes. As a direct reflection of market sentiment heading into the next trading day, the closing price is deemed as important in technical analysis. For instance, short term traders make their trade based on where the candlestick closing price sits relative to the moving averages and trend lines. A close below the uptrend trend line support triggers selling across the market. A close on/above support during a downward price movement would trigger dip buying.

The variable S_Close(t-1), which signifies the price exactly one hour before the closing bell, also exhibited a p-value of less than 0.001. However, it was observed during the stepwise regression stage that (t) had a standard error of 0.161 while (t-1) had a standard error of 0.168. The standard error signifies the average distance the observed value falls from the regression line; thus a similar standard error implies high level of similarity between the two variables. To further investigate, the correlation matrix was referenced to visualize the correlation relationship between variables in the regression model. It turned out that close(t) and (t-1) had a correlation of 1, which denotes perfect correlation. Since multicollinearity is occurring between the two variables, the removal of one of the variables is necessary to optimize the model's efficiency.

Multicollinearity Checking

VIF	R-Square
42.88325647	0.976680875
32.16018072	0.968905647
119.5386374	0.991634504

Dow Jones Daily Intraday Mean is obtained through feature engineering. In this experiment, it was derived from the average of $DOW closing prices exactly 1 hour (t-1) and exactly 5 hours (t-5) before the closing bell. This variable is created to minimize the effects of hourly fluctuations in the stock price, essentially taking the average value of two of the highest volume trading times in a day. The Dow Jones Industrial Average is an exchange-traded fund (ETF) containing 30 of the largest publicly traded US companies. These companies span multiple sectors, including Technology, Finance, Consumer Services, and Real Estate, and are collectively known as the blue chip stocks. As they represent the bedrock of the American economy, their performance closely mirrors the macroeconomic state of America. During economic downturns, the $DJIA companies exhibit greater resilience to recessionary pressures. According to historical data, Apple Inc. ($AAPL) consistently outperforms the Dow Jones Industrials Index in periods of economic expansion. However, during economic recessions, $AAPL tends to underperform. Despite that the Dow Jones only has a correlation of 0.918 with the target $AAPL closing price, keeping the variable in the model actually decreases the standard error of the resulting model.

The Nasdaq 100 10-day moving average takes into account the closing price over the 50-day time period and divides it by the number of days. $AAPL holds 11.418%-- the most weight in terms of market capitalization in the Nasdaq 100 ETF. The inclusion of the 10-day moving average of the $QQQ derived from past research has suggested that during a market correction, the top five companies in market cap FAANG typically start collapsing after small cap stocks. This phenomenon is notably exemplified in the case of $AAPL compared to the Nasdaq Index. During an economic expansion, small cap stocks observe substantial growth due to increased borrowing and heightened consumer confidence, meaning more retail traders have their money invested into small to mid cap stocks for the highest possible return in speculative trades. Under economic conditions where businesses experience liquidity contractions, the risk-averse institutions tend to let the fund reside in large-cap tech stocks or ETFs to minimize risk. This implication can be supported by data from Yahoo Finance: Between August 1st and October 1st of 2023, the Russell 2000 ETF, which consists of 2000 small-cap U.S companies, has fallen around 15%. On the other hand, the Dow Jones ETF, fell around 7%. Thus, when the broad market ($QQQ) does correct, Apple would be expected to have a lagged correction in the case that the large amounts of financial capital have yet to be pulled out to cash form by financial institutions. Similar to $AAPL, the $QQQ 10-day moving average would also lag because it takes the mean of the past 10 days of trading (all up) and will only flatten out/change direction if the $QQQ continues to fall.

Results- Stepwise Regression

Equation: Close forecast = 0.97559516 + 0.99388407 Close(t) - 0.00017797 Dow intraday mean + 0.02895838 QQQ_MA10

Stepwise Regression for Close_forecast Summary	**Multiple R**	**R-Square**	**Adjusted R-square**	**Std. Err. of Estimate**
	0.9995	0.9991	0.9991	2.447571783

Regression Table	**Coefficient**	**Standard Error**	**t-Value**	**p-Value**
Constant	0.975595162	0.256792769	3.799153555	0.0001
Close(t)	0.993884066	0.003313582	299.9424761	< 0.0001
Dow intraday mean	-0.00017797	4.06129E-05	-4.382103159	< 0.0001
QQQ_MA10	0.028958378	0.007837136	3.695020708	0.0002

Correlation Matrix	**Close forecast**	**Close(t)**	**Dow intraday mean**	**QQQ_MA10**
Close forecast	1.000	1.000	0.918	0.978
Close(t)	1.000	1.000	0.919	0.979
Dow intraday mean	0.918	0.919	1.000	0.972
QQQ_MA10	0.978	0.979	0.972	1.000

Following multiple rounds of stepwise regression, the top five independent variables are close, close(t-1), 50-day moving average, Dow intraday mean, and QQQ 10-day moving average. The regression model yields multiple statistical measures, including the multiple R, R-Square, Adjusted R-Square, and Standard Error. The R square ranges from 0 to 1 and is expressed in percentage. An R square value of 1 suggests a perfect correlation of movement between the predicted value and the target value. From a statistical standpoint, a model with a R-Square value of more than 0.90 would be considered strongly correlated. The objective of this experiment is to obtain a multiple regression equation with the highest R square possible, where the forecasted stock price is most in sync with the targeted predictive value. The final model resulted in forecasted values with an adjusted R square is 0.9991, which implies that 99.91% of the variation in Apple's future stock price positively corresponds to the forecasted price based on its closing price, 50-day simple moving average, the Dow Jones intraday mean, and the QQQ 10-day moving average.

The closing price has a coefficient of 0.993884006, which signifies that a $1 dollar increase in the closing price (t) corresponds to an increase of $0.993884006 in the closing price the next day (d+1). Additionally, the R-square of 0.999 indicates a near-perfect correlation between the closing price and the closing price (t) of the following day (d+1). The instances where they are negatively correlated are when the closing price reverses in direction. With the R-square value of 0.976, it can be implied that 97.6% of the variation in Dow Jones corresponds to the variation in (d+1). Lastly, the P-value of less than 0.0001 is lower than 0.01. Thus, the null hypothesis can be rejected at the 1% level of significance.

The coefficient of Dow Jones intraday mean is -0.00017797, which implies that every time the Dow Jones intraday average increases by $1, the closing price for $AAPL the next day decreases by $0.00017797. With the R-square value of 0.968905647, it can be implied that 96.9% of the variation in Dow Jones corresponds to the variation in (d+1). Lastly, the P-value of less than 0.0001 is lower than 0.01, thus the null hypothesis can be rejected at the 1% level of significance.

The QQQ 10-day moving simple moving average has a coefficient of 0.0363 which implies that every time the $QQQ 10-day moving average increases by $1, the closing price for $AAPL next day increases by $0.0363. The R-square of 0.991634504 indicates that 99.2% 0f the variation in Dow Jones corresponds to the variation in (d+1). The P-value of less than 0.0001 is lower than 0.01. Thus, the null hypothesis can be rejected at the 1% level of significance.

Analysis of the Quality of Forecast

The residual standard error represents the average distance that the observed values fall from the regression line. In this case, the derived function has a standard residual of 2.44556269. Considering $AAPL has had a price range of $6.6 to $460.04, the error is of low significance relative to the stock price. To assess the model's goodness of fit, the residual for each datapoint was also observed in a scatterplot. It was observed that as the value of the stock increased, the residual also increased. To formulate a more accurate conclusion, the residual was divided by the target price to find the relative margin of error. The calculation revealed that on 3193 out of 3732 of the days, the value of the relative error fell between -0.02 and 0.02, which means the predicted price deviated less than 2% from the target price 85.6% of the time. On the other hand, during March of 2020, however, it was observed that the model experienced a notably high standard error. That is due to the large magnitude correction across multiple sectors across the market driven by consumer fears of revenue shock caused by manufacturing shut down and supply bottlenecks. The inconsistency of this model under high volatility conditions is evidence that advanced models involving regression algorithms which take in market factors of potential influence are necessary for producing highly accurate price predictions.

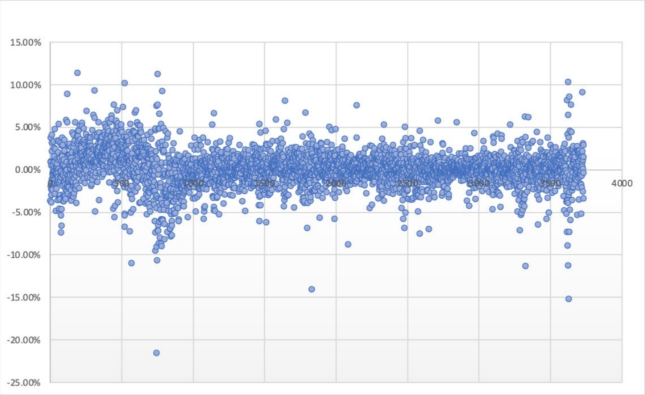

Figure 1. Residual graph of resulting equation

Conclusion

To conclude, the model experienced collinearity following forward stepwise regression, after removing the excess variable, c(t-1), the standard error of the model increased by $0.038. However, the elimination of the variable resulted in a less complex equation. The final model has an adjusted R-square value 0.9995, meaning 99.95% variation can be attributed to closing, 50d ma, Dow Jones intraday mean, and Nasdaq 10-day moving average. There is a slight problem as not all equities listed on the Nasdaq exchange have such strong positive correlation with the market indexes like Nasdaq and S&P 500: According to an analysis by the SPDR Research and Strategy team with data from Bloomberg Finance [7], only 52% of S&P 500 stocks have outperformed the index in the past ten years; small cap stocks are subjected to institutional manipulation and thus have price action inconsistent with market leaders; and commodity equities have low volatility due to low trading volume relative to the stock market. The model has limitations when it comes to representing other equities on the market and it's hypothesized that it would only have high accuracy when applied to large cap stocks that correlate strongly with the Nasdaq 100 Index. Several other shortcomings of the forecasting model have been addressed throughout the paper, including the model's predictive accuracy under high volatility. Even though the results show that 85.5% of the predicted values deviate less than 2% from the target price, the model is still subjected to relatively high standard error when the equity is experiencing high-volume after hours trading and high magnitude market corrections. Time series is a model which forecasts with respect to time: By taking into account seasonality, trends, cycles, and unexpected variations, the method provides the forecaster with greater flexibility in choosing time frames without compromising predictive accuracy. The combination of time series modeling and algorithmic regression presents opportunities for making a model that forecasts more accurate long-range predictions while using less input data points. For now, the multiple regression model examined in this paper is a simple and accessible forecasting method that yields accurate values which short-term traders can reference as a part of their risk management.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

[1] Yu, Minhui. "Linear regression model for stock price of Pfizer." *Proceedings of the 5th International Conference on Economic Management and Green Development*. Singapore: Springer Nature Singapore, 2022.

[2] Rishi, Taran. "Stock Market Analysis Using Linear Regression." *Proceedings of the Jepson Undergraduate Conference on International Economics*. Vol. 4. No. 1. 2022.

[3] Nunno, L., 2014. Stock market price prediction using linear and polynomial regression models. *Computer Science Department, University of New Mexico: Albuquerque, NM, USA*.

[4] De Bondt, Werner FM, and Richard H. Thaler. "Financial decision-making in markets and firms: A behavioral perspective." *Handbooks in operations research and management science* 9 (1995): 385-410.
[5]https://www.kaggle.com/code/nikhilkohli/stock-prediction-using-linear-regression-starter

[6] Galton, Francis. "Regression towards mediocrity in hereditary stature." *The Journal of the Anthropological Institute of Great Britain and Ireland* 15 (1886): 246-263.

[7]https://www.ssga.com/us/en/intermediary/etfs/insights/narrow-market-leadership-in-the-us-makes-international-stocks-attractive