Using the Nasdaq Index to Predict AAPL Price by Linear Regression Analysis

Junqi Jin[1, †, *], Haochen Ma[2, †]

[1]Department of Art and Science, University of Toronto, Toronto, Canada

[2]School of Small Business, Miami University-Oxford, Oxford, Ohio, United States

[*]Corresponding author: haochem@miamioh.edu

[†]These authors contributed equally.

Abstract. In this project, we want to predict AAPL's stock price by the NASDAQ index by the regression model. The dependent variable is AAPL's stock price, and the independent variable is the NASDAQ index. First, we do some descriptive statistics for the two variables and measure the distribution from the central tendency, variation tendency, and distribution to acknowledge the character of distributions. Based on the strong linear relationship between AAPL stock price and the NASDAQ index, we constructed a simple linear regression model. Considering the scale of the two variables, we tried the other three models with log transformation. And then, it shows that the log-log model has the best performance. However, in the residual analysis of the log-log model, it shows an autocorrelation in the residual, then we generate a new variable that is the one-order term for AAPL and add it into the model, and it surprisingly performs very well, whose R square is up to 99.72%. Therefore, we think combining the linear relationship with the market and the autocorrelation itself can construct a good model, and it can apply to predict much other stock's prices in the market.

Keywords: Stock price prediction; Linear regression model; Autocorrelation.

1. Introduction

1.1 Background

In the capital asset pricing model, we know the expected return risk-free interest rate of single security and the compensation for risk, and the size of the risk premium depends on the beta value. From the form of the model, we can also see that the market portfolio is related to a single security. There is a linear relationship between securities, and we can use regression models to describe this linear relationship. Furthermore, as it is also a time series data, there is an auto-correlated relationship for the financial asset price, and I want to take this autocorrelation into consideration.

Affected by the new crown pneumonia and the superposition of many factors, the global economy is facing huge challenges at present. For listed companies, stock prices play an important role in company management and market confidence. As the company with the highest market value in the world, Apple reflects the operating status and living conditions of high-tech manufacturing in the United States to a certain extent and reflects the confidence of the market from the consumer side. Therefore, it is of practical significance to study the stock price trend of Apple's formula.

1.2 Related Research

Stock price prediction is a famous domain in current financial research. Many models have been developed to predict a company's stock price. Based on traditional financial theory, Capital Asset Pricing Model and Fama-French Three-Factor Model have been wildly used in the past. EUGENE F. FAMA and KENNETH R. FRENCH use CAPM to explain how premiums vary with a company's size [1,2]. Michael J. Cooper, Huseyin Gulen, and Michael J. Schill tested the factors model's predictions for firm-level asset investment effects on returns by examining the cross-sectional relationship between firm asset growth and subsequent stock returns [3]. They discovered that a firm's annual asset growth rate emerges as an economically and statistically significant predictor of the cross-section of U.S. stock returns [4]. Additionally, statistical methods like time series tools and grey forecasts are utilized to anticipate stock prices. Ayodele A. Adebiyi and Aderemi O. Adewum (2014) describe a detailed procedure for developing an ARIMA-based stock price prediction model. Results showed that the ARIMA model can successfully compete with other stock price prediction methods and has a high potential for short-term prediction [5,6],

These days, machine learning and deep learning have advanced extremely swiftly, introducing several new models, and they exhibit great predictive capabilities. Numerous fields, including stock price prediction, have benefited from the use of machine learning and deep learning [7]. Mehar Vijh (2019) predicted the closing prices of five firms in various operating sectors the next day using Artificial Neural Network and Random Forest methods [8]. RMSE and MAPE, two common strategic measures, are used to assess these models. The models are effective in forecasting stock closing prices, as seen by the low levels of these two indicators [9]. Wenjie Lu proposes a CNN-BiLSTM-AM method to predict the stock closing price of the next day. Convolutional neural networks (CNN), bi-directional long short-term memory (BiLSTM), and attention mechanism (AM) make up this technique. The CNNBiLSTM-AM method is more suited than other methods for predicting stock prices and gives investors a reliable way to select which stocks to buy [10]. Although deep learning methods can provide "beautiful" results for prediction, it is hard to explain the model. Also, for deep learning method, it is complex. A large number of parameters are estimated, and it will be time-consuming. However, simple regression analysis for short-term prediction of stock price can also perform well, and in our final model, the R square is higher than 99%.

1.3 Objection

In our project, we want to predict AAPL's stock price by the NASDAQ index by the regression model. The dependent variable is AAPL's stock price, and the independent variable is the NASDAQ index. Besides, we will add the one-order lag term for AAPL's stock price into our regression model, which explains the auto-correlated relationship for the financial asset price. It assumes there exists a linear relationship between the two variables because AAPL is a global manufacturing company with advanced technology, and the NASDAQ index is made up of many global high-tech companies and is synonymous with the "new economy." Modeling AAPL's stock can help investors to make better decisions. Also, as our model is efficient and easy to construct, there are not too many thresholds for implementing the model. Everyone can use our model to predict the stock price in the NASDAQ market and get a good result. It is the real meaning of our model.

2. Methodology

2.1 Data Source

We collect data from the yahoo website and the data's time range is from August 1st, 2010, to August 1st, 2022, which includes 12 years of observations. There are three main variables, Date, AAPL closed price, and NASDAQ market index. In our further analysis, we will generate a new variable, which is the one-order lag term for AAPL stock price. Regression models are very useful models that we will use to explore the relationship between AAPL's stock price and the NASDAQ index. In our research, the dependent variable is AAPL's stock price.

2.2 Descriptive for AAPL and NASDAQ

There are three variables in our dataset, which are the date, AAPL's stock price, and the NASDAQ index. After merging the dataset, we get 3166 valid observations. In the following, we want to make some descriptive statistics for the variables.

Table 1. Descriptive table for AAPL and NASDAQ

Statistic	AAPL	NASDAQ
Mean	48.02	6021.09
Median	29.07	4565
Min	6.86	1728
Q1	18.77	2797
Q3	52.62	7599
Max	182.01	16573
IQR	33.85	4802
Range	175.15	14845
SD	44.46	3894.54
Skew	1.50	1.09
Kurtosis	1.02	0.15

From above Table 1, we can see in the measurement of central tendency, the mean and median of APPL's stock price are 48.02 and 29.07 dollars. The mean and median for the NASDAQ index are 6021.09 and 4565. In the measurement of variation tendency, the range and standard deviation of APPL's stock are 175.15 and 44.46, and the of NASDAQ are 14845 and 3894.54. Furthermore, from the skewness and kurtosis, we know both APPL and NASDAQ are positively skewed with leptokurtic.

We can see that from 2010, the NASDAQ index continued to increase until 2021 November (Fig. 1).

Also, the NASDAQ index continued to increase from 2010 until 2022 January, and it increased rapidly in 2019 due to the surge in demand for electronics due to the Covid-19 lockdown (Fig. 2).

From the scatter plot (Fig.3), as the NASDAQ index increases, the AAPL price tends to increase, and it shows a very strong positive relationship between the NASDAQ index and the AAPL price. It also implies that a simple linear regression model will have a good fit for the dataset.

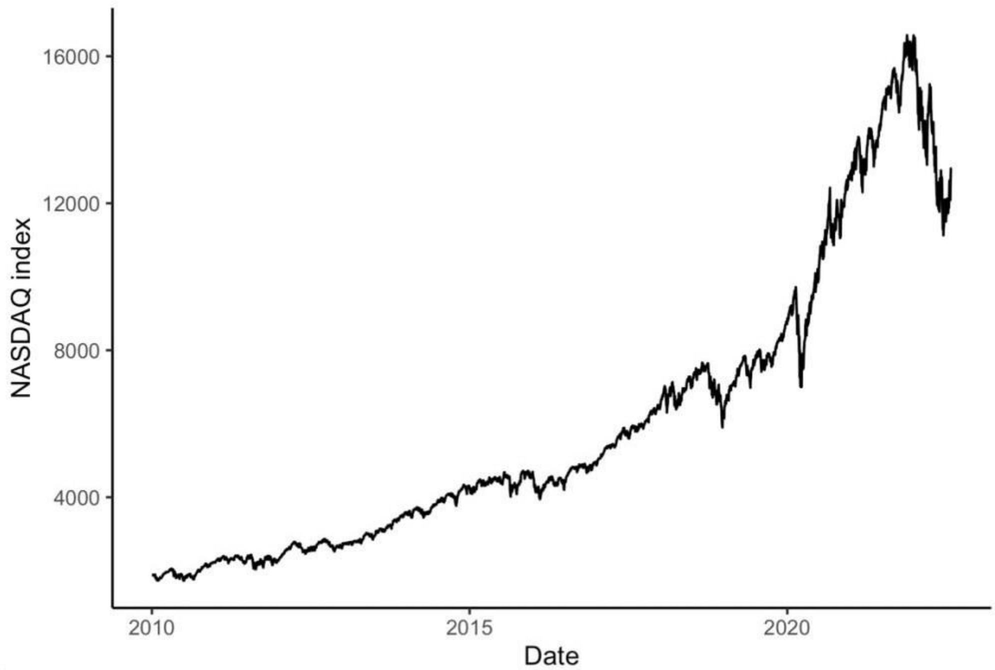

Fig. 1 The line chart for NASDAQ (Photo credit: Original)

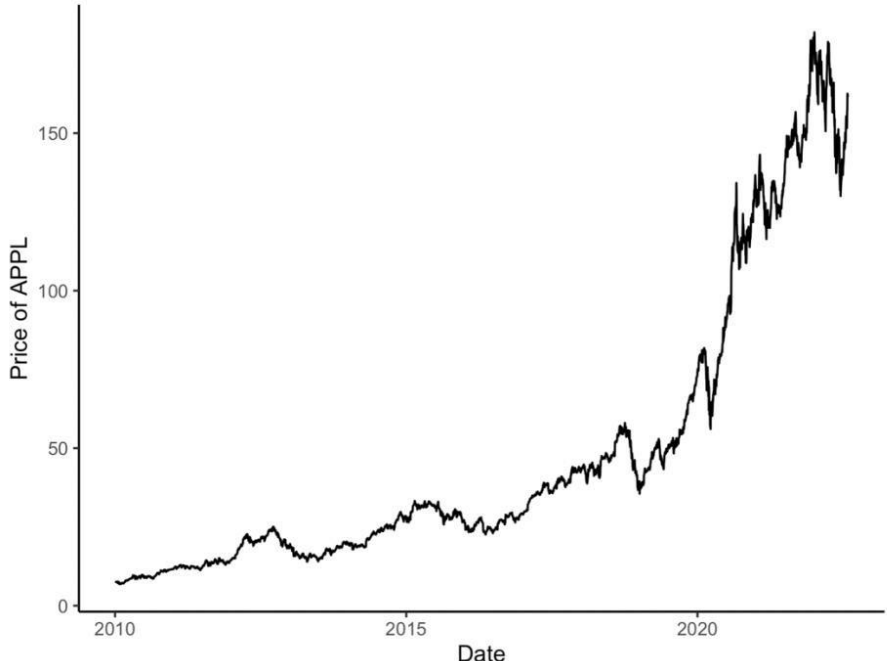

Fig. 2 The line chart for AAPL price (Photo credit: Original)

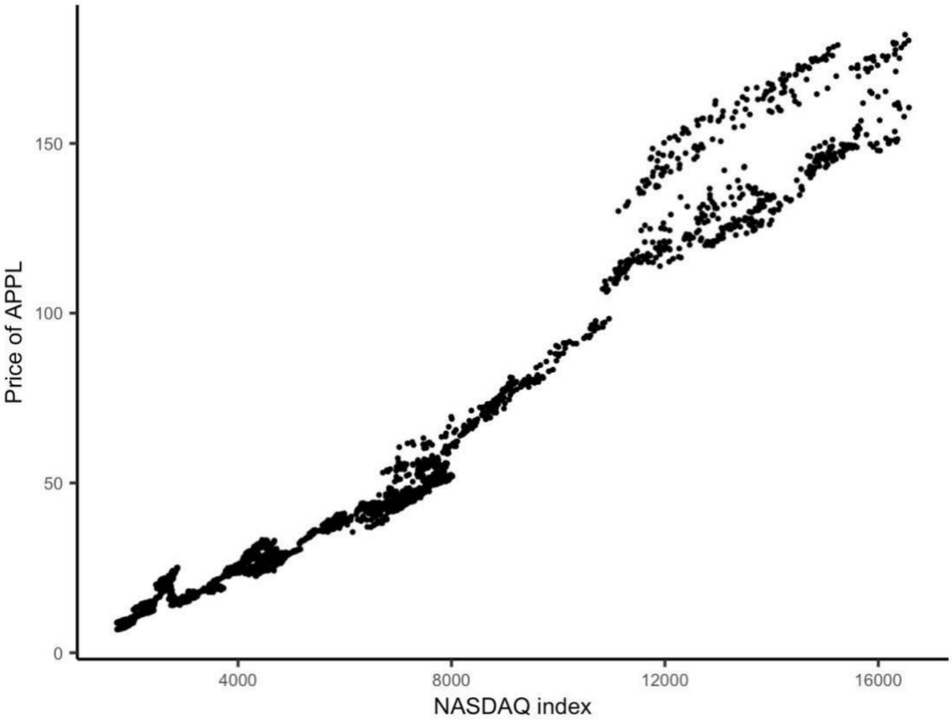

Fig. 3 The scatter plot between the NASDAQ index and AAPL price (Photo credit: Original)

2.3 Descriptive for AAPL and NASDAQ

The main purpose of the out project is to predict AAPL's stock by the NASDAQ index and I will use simple regression analysis to model this. And the model is as follows.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \quad (1)$$

Where $i = 1, 2, \ldots, n$,

And there are four basic assumptions for the linear regression model.

(i) linearity of the relation between the true mean of Y and X

(ii) independence of the error
(iii) constant variance of the errors
(iv) normality of the errors

Equation (1) implies the multiple linear regression model, which is widely used. However, in many situations, we can do the transformation for the model such as log transformation or power transformation. As the scale of involved variables is large, we use log transformation for both variables, and it performs better than the simple regression model by our comparison. Also, we add a one-order lag term of AAPL stock price into our model. And our final model is

$$\ln(AAPL_t) = \beta_0 + \beta_1 * \ln(NASDAQ_t) + \beta_2 \ln(AAPL_{t-1}) + \epsilon_t \tag{2}$$

This is an auto-regression model including a linear relationship with the NASDAQ.

2.4 Evaluation indicators

A good model should have good prediction and simple architecture. On the one hand, a model should have a good performance on our aim. On the other hand, we don't want our model to be more complicated because it has the risk of overfitting. So, we need a balance between performance and complexity.

R square is usually used to measure the fitness of the model, and it measures the proportion of variation of the response variable, which is complained by the model. But it doesn't consider the number of parameters. In our model, we will use adjusted R square and AIC criterion to measure.

$$R^2 = \frac{SSR}{SST} \tag{3}$$

Where SSR implies the total sum of regression square, and SST implies the total sum square.

$$Adj.R^2 = 1 - (1 - R^2) * \frac{n-1}{n-k-1} \tag{4}$$

n is the sample size and k is the number of parameters. A higher adjusted R square implies a better model.

And AIC criterion is

$$AIC = -2\ln\left(\hat{L}\right) + 2k \tag{5}$$

Where $\hat{L}$ implies the likelihood function, and k implies the number of parameters. Lower AIC implies a better model.

3. Model Development and Analysis

3.1 Model Development

The main aim of our paper is to predict AAPL's stock price. First, I constructed a simple regression model, the response variable is the daily AAPL stock price, and the predictor is the NASDAQ index. Then I try to do the transformation to variables, and I found that log transformation is appropriate compared with the simple regression model.

After building this model, I plot the residual plot (Fig. 4), which shows obvious autocorrelation. It means that we violated the basic assumption for the linear model, also, it means that we can extract more information from the residual. I also plot the ACF plot and PACF plot, the ACF plot show tail on and PACF tails off. It implies that AR(1) model is appropriate if we don't consider the linear relationship between AAPL stock price and the NASDAQ index. Therefore, I take the one-order lag term for AAPL stock price as another predictor and combine the information of linear relationship and autocorrelation.

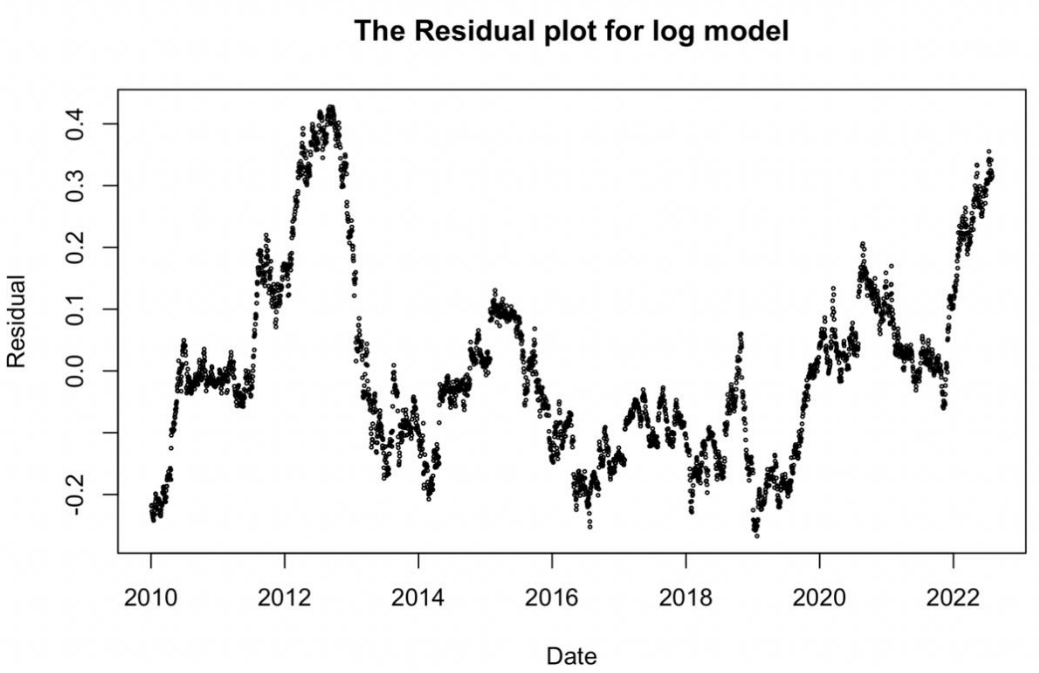

Fig. 4 The residual plot for the log-log model (Photo credit: Original)

We use R software to conduct our analysis and all the models are as follows.

Table 2. The output for multiple estimated model

	AAPL		ln(AAPL)		
	(1)	(2)	(3)	(4)	(5)
NASDAQ	0.0114*** (4.429e-05)		2.0e-04*** (1.02e-06)		
ln(NASDAQ)		64.2208*** (0.5725)		1.307*** (0.0043)	0.098*** (0.0067)
AAPL_lag					0.926*** (0.005)
Constant	19.06*** (0.3176)	-498.45*** (4.8845)	2.291*** (7.28e-03)	-7.608*** (0.0365)	-0.574*** (0.040)
Observations	3166	3166	3166	3166	3165
R^2	0.9524	0.7991	0.9269	0.9672	0.9972
$Adj.R^2$	0.9524	0.799	0.9269	0.9672	0.9972
AIC	14392.14	18949.81	-9515.304	-12048.88	-19794.69

The estimated model (1) is

$$E(AAPL) = 19.06 + 0.011 * NASDAQ \tag{6}$$

The F statistic is 6327 with a degree of freedom of 1 and 3164, and the p-value of the overall significance test is 0.000. At the 1% significance level, we do reject the null hypothesis and conclude that the overall model is significant and there really exists a linear relationship between NASDAQ and AAPL. The R square for the model is 0.9524, which means that a 95.24% variation of AAPL can be explained by the model. It's larger than 90% and implies that NASDAQ has a good prediction for AAPL price. The estimated slope is 0.011, which can be interpreted that when NASDAQ increases 1 unit, on average the AAPL stock price will increase by 0.011 dollars. For testing the significance of the slope, the t-statistic is 251.54 and the p-value is 0.000. At the 1% significance level, we fail to reject the null hypothesis and conclude the slope is significant.

Then we conduct log transformation for the response variable or independent variable or both and then constructed (2)-(4) models. By comparing the above model, we find out that only doing log transformation for AAPL or NASDAQ can't improve the performance of the model. The log-log model (4) can perform best among the four models.

The estimated model (4) is

$$\mathrm{E}\big(\ln(\mathrm{AAPL})\big) = -7.6076 + 1.3072 * \ln(\mathrm{NASDAQ}) \tag{7}$$

The F statistic is 9319 with a degree of freedom of 1 and 3164, and the p-value of the overall significance test is 0.000. At the 1% significance level, we do reject the null hypothesis and conclude that the overall model is significant, and the overall model is very significant. The R square for the model is 0.9672, which means that 96.72% variation of AAPL can be explained by the model. The adjusted R square for model (4) is higher than the adjusted R square for model (1), which implies that the log-log model is better than model (1)

The estimated slope is 1.3072, which can be interpreted that when NASDAQ increases by 1%, on average the AAPL stock price will increase by 1.3072%. For testing the significance of the slope, the t-statistic is 305.3 and the p-value is 0.000. At the 1% significance level, we fail to reject the null hypothesis and conclude the slope is significant.

Taking the autocorrelation relationship into consideration, I construct a model (5). And the estimated model is

$$\mathrm{E}\big(\ln(\mathrm{AAPL_t})\big) = -0.5744 + 0.0982 * \ln(\mathrm{NASDAQ_t}) + 0.9259 * \ln(\mathrm{AAPL_{t-1}}) \tag{8}$$

It is surprising that the R square for the model is up to 0.9972, which means that a 99.72% variation of AAPL stock price can be explained by this log-log model with autocorrelation term. It performs very well. The F statistic of testing the overall model significance is 5555, which implies that the overall model is very significant. Among the five models, model (5) has the highest adjusted R square and lowest AIC, which means the best model.

In this situation, the estimated coefficient for log NASDAQ is 0.0982, which can be interpreted that when NASDAQ increases by 1%, on average the AAPL stock price will increase by 0.0982%. The p-values for coefficients of ln (NASDAQ) and ln (AAPL_(t-1)) are all 0.000, which means that both coefficients are very significant, and the effect of autocorrelation does exist in AAPL stock price. Then we will come model diagnoses for our final model.

3.2 Model Diagnosis

From the above residual plot (figure5), we can see that the horizontal axis is the fitted value in the model (5) and the vertical axis is the residual of the model (5). We can see that all the points are distributed around 0. It's a good sign, which means that the assumptions of homogeneity and non-autocorrelation weren't violated. Compare with figure 4, we can conclude that we have extracted the information on autocorrelation in the model (4).

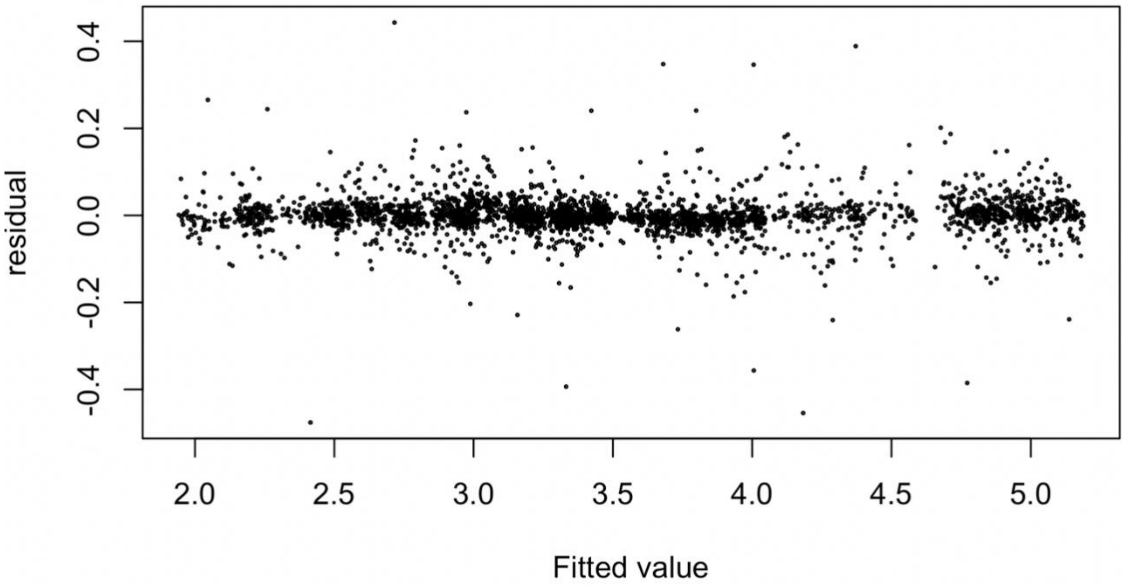

Fig. 5 The residual plot for model (5) (Photo credit: Original)

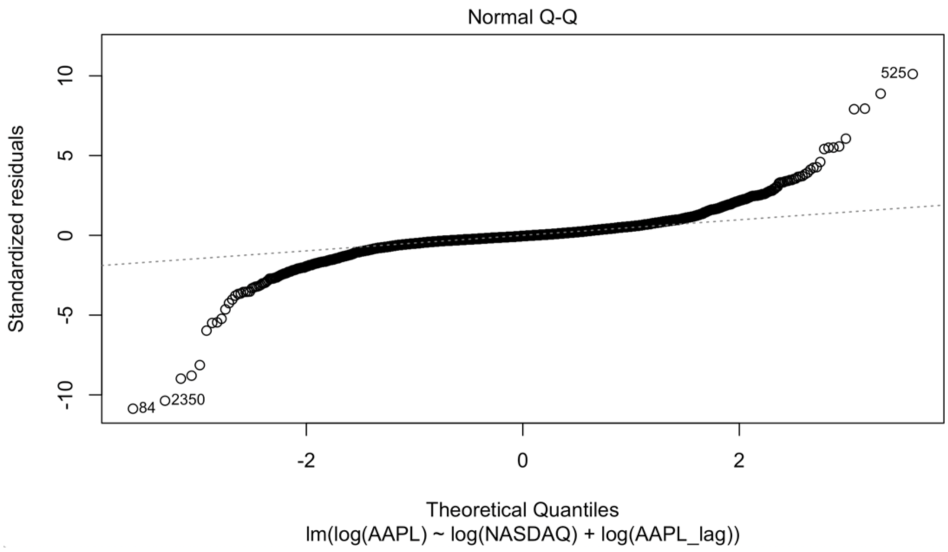

Fig. 6 The Q-Q plot for model (5) (Photo credit: Original)

From the above Q-Q plot for model (5), we can find out that the points show an "S" shape. It implies that the normality assumption is violated, and the distribution of log(AAPL) may be spike distribution.

As the number of independent variables is only two, there is no need to test the issue of multilinearity.

4. Conclusion

In this paper, we want to explore the relationship between the AAPL stock price and the NASDAQ index. We collected about 12 years of data and use simple regression to get the quantitative relationship between the two variables. First, we do some statistics for the two variables and measure the distribution from the central tendency, variation tendency, and distribution. Also, the two line charts show that the AAPL stock price and NASDAQ index continued to increase until about 2021. Besides, the scatter plot shows a strong positive relationship between the NASDAQ index and AAPL price. Then we constructed a simple regression model, knowing that the overall model and slope are significant. The R square is very high, which implies a good fit for the model. Also, the estimated slope is 0.011, which can be interpreted that when NASDAQ increases by 1 unit, on average, the AAPL stock price will increase by 0.011 dollars.

As the scale of variables is very large, I tried to do the transformation to variables, and I found that log transformation is appropriate compared with a simple regression model. Following this, I constructed four models with or without log transformation and found that the log-log model performs best. In the log-log model, 96.72% variation of AAPL can be explained by the model. The adjusted R square for model (4) is higher than that for model (1). And the estimated coefficient for log NASDAQ is 1.307, which can be interpreted that when NASDAQ increases 1%, the AAPL stock price will increase by 1.307%. Then we plot the residual plot for model 4 and apparent autocorrelation. It means that we violated the basic assumption for the linear model. Also, it means that we can extract more information from the residual. Combining the autocorrelation and log-log model together, we constructed model (5) and surprisingly found out that the R square for this model is 0.9972, which is super high. The p-value of the lag term for AAPL is 0.000, which means that the effect of the one-order lag term is significant. Finally, we make some model diagnoses for model (5), showing that the assumption of homogeneity and autocorrelation weren't violated.

In our final model, it gets a very high goodness of the data. It shows the powerful ability of our model to predict the AAPL stock price. A combination of the linear relationship with the market and the autocorrelation itself can construct a good model, and it can apply to predict much other stock's prices in the market, and we believe it will have outstanding performance.

References

[1] Fama, E.F. and French, K.R., 2006. The value premium and the CAPM. The Journal of Finance, 61(5), pp.2163-2185.

[2] Ican, O. and Celik, T.B., 2017. Stock market prediction performance of neural networks: A literature review. International Journal of Economics and Finance, 9(11), pp.100-108.

[3] Cooper, M.J., Gulen, H. and Schill, M.J., 2008. Asset growth and the cross-section of stock returns. the Journal of Finance, 63(4), pp.1609-1651.

[4] Kisman, Z. and Restiyanita, S., 2015. M. The Validity of Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT) in Predicting the Return of Stocks in Indonesia Stock Exchange. American Journal of Economics, Finance and Management, 1(3), pp.184-189.

[5] Wang, M. and Wang, Y., 2019, September. Evaluating the Effectiveness of Candlestick Analysis in Forecasting U.S. Stock Market. ICCDA 2019: Proceedings of the 2019 3rd International Conference on Compute and Data Analysis, pp.98-101.

[6] Ariyo, Adebiyi A., Adewumi O. Adewumi, and Charles K. Ayo. "Stock price prediction using the ARIMA model." 2014 UKSim-AMSS 16th international conference on computer modelling and simulation. IEEE, 2014.

[7] Vijh, M., Chandola, D., Tikkiwal, V.A. and Kumar, A., 2020. Stock closing price prediction using machine learning techniques. Procedia computer science, 167, pp.599-606.

[8] Patil, P.R., Parasar, D. and Charhate, S., 2021. A Literature Review on Machine Learning Techniques and Strategies Applied to Stock Market Price Prediction. Emerging Trends in Data Driven Computing and Communications, pp.121-135.

[9] Yu, P. and Yan, X., 2020. Stock price prediction based on deep neural networks. Neural Computing and Applications, 32(6), pp.1609-1628.

[10] Lu, W., Li, J., Wang, J. and Qin, L., 2021, September. A CNN-BiLSTM-AM method for stock price prediction. Neural Computing and Applications, 33(3), pp.30-33.