Estimation of Close Price Netflix's Stock in US Market using Multiple Regression Analysis and Appropriate Regression Model in Correction of the Residual Assumptions

Patrick Jonathan
Statistics Department
School of Computer Science,
Bina Nusantara University
Jakarta, Indonesia 11480
patrick.jonathan001@binus.ac.id

Devin Augustin
Statistics Department
School of Computer Science,
Bina Nusantara University
Jakarta, Indonesia 11480
devin.augustin@binus.ac.id

Rafi Muhammad
Statistics Department
School of Computer Science,
Bina Nusantara University
Jakarta, Indonesia 11480
rafi.muhammad001@binus.ac.id

Grivaldo Mahesa Wiwenar
Statistics Department
School of Computer Science,
Bina Nusantara University
Jakarta, Indonesia 11480
grivaldo.wiwenar@binus.ac.id

Margaretha Ohyver
Statistics Department
School of Computer Science,
Bina Nusantara University
Jakarta, Indonesia 11480
mohyver@binus.edu

Abstract—During the COVID-19 pandemic, the U.S. Stock market has seen massive growth. The cause of this growth is people staying in their homes because of the lockdown issued by the government. People have nothing to do, so they pick on an interest in the stock market. Many people start investing in stocks they like which causes the stock's price to go up. One of these stocks is Netflix. This study aims to develop a multiple regression model using the Open Index, High Index, Low Index, and Volume Index to help estimate the closing price of Netflix's stock. We also developed the model by using Ridge and Weighted Least Squares Regression to handle some of the residual assumptions. At the end of our research, we find out that the Ridge Regression model is the best model to estimate the Close Price, where the model is selected based on the comparison of AIC, BIC, and Adjusted R squared values of each model.

Keywords— Netlflix, Stock, Multiple Linear Regression, Weighted Least Square, Ridge Regression.

I. INTRODUCTION

Since the COVID-19 pandemic began, many people have their sights on the stock market and begin buying stocks of their preferences. This happened because a lot of people are stuck in their own home doing nothing, some got laid off from their jobs and some started to look for new hobbies. As a developing country, Indonesia in the economic sector is also affected by the Covid-19 pandemic in the stock market sector which is one type of securities traded on Bursa Efek.

According to Rusdin Shares are certificates that show proof of ownership of a company, and the owner has the right to the company's income and assets[1]. A stock (also known as equity) is a security that represents the ownership of a fraction of a corporation[2]. This entitles the owner of the stock to a proportion of the corporation's assets and profits equal to how much stock they own. Units of stock are called "shares".

Stocks are usually bought and sold from online stockbrokers. These types of transactions are regulated by the governments as a means to protect investors from fraudulent practices. This type of investment usually beats other types of investment in the long run.

The stock market was first created in Amsterdam in 1611 and the only stock available at the time was Dutch East India Company. Meanwhile, the U.S. stock market was founded in the late 1700s and known as The New York Stock Exchange(NYSE)[3].

The U.S. stock market or NYSE, generally opens at 9.30 a.m. - 4.00 p.m. Eastern time. People can buy stocks on Weekdays (Monday - Friday), while on Saturday, Sunday, and public holidays, they can't[4].

There are many components in Stocks, such as open, low, high, close, and volume. In stock trading, the high and low refer to the maximum and minimum prices in a given time period. Open and close are the prices at which a stock began and ended trading in the same period. Volume is the total amount of trading activity—adjusted values factor in corporate actions such as dividends, stock splits, and new share issuance.

Netflix's stock is one of the most traded stocks since the pandemic began. Netflix is a subscription based streaming service company that offers a lot of variety shows and movies on their platform. By 2020, Netflix has around 209.18 million subscribers worldwide [5]. As one of the most traded stocks, we are curious to find out the estimation of the close price using other variables given, by using Multiple Regression

We can also handle the assumption violations if there are any. When the multiple regression model is obtained, we must check the normality, the homoscedasticity, and the nonautocorrelation assumption. We can also check the multicollinearity that shows a strong correlation or relationship between two or more independent variables in a multiple linear regression model. Unfortunately, the model did not fit the homoscedasticity assumption and there is multicollinearity in this model. Therefore, we developed the model using Ridge and Weighted Least Squares Regression to handle these violations.

II. METHODOLOGY

A. Multiple Linear Regression

Multiple linear regression is a linear regression model that includes one dependent or response variable with two or more independent or predictor variables [6]. The equation of the multiple linear regression has the same form as the simple linear regression, but with more than one parameter and predictor variables. The multiple regression model is:

$$y = {}_0 + {}_1 x_1 + {}_2 x_2 + \ldots + {}_p x_p + \varepsilon$$

Where,
y = Dependent variable
x = Independent variable
β = regression coefficients.
ε = Residual value

We can solve the Multiple Linear Regression by using the Ordinary least-squares (OLS) method. This method can be used to estimate the value of a response variable by using two or more predictor variables. We can also use this method to check if there is any influence between the response variable and predictor variables. The Ordinary Least Square method will give the best estimation if all the classical assumptions are met. The formula of the Ordinary least-squares regression method is :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Where,
X = Matrix of independent variable
X^T = Matrix of independent variable in transpose form
Y = Matrix of dependent variable

B. Residual Assumptions

There are several residual assumptions that must be met to build a good regression model [7], which are :

1. The errors must have a normal distribution or met the normality assumption
2. The errors must have a constant variance or met the homoscedasticity assumption
3. The errors must be independent from each other or non autocorrelated

C. Multicollinearity Test

Multicollinearity is a statistical phenomenon in which two or more predictors variables in a multiple regression model are highly correlated. If there is no linear relationship between predictor variables, they are said to be orthogonal [10].

Multicollinearity may be present if The algebraic signs of the estimated coefficients do not conform to the prior expectation or the coefficients of variables that are expected to be important have large standard errors (small t-values).

Checking the correlation between predictors is not the best approach since it has some downsides and limitations. That is why most of the time people use Variance Inflation Factors (VIF) with criteria if the VIF value is more than 10, then the regression model has multicollinear. We could calculate VIF by using this formula:

$$VIF_i = \frac{1}{(1-R_i^2)}$$

Where,
i = The number of predictor variables
R_i^2 = Multiple R-square

The bigger R_i^2, the smaller VIF value obtained and vice versa.

D. Ridge Regression

Ridge regression is a method to tune a model that is used to analyze any data that suffers from multicollinearity. It is a method to estimate the coefficients of multiple-regression models where the variables are highly correlated. [8]

The ordinary least squares estimator for the regression coefficient use this following formula:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

And by adding a small constant value of λ to the diagonal entries of the matrix $X^T X$, this following formula could be improved resulting the ridge regression estimator :

$$\hat{\beta}_{ridge} = (X^T X + \lambda I_p)^{-1} X^T Y$$

Where,

$I_p = p \times p$ identity matrix
β = Regression Coefficient
X = Independent Variable
λ = Penalty Term

In Ridge Regression we want to minimize both the error and the size of the coefficient, so we add the penalty term to find the balance between these two objectives.

$$L_{ridge}(\widehat{\beta}) = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

E. Weighted Least Squares

To find out whether the data has heteroscedasticity error, the WLS method can be used. The WLS method is a development of the Ordinary Least Square (OLS) method, namely by adding a weighting function to the linear regression model of least squares, to determine the estimator of the model parameters. Heteroscedasticity error can be corrected by either transforming the predictor variable or by transforming both sides [12]. The WLS method presented here is used as a way to handle heteroscedastic errors. The WLS equation is shown below.

$$\widehat{\beta}_{WLS} = (X'WX)^{-1} X'WY$$

Where,
$W = Weighted\ Matrix$
$X = Independent\ Variable\ Matrix$
$Y = Dependent\ Variable\ Matrix$

With a weighted matrix using the equation below.

$$W = \frac{1}{\alpha_i^2} I$$

Where,
$\alpha_i^2 = Variety\ Group$
$I = Identity\ Matrix$

F. Dataset

In this paper, the dataset is obtained from investing.com with a total of 253 data and 6 features from the 2021 Netflix's stock in the US market. Investing.com is a financial markets platform providing real-time data, quotes, charts, and many more across 250 exchanges from around the world. The data contains Date, Open, Close, High, Low, and Volume.**"Date"**, it contains all the dates when the market is opened. **"Open"**, it contains all the opening price of the share on that day. **"Close",** it contains all the closing price of the share on that day. **"High"**, it contains all the highest price of the share on that day. **"Low"**, it contains all the lowest price of the share on that day. And **"Volume",** it contains all the number of shares traded on that day.[11]

Here is the Exploratory Data Analysis (EDA) performed on the dataset. Fig. 1 shows the summary of the dataset calculated using R software.

Date	Open	High	Low	Close	Volume
Length : 252	Min. : 479.8	Min. : 488.6	Min. : 478.5	Min. : 485.0	Min. : 1287513
Class : character	1st Qu. : 512.6	1st Qu. : 518.0	1st Qu. : 505.6	1st Qu. : 512.3	1st Qu. : 2597019
Mode : character	Median : 544.2	Median : 551.7	Median : 538.1	Median : 543.3	Median : 3241963
	Mean : 558.5	Mean : 565.1	Mean : 551.6	Mean : 558.2	Mean : 3915339
	3rd Qu. : 598.3	3rd Qu. : 609.4	3rd Qu. : 593.7	3rd Qu. : 599.8	3rd Qu. : 4325963
	Max. : 692.4	Max. : 701.0	Max. : 686.1	Max. : 691.7	Max. : 32637449

Fig. 1. Summary of dataset

From here, we can see that **"Date"** is the only variable with a character class. Therefore we will not be using it during our model building and testing.

III. RESULTS AND DISCUSSION

In this dataset, we want to find out the estimation of the closing price using other variables given so that the close variable becomes the response and other variables become the predictor variables. First of all, we want to see what variables have a significant effect on the close variable. Then we do the variable selection method to eliminate variables that do not significantly affect the response variable in order to obtain a good regression model to estimate the close variable later on [13]. In this research, we are going to use the backward elimination method where we enter all the predictors and then we eliminate one by one until there are only significant predictors remaining.

We are going to use the R software to help us calculate the significance value (t-test) for every predictor variable. If the significance value is bigger than **α**(0.05), then the variable has no significant effect on the response variable, so that we can eliminate it.

TABLE I. P-VALUE OF PREDICTOR VARIABLES

Predictor Variables	**P-Value**
Open	$< 2 * 10^{-6}$
High	$< 2 * 10^{-6}$
Low	$< 2 * 10^{-6}$
Volume	0.286

From the table above, we can see that almost all p-value or significant value from every predictor is bigger than α(0.05) except the volume variable. Therefore, we will eliminate the **“Volume”** variable and we are going to check the significant value again for the remaining predictor variable.

TABLE II. P-VALUE OF PREDICTOR VARIABLES AFTER ELIMINATE VOLUME

Predictor Variables	**P-Value**
Open	$< 2 * 10^{-6}$
High	$< 2 * 10^{-6}$
Low	$< 2 * 10^{-6}$

After we checked again, it turns out that all other predictor variables have a significant effect on the response variable. This means that we can use the **“Open”**, **“High”**, and **“Low”** variables to create Multiple Linear Regression using the Ordinary Least Squares method. Then we get the following equation:

New Model (After Elimination of Volume):
Close = 0.40953 + 0.70173(*High*) + 0.87185(*Low*) + (-0.57229)(*Open*)

Next, we want to check the residual assumption tests using the Kolmogorov-Smirnov test for normality test, Durbin-Watson test for non-autocorrelation test, and Breusch-Pagan test for homoscedasticity test. With the help of R software, we get the p-value for each test. For each test, we want the p-value to be greater than α (0.05) to meet all the residual assumptions.

TABLE III. P-VALUE OF REGRESSION EQUATION IN RESIDUAL ASSUMPTION TESTS

Assumption Tests	**P-Value**
Kolmogorov-Smirnov	0.1963

Durbin-Watson	0.4749
Breusch-Pagan	$1.671 * 10^{-7}$

Based on the p-value results from table III, it can be concluded that this model fulfills the assumption of normality and the assumption of non-autocorrelation. However, this model does not meet the assumption of homoscedasticity because only the Breusch-Pagan p-value (Homoscedasticity Test) is smaller than (0.05). Therefore, we have to handle this violation with other regression models in order to produce a more optimal regression model. One of the regression models that can handle heteroscedasticity is Weighted Least Squares Regression. In this method, a weighting process is used on the error from the OLS Regression which produces a new error variant of the homoscedastic WLS regression.

Next, we check whether the OLS regression model that has been made is multicollinear by checking the VIF value of each predictor variable. The following is the VIF value of each predictor variable:

TABLE IV. VIF VALUES OF PREDICTORS

Predictor Variables	**VIF Value**
Open	126.6115
High	127.1421
Low	114.0850

Based on the value of table IV, it can be seen that all predictor variables both **"Open"**, **"High"**, and **"Low"** have a VIF value greater than 10, so it can be concluded that there is multicollinearity in this model. Therefore, we need another regression model to handle this violation. One of them is to use the ridge regression model. Ridge regression method can be used to overcome the high correlation between several predictor variables. Ridge regression is a method of estimating the regression coefficient obtained by adding λ bias constant to the X'X diagonal. This method produces a biased regression coefficient estimator, but can approach the true parameter value [14].

After being calculated using R software, the coefficient values of each model are obtained as follows:

TABLE V. REGRESSION COEFFICIENTS FROM EACH MODEL

Regression Model	**Intercept**	**Open**	**High**	**Low**
Ordinary Least Squares Regression (OLS)	0.40953	0.70173	0.87185	-0.57229
Ridge Regression	0.40950	- 0.57320	0.70170	0.87180
Weighted Least Squares Regression (WLS)	0. 44826	- 0.55292	0.69463	0.85943

TABLE VI. AKAIKE INFORMATION CRITERION (AIC) AND BAYESIAN INFORMATION CRITERION (BIC) VALUES FOR EACH MODEL

Regression Model	**AIC**	**BIC**
Ordinary Least Squares Regression (OLS)	1364.908	1382.555
Ridge Regression	645.76319	2049.76754
Weighted Least Squares Regression (WLS)	1357.43	1375.077

When comparing the AIC and BIC values of the Ridge Regression with Ordinary Least Squares Regression, it can be seen that the Ridge Regression has a smaller AIC value, but a larger BIC value than Ordinary Least Squares Regression. Meanwhile, when comparing the AIC and BIC values of Weighted Least Square Regression with Ordinary Least Squares Regression, it can be seen that both AIC and BIC values of Weighted Least Square Regression are smaller than Ordinary Least Square Regression. Because Weighted Least Square Regression excels in 2 categories which is AIC and BIC, we can conclude that Weighted Least Squares Regression is the best model to estimate the Close price, followed by Ridge Regression, and last Ordinary Least Squares Regression. Hence, the estimation of the Close Price by using Weighted Least Square Regression Model as follows :

$$\text{Close (WLS)} = 0.44826 - 0.55292(Open) + 0.69463(High) + 0.85943(Low)$$

IV. CONCLUSION

In this research, we can conclude that the closing price of Netflix's stock could be estimated using Weighted Least Square Regression with Open Index, High Index, and Low Index as the regressor variables. The developed model is selected based on the comparison of AIC and BIC values of the three models, Ridge Regression, Weighted Least Squares, and Ordinary Least Squares Regression. In which the Weighted Least Square Regression showed a better accuracy and fit towards our data. The estimated closing price could be used to help determine the closing price for Netflix’s stock price in the future, so that people could determine the price they want to buy the stock at.

REFERENCES

[1] Rusdin, “Pasar Modal: Teori, Masalah, dan Kebijakan dalam Praktik,” 2008.

[2] https://www.investopedia.com/terms/s/stock.asp

[3] https://www.sofi.com/learn/content/history-of-the-stock-market/

[4] https://www.nyse.com/markets/hours-calendars

[5] https://www.whats-on-netflix.com/news/netflix-library-by-the-numbers-2020/

[6] K. Marill, "Advanced Statistics: Linear Regression, Part II: Multiple Linear Regression", Academic Emergency Medicine, vol. 11, no. 1, pp. 94-102, 2004.

[7] S. Chatterjee and A. Hadi, Regression analysis by example, 4th ed. 2006.

[8] Hilt, Donald E.; Seegrist, Donald W. (1977). *Ridge, a computer program for calculating ridge regression estimates*. doi:10.5962/bhl.title.68934.

[9] https://online.stat.psu.edu/stat857/node/155/

[10] https://iopscience.iop.org/article/10.1088/1742-6596/949/1/012009/pdf

[11] https://analyzingalpha.com/open-high-low-close-stocks

[12] R. Dennis Cook and Sanford Weisberg, “Diagnostics for Heteroscedasticity in Regression”, 1 April 1980.

[13] Variable Selection in Regression Tutorial

[14] Ohyver, M. Metode Regresi Ridge untuk Mengatasi Kasus Multikolinear. ComTech. (2011).