Journal of Engineering Sciences (Improsci)
e-ISSN: 3031-7088
p-ISSN: 3032-3452

Evaluation of *the Multiple Regression Analysis* Algorithm on Stock Market Prediction

Mohamat Setiawan [1], Ade Johar Maturidi [2*], Dian Novianti [3]
[1]Polytechnic of Indonesian Institute of Education and Professional Development, Cirebon, Indonesia
Email setiawan.mohamat@gmail.com
[2 *]Polytechnic of Indonesian Institute of Education and Professional Development, Cirebon, Indonesia
Email ade.johar@gmail.com
[3] University of Muhammadiyah Cirebon, West Java, Indonesia
Email dian.novianti@umc.ac.id
***Corresponding Author Email ade.johar@gmail.com**

Abstract. Financial time series is one of the most challenging applications of modern time series forecasting. The financial time series is closely related to noise, non-stationary, and deterministic chaos. The characteristics suggest that no complete information can be obtained from the past behavior of financial markets to fully capture the dependency between future prices and that of the past. The data collection method was collected from the Stock Market Online Application "MetaTrader version 4" type "Daily" with a time range from "03/09/2001 to 25/07/2012", as many as 2052 data", with the attributes "Date, Open, High, Low, Close, Volume" with the main attribute "Close" using the Support vector machine algorithm, artificial neural network, and multiple linear regression. The conclusion of the value that is close to the series value is the value by testing on the support vector machine algorithm, with the parameter for the RMSE value that is close to the "0" value obtained from the measurement results on the SVM algorithm on the RBF kernel (radial base function) with a value of "gamma" $\gamma = 100$ with the value of RMSE = 0.000, and SE = 0.000. with prediction accuracy error = 0.976

Keywords: stocks, algorithms, multiple, regression, analysis

INTRODUCTION

Regression analysis in statistics is one of the methods to determine the cause-and-effect relationship between one variable and another variable(s). The "causation" variable is referred to by various terms: *explanatory variable*, *explanatory variable*, *independent variable*, or independent *variable X* (as it is often depicted in graphs as abscess or X-axis). The affected variable is *the affected variable*, the *dependent variable*, *the bound variable*, or *the Y variable*. These two variables can be random, but the influenced variable must always be a random variable.

Regression means forecasting, estimation, or estimation was first introduced in 1877 by Sir Francis Galton (1822-1911). Regression analysis is used to determine the shape of the relationship between variables. The primary purpose of using the analysis is to forecast or estimate the value of a variable to other variables. In addition to the linear relationship of two variables, the linear relationship of two variables can also occur, for example, the relationship between sales results and price and purchasing power.

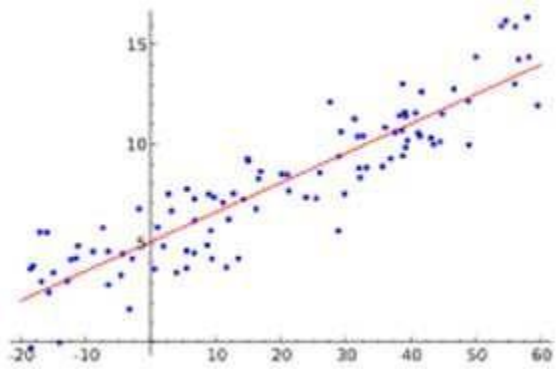

Figure 1 Regression analysis
(source: http://en.wikipedia.org/wiki/Linear_regression)

The linear relationship of more than two variables, when expressed in the form of mathematical equations, is:

$Y = a + b1x1 + b_{2x2} +$**.....**BKXK

Information:

x, x1, x2.....xk = variables

a, b1, b_2..... bk = constant number (constant) of variable coefficient

Common stock can be defined as securities as evidence of a statement or selection of individuals or institutions in a company. If an investor buys shares, he will become the owner and be referred to as a company shareholder. There are usually two types of shares: in the name and in the designation. For shares in the name, the name of the owner of the shares is listed on the shares, while the shares on designation, namely the name of the owner of the shares, is not listed on the shares, but the owner of the shares is the one who holds the shares. All rights of shareholders will be given to the custody of the shares.

Investing in stocks, also known as stocks, reflects an individual or business entity's

efforts to own a company or limited liability company. Stocks are one of the essential instruments in the capital market that can provide profits for investors through capital gains, namely the difference between the selling price and the buying price of shares (Hardiningsih, 2001). However, stocks also have risks in the form of capital losses, namely a decline in stock prices. Therefore, investors must conduct careful analysis to minimize such risks, considering that stock prices fluctuate in time series data. To overcome these challenges, investors can apply the science of forecasting or prediction, a systematic process of estimating the most likely future possibilities based on today's information. Although forecasts do not provide definitive answers about the future, the goal of predictions is to provide the most accurate forecasts possible. In the context of the stock market, predicting trends in the stock market is not a simple process; therefore, in this study, technical analysis is used, such as the use of the moving average method as an approach that helps improve the quality of prediction by using historical stock data. This research was conducted for stock market prediction using the evaluation of multiple regression analysis algorithms to predict the stock market.

LITERATURE REVIEW

Regression Analysis

Regression analysis is one of the most popular and widely used analyses. Regression analysis is widely used to make predictions and predictions, with uses that complement the field of machine learning. This analysis is also used to understand which independent variables are related to the bound variable and determine the forms of these relationships.

Multiple Linear Regression In statistics, linear regression is an approach to modeling the relationship between the scalar of the dependent variable *y* and one or more explanatory variables denoted by *X*. The case of one explanatory variable is called *Simple Regression Linear* (simple linear regression). For more than one explanatory variable, it is called *Multiple Regression Linear.* (This should be distinguished from *the double linear regression model*, in which multiple dependent variables are correlated rather than a single scalar variable).

Double Linear Regression Equation

Multiple linear regression is a regression in which its bound variable (Y) is linked or described by more than one variable, possibly two or three, following the independent variable (x, x1, x2, xn) but still showing the diagram linear relationships. The addition of

this independent variable is expected to better explain the characteristics of the existing relationship, even though some variables are still overlooked.

The general form of a multiple linear equation can be written as follows:

a. Stochastic shape

$\hat{y}$= a + b1x1 + b_{2x2} + b3x3 **.....**BKXK+C

b. Non-stochastic form

$\hat{y}$= a + b1x1 + b_{2x2} + b3x3**.....**BKXK

Information

$\hat{y}$ Bound variable (expected value y)

a, b1, b2 b_3.....bk : regression coefficient

x1, x_2 x3.....xk : free variable

e : bully error

Forecasting and Testing Regression Coefficients

Regression standard error and multiple regression coefficients

The standard error or difference in the standard estimate of regression is a value that expresses how far the regression value deviates from the actual value. This value is used to measure a forecast's accuracy level in predicting value. If this value equals zero, then the forecast has a 100% accuracy rate.

The standard error or difference of the standard estimate of multiple regression is formulated.

$$Se = \sqrt{\frac{\sum y^2 - b_1\left(\sum X_1 Y\right) + b_2\left(\sum X_2\right)}{n - m}}$$

Information

Se : Multiple regression standard error n : Number of observation pairs

m : the number of constants in multiple regression equations.

For coefficients b1 and b_2, the default error is formulated

$$Sb1 = \frac{O}{\sqrt{\left(\sum x_1^2 - NX_1^2\right)\left(1 - r^2 y_1\right)}}$$

$$Sb2 = \frac{O}{\sqrt{\left(\sum x_2^2 - NX_2^2\right)\left(1 - r^2 y_1\right)}}$$

Forecasting of multiple regression coefficient intervals (parameters B1 and B_2) Parameters B1 and B_2 It is often also referred to as the partial regression coefficient. Estimation of parameters B1 and B_2 Using a t distribution with free degrees db = n – m in general estimation

of parameters B1 and B_2 are : $B1 - T_{A/2N-M}\,SBI \leq Bi \leq Bi + T_{A/2N-M}\,SBI$ where i = 2.3
Multiple regression coefficient hypothesis testing (parameters B1 and B2) Hypothesis testing for multiple regression coefficients or partial regression of parameters B1 and B2 can be distinguished into two forms: simultaneous hypothesis testing and individual hypothesis testing.

Individual hypothesis testing is a test of the multiple regression coefficient hypothesis with only one B (B1 and B2) influencing Y. Simultaneous hypothesis testing is a test of the multiple regression coefficient hypothesis with B1 and B_2 simultaneously or together affecting Y.

METHOD

The methods used in this study include research design, experimental design, source information, and a summary of the methodology are determined. The research design leads to the steps taken in conducting this research. The experimental design is set to determine the experimental settings to be carried out through this study. The information collected is the source of information related to this research problem. The final part of this chapter is a summary of the overall research methodology.

The data collection method was collected from the Stock Market Online Application "MetaTrader version 4" type "Daily" with a time range from "03/09/2001 to 25/07/2012", as many as 2052 data", with the attributes "Date, Open, High, Low, Close, Volume" with the main attribute "Close" using the Support vector machine algorithm, artificial neural network, and multiple linear regression.

The method used in this study is an experiment using secondary data and data mining techniques. Secondary data is obtained from previously collected sources. Furthermore, data mining techniques are used to analyze the data to identify patterns, trends, and relationships relevant to stock price predictions. This approach allows researchers to dig deep insights from existing data and apply appropriate data mining algorithms to build accurate and informative prediction models.

DISCUSSION

Forecasting with Double Linear Regression

Forecasting the value of Y using multiple linear regression can be done if the regression line equation has been estimated and the values of the independent variables

x1, x2 are known.

A multiple linear regression line equation can be used in forecasting by first testing the hypothesis on its partial regression coefficients. The goal is to find out whether the independent variables used have a real influence on the y . The independent variables x1 and$_{x2}$ are said to have a real influence if in the hypothesis test the partial coefficients H0 : B_1 = B2 = 0 are rejected or H_1 : B1 ≠ B2 ≠ 0 is accepted, especially at the real level of 1%

The advantage of forecasting y using multiple linear regression is that it can be known quantitatively the magnitude of the influence of each independent variable (x1 or x2) if the influence of the variable is considered constant. For example, a multiple regression equation

$y = a + b1x1 + b_{2x2}$

Information:

y : Student statistical score

x1 : Student intelligence value

x2 : Frequency of truancy of students

B1 : Effect of x1 on y if x_2 constant

B2 : Effect of x2 on y if x_1 constant

The writing of multiple linear regression line equations is usually accompanied by the default error of each free variable and the multiple determination coefficient r2, as a measure of the accuracy of the line so that the approach.

Mean Square Error

An estimator's goodness can be seen from the magnitude of the error rate. The smaller the error rate, the better the estimator. One measure of the goodness of an estimator is MSE. An estimator $\hat{f}(\ )$ has an MSE:

$(\hat{f}(\ ))$ = var($\hat{f}(x)$) + bias2 ($\hat{f}(x)$) with bias ($\hat{f}$) = E(-f(x)). Because f(x) is unknown, the MSE value cannot be known either, so it is necessary to estimate the MSE. provide an estimate from MSE as

$$MSE = n^{-1} \| y - \hat{y} \|^2 = n^{-1} \sum_{i=1}^{n} (y - \hat{y})^2$$

Symbol description:

Yi : Prediction value for the ith data

Yi : The actual output value for the ith data

N : the amount of data

Bootstrap is a resampling method with returns. Bootstraps in regression estimation can be done through resampling on data, residual or otherwise. In this paper, the bootstrap was performed by resampling the wavelet koe-fission from the residual. To find the best wavelet thresholding regression estimator using the bootstrap method, if the bootstrap sample is generated M times, the Mean Square Error (MSE) will be obtained as much as M. From the number of M MSEs, the minimum MSE is selected. This MSE-minimizing estimator is the best estimator of the M resampling bootstrap.

Root Mean Square Error

The RMSE conformance test indicator is an error indicator based on the total square of the deviation between the model results and the observation results, which can be defined as an equation.

$$R = \sqrt{\sum_{i=1}^{N}\sum_{d=1}^{N}\left[\frac{(T_{ij}-T_{ij})^2}{N(N-1)}\right]} \quad \text{for } i \neq d$$

N = number of rows or columns of the matrix

T_{ij} T_{ij} = Matrix cell values from the model and observations

Some studies use standard deviations from divergences that can be defined as equations

$$S = \sqrt{\sum_{i=1}^{N}\sum_{d=1}^{N}\left[\frac{(T_{ij}-T_{ij})^2}{N(N-1)-1}\right]} \quad \text{for } i \neq d =$$

From the equation above, it can be seen that the greater the value of N, the RMSE value will be approximately equal to the elementary school value. The %RMSE indicator compared 2 MATs with different cell counts.

The plot in Figure 2 shows that the FOREX time series is non-stationary. It is transformed using natural logarithms, and then differencing is applied.

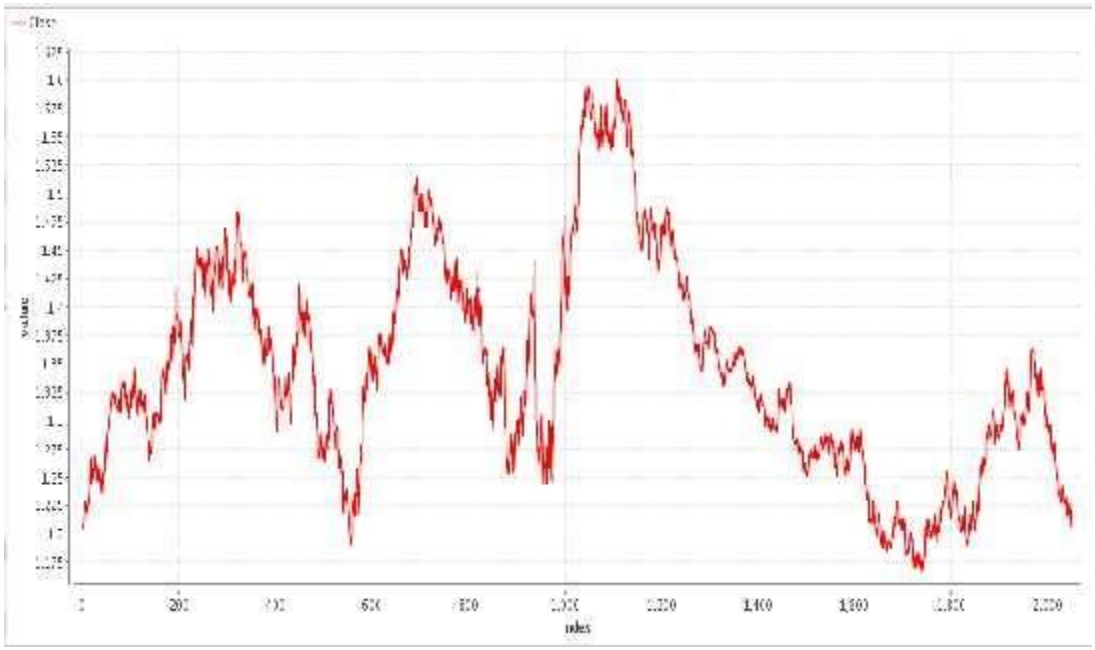

Figure 2. The movement pattern of the FOREX data used (value "Close")

Multiple linear regression fittings

Several linear regression (MLR) models developed and tested with the same set of data are used to create ANNs. The regression equation developed is referred to as a trained model. This model is then validated with the same data set used to test the JST model, thus making comparable results, and is referred to as the validated model.

CONCLUSION

1. The value that is close to the series value is the value by testing on the support vector machine algorithm, with the parameter For the RMSE value that is close to the value of "0" obtained from the measurement results on the SVM Algorithm on the RBF kernel (radial base function) with a value of "gamma" $\gamma = 100$ with the value of RMSE = 0.000, and SE = 0.000. with prediction accuracy error = 0.976
2. The evaluation results showed good accuracy, with strong correlation and low Mean Absolute Percent Error (MAPE) values. In addition, the model was tested on historical data, and significant profits were generated based on its predictions. According to the findings obtained from the assessment, predicting stocks using the moving average and linear regression methods can help investors earn profits and reduce risk.

BIBLIOGRAPHY

Categories of forecasting methods. (n.d.).*en.wikipedia.org/forecasting*.

Chen, R. (2007). Using SVM with Financial Statement Analysis for Prediction of Stocks, *7*(4), 63–72.

Collins, M., Avenue, P., Room, A., Park, F., & Schapire, R. E. (2000). Logistic Regression, AdaBoost and Bregman Distances. *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, *13*, 1–26.

Cuaresma, J. C. (2010). Modeling and Predicting the EUR/USD Exchange Rate : The Role of Nonlinear Adjustments to Purchasing Power Parity, 64–76.

Farrell, M. T., & Correa, A. (2007). Gaussian Process Regression Models for Predicting Stock Trends. *Mit. Media Edu*, 1–9.

Fitting a trend : least-squares. (n.d.).*en.wikipedia.org/Trend Estimation*.

Fox, J. (2010). Appendices to Applied Regression Analysis, Generalized Linear Models, and Related Methods, Second Edition.

From, F. (n.d.). Polynomial kernels, 2–4.

Hanias, M. P. (2008). Time Series Prediction of Dollar \ Euro Exchange Rate Index.

Hardiningsih, P. (2001). *The Influence of Fundamental Factors and Economic Risk on the Return of Company Shares on the Jakarta Stock Exchange*. Diponegoro University. http://europa.eu/legislation_summaries/economic_and_monetary_affairs/introducing_euro_practical_aspects/l25007_en.htm. *International Research Journal of Finance and Economics*, *15*(15), 232–239.

Linear prediction. (n.d.).*en.wikipedia.org/linear prediction*.

Multinomial, P., Multilevel, P., Semiparametric, N., Quantile, R., Principal, I., & Errorsvariables, L. S. (n.d.). Linear regression. *en.wikipedia.org/Linear regression Regression*.

Suparti, A. M. and A. R. (2007). Wavelet regression estimation thresholding with the bootstrap method. *Vol*, *Journal of Mathematics*, *10* (2), 43–50.