Analysis of stock market predictor variables using linear regression

Ramaswamy Seethalakshmi
SASTRA University
15 PUBLICATIONS **34** CITATIONS

International Journal of Pure and Applied Mathematics
Volume 119 No. 15 2018, 369-378
ISSN: 1314-3395 (on-line version)
url: http://www.acadpubl.eu/hub/
Special Issue
http://www.acadpubl.eu/hub/

Analysis of stock market predictor variables using Linear Regression

R.Seethalakshmi
School of Humanities and Sciences, SASTRA Deemed to be University, India

Abstract

Technological advancement increases the study on stock and share market industry. Decision making is enhanced by various statistical and machine learning algorithms. Enormous research work have been concentrated on the feature prediction of stock prices based on historical prices and volume. Performance measures are analyzed in this work with S&P 500 Index using statistical methods in R environment. Results obtained in this study are superior than the existing methods. The conventional methods for financial market analysis is based on linear regression. This paper focuses on best independent variables to predict the closing value of the stock market. This study is used to determine specific factors which are providing most impact on prediction of closing price.

Keywords: Stock market, Closing price, S&P 500 Index, Linear Regression, AIC

1. Introduction

History has revealed that stock prices and other resources is an essential part of the important forces of economic activity, and can control or be a pointer of communal mood. In a financial system where the stock market on an increase is measured to be a flourishing economy. Often the stock market is measured the principal pointer of a country's financial power and progress. Stock market research is required in order to make a smart speculation result. Stock market research is necessary if one wants to earn a major return on stocks. Before putting money in the stock market one should be alert of the company and its return patterns. Stock market research will help to make a decision which industry should invest in.

Stock market forecasting is the act of demanding to conclude the future price of a company stock or other financial instrument traded on an exchange. The successful forecast of a stock's future value might give up important profit. The efficient-market hypothesis suggests that stock prices reveal all currently existing information and any price changes that are not based on newly exposed information thus are intrinsically unpredictable. Stock price prediction is possible by Data mining Algorithms. Data mining can be defined as “making better use of data”. Every human being is more and more faced with uncontrollable amounts of data; hence, data mining or knowledge discovery it seems that affects all of us. It is so known as one of the key research areas. Preferably, we would like to build up techniques for “making improved use of any kind of data for any purpose”. On the other hand, we dispute that this goal is too challenging yet. Over the last three decades, more and more large amounts of historical data have been stored by electronic means and this amount is likely to continue to develop significantly in the future.

Data mining technique have been effectively revealed to produce high forecasting accurateness of movement of stock price. Now a days, as an alternative of a particular method, traders have to use various predicting methods to increase several signals and more information about the markets future. Data mining methods have been introduced for forecasting of movement indication of stock market index. Data mining techniques have a more successful act in predicting various fields such as policy, economy and engineering compared to usual statistical techniques by discovering unknown information of data .

Data mining is systematic method plan to explore data (usually large amount of data usually business or market related also recognized as “Big Data”) in search of reliable pattern and/or organized relations between variables, and then to confirm the result by applying the detected patterns to new. Stock market is very unpredictable in nature. Changes of stock prices almost instantly. Financial analysts who purchases stocks are not conscious of all factors like economic growth , inflation affecting stocks prices. They do not have idea in which stocks to spend and sell. The stock brokers can easily manipulate them. Stock prices depend on news appearing in news articles. It is not achievable for an average buyer to investigate such large amount of information . Data Mining technique can be used to deal with this problem. Data mining can automatically take out significant information from large amount of data that is disturbing the stock prices. Predicting the stocks prices precisely can be done by Artificial Neural Network (ANN). The benefit of using ANN is that it can agreement with both linear and non linear data for predicting the stock prices. Price will move up and down and the linear regression channel also experience changes as old prices fall off and new prices appear.

Trade in stock market deals the movement of money of a security or stock from a trader to a buyer. This require these two parties to have the same opinion on a price. Equities (stocks or shares) present an rights interest in a specific company. Stock market participants range from small individual stock investors to larger traders investors, who can be based wherever in the world, and may contain insurance companies or pension funds, banks and hedge funds. Their buy or sell orders may execute on their behalf by a stock exchange dealer. Stock trading volume includes the number of lots bought and sold which is express in daily basis . The more trading volume of a stock is higher, the more the stock is active. Trading volume is an appreciative to price patterns in practical testing and it's additional vital than stock price.

Stock market contribution refers to the number of agents who buy and sell equity backed securities either directly or indirectly in a financial trade. Participants are normally subdivided into three distinct sectors; households, institutions, and foreign traders. Direct participation occur when any of the above entities buys or sells securities on its own behalf on a trade. Indirect participation happens when an institutional investor exchanges a stock on behalf of an individual or household. Indirect investment takes in the form of pooled investment accounts, retirement accounts, and other managed financial accounts.

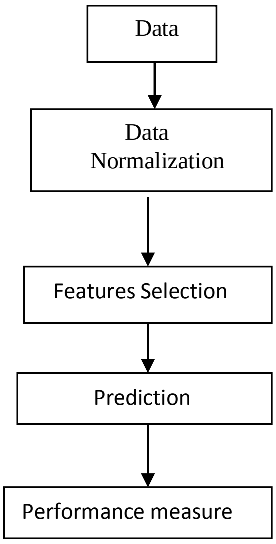
2. Literature Review

Box–Jenkins[1] used Time series analysis for forecasting and control. White [2,3,4] used Neural Networks for stock market forecasting of IBM daily stock returns. Following this, a range of studies reported on the efficacy of different learning algorithms and forecasting methods using ANN. Henry [5] used ARIMA model, to predict the daily close and morning open price,. But all these predictable methods had troubles when non linearity exists in time series. Chiang et al.[6] have used ANN to predict the end-of-year net asset value of mutual funds. Kim and Han [7] found that the complex dimensionality and hidden noise of the stock market data make it difficult to re calculate the ANN parameters. Romahi and Shen [8] also found that ANN rarely suffers from over fitting problem. They developed a budding rule based expert system and obtained a method which is used to predict financial market behaviour. There were also hybridization models successfully used to predict financial behaviour. The disadvantage was prerequisite of expert knowledge.

Logistic regression(LR), which is useful for predicting the occurrence or non occurrence of a quality or outcome based on values of a set of forecaster variables, is a multivariate analysis model [Lee, 2004][9]. In the area of banking, corporate finance and investments, LR applications have frequently been used. For the default-prediction model, many researchers used Multivariate discriminant analysis (MDA). Öğüt and Aktaş [2009][10] found that data-mining techniques (ANN and SVM) are better suitable to detect stock-price manipulation than multi variate statistical techniques such as discriminate analysis or LR, because the performances of data-mining techniques in terms of classification accuracy are better than those of multivariate techniques. They proposed a new binary classification method for predicting corporate failure based on genetic algorithm, and proposed to validate its prediction power through empirical analysis. Minand Jeong [2009][11] compared prediction accuracy with other methods such as multi-discriminant analysis, logistic regression, decision tree, and artificial neural network and showed that the binary classification method they proposed can serve as a promising alternative to existing methods for bankruptcy prediction.

3. Proposed Algorithm

Fig 1 : Work flow Diagram



3.1 Data Description

The S&P 500index provides the market value of 500 stocks. Though it includes stocks from different industries S&P 500 has some distinguishing characteristics such as sensitivity, predictability, scalability ,to name a few. We have used all the indices for future calculations. A sample of data obtained from [www.yahoofinance.com] is shown in table. Open, High, Low, Close, Volume, Adj Volume are considered as attributes for this study

Table : 1 Sample data

Date	Open	High	Low	Close	Volume	Adj Close
6/9/1998	1115.72	1119.92	1111.31	1118.41	5.64E+08	1118.41
6/8/1998	1113.86	1119.7	1113.31	1115.72	5.43E+08	1115.72
6/5/1998	1095.1	1113.88	1094.83	1113.86	5.58E+08	1113.86
6/4/1998	1082.73	1095.93	1078.1	1094.83	5.77E+08	1094.83
6/3/1998	1093.22	1097.43	1081.09	1082.73	5.84E+08	1082.73
6/2/1998	1090.98	1098.71	1089.67	1093.22	5.91E+08	1093.22
6/1/1998	1090.82	1097.85	1084.22	1090.98	5.38E+08	1090.98

3.2 Data Normalization

There are a couple important details to note about the way the data must be preprocessed in order to be fit into regression models. Firstly, dates are normally represented as strings of the format ”YYYY-MM- DD” when it comes to database storage. This format must be converted to a single integer in order to be used as a column in the feature matrix. This is done by using the date’s ordinal value.

3.3 Features

Stock market close price is an important piece of information that is very useful for every short-term trader. The close prices are **very important**, especially for swing traders and position traders. It also has implications for practical day trading in many day trading systems. The stock market close price level provides very important information about the general mood of investors. It tells a lot about the thinking of big investors that allocate large amount of money into the stock market for their asset management purposes.

3.4 Regression

If the goal is prediction, or forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of y and X values. After developing such a model, if an additional value of X is then given without its accompanying value of y, the fitted model can be used to make a prediction of the value of *y*. Regression predicts a numerical value [12]. Regression performs operations on a dataset where the target values have been defined already. And the result can be extended by adding new information [13]. The relations which regression establishes between predictor and target values can make a pattern. This pattern can be used on other datasets which their target values are not known. Therefore the data needed for regression are 2 part, first section for defining model and the other for testing model. In this section we choose linear regression for our analysis. First, we divide the data into two parts of training and testing. Then we use the training section for starting analysis and defining the model.

Model 1: It includes all the available features . The features are described below.

Opening price

The opening price is the value that each share has when the S&P 500 stock exchange opens for trading. The opening price gives a good indication of where the stock will move during the day. Since the Stock exchange can be likened with an auction market i.e. buyers and sellers meet to make deals with the highest bidder, the opening price does not have to be the same as the last day’s closing price.

Highest/lowest price of the day

The highest and the lowest price of the day are taken the day before and gives an indication of how much the shares usually move during a day and how this in the end will affect the closing price. It also shows the general cyclical movement for each share.

An adjusted closing price is a stock's closing price on any given day of trading that has been amended to include any distributions and corporate actions that occurred at any time prior to the next day's open. The adjusted closing price is often used when examining historical_returns or performing a detailed analysis on historical returns.

Volume

Volume is one of the most basic and beneficial concepts to understand when trading stocks. Volume is defined as, **“the number of shares or contracts traded in a security or an entire market during a given period of time.”**

Model 2: In model 2, two features such as volume and adjusted close are omitted

3.5 Performance measure

Models are evaluated through standard performance measures and its description is given in Table2.

Table 2 Performance measure description

Name	Formula	Remark
R²	$R^2 = 1 - \frac{SS\,Error}{SS\,Total} = \frac{SS\,Total - SS\,Error}{SS\,Total} = \frac{SS\,regression}{SS\,Total}$ $SSE = \sum_{i=1}^{n} e_i^{\ 2} = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$ $SS\,Total = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 =$ $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 = SSE + SS\,regression$	The percent of the variance in the dependent variable that can be explained by all of the independent variables taken together
Adjusted R²	$R^2{}_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$	A version of R-Squared that has been adjusted for the number of predictors in the model. R-Squared tends to over estimate the strength of the association especially if the model has more than one independent variable.
F Test	***F* = test statistics for ANOVA for Regression= *MSR/MSE,*** where MSR=Mean Square Regression, MSE = Mean Square Error The null and alternative hypotheses for simple linear regression for the F-test statistic are	can be used in Simple Linear Regression to assess the overall fit of

	H_o: b_1=0; where b_1 is the coefficient for *x* (i.e. the slope of *x*) H_a: b_1 is not 0	the model.
Akaike information criterion AIC	AIC = 2 k – 2ln(L) where k is the number of estimated parameters and L be the maximum value of the likelihood function.	measure of the relative quality of statistical models for a given set of data.
Bayesian Information Criterion BIC	$\mathrm{BIC} = -2LogLikelihood + k\ln(n)$ where k is the number of estimated parameters in the model and n is the number of observations in the data set.	information-based criteria that assess model fit

4. Results and Analysis

Six attributes of the Stock data set is considered for Model 1. After identifying the most insignificant attribute and eliminate it from data set gives model 2.

This work compares the model 1 and model 2 using AIC, BIC and R^2 values. The model outputs are tabulated in Table 3. Model 1 and Model 2 are in Figure 2 and 3 respectively.

Fig 2: Model 1 with all Features

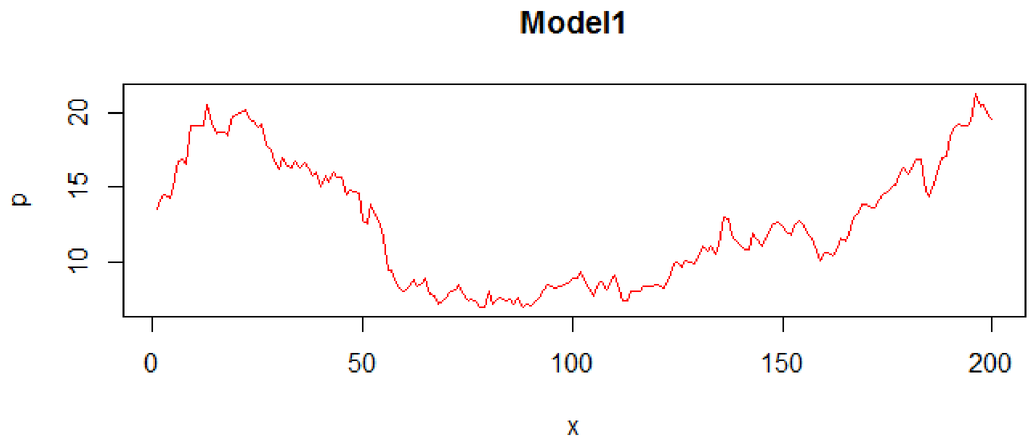


Fig 3 : Model 2 with Selected Features

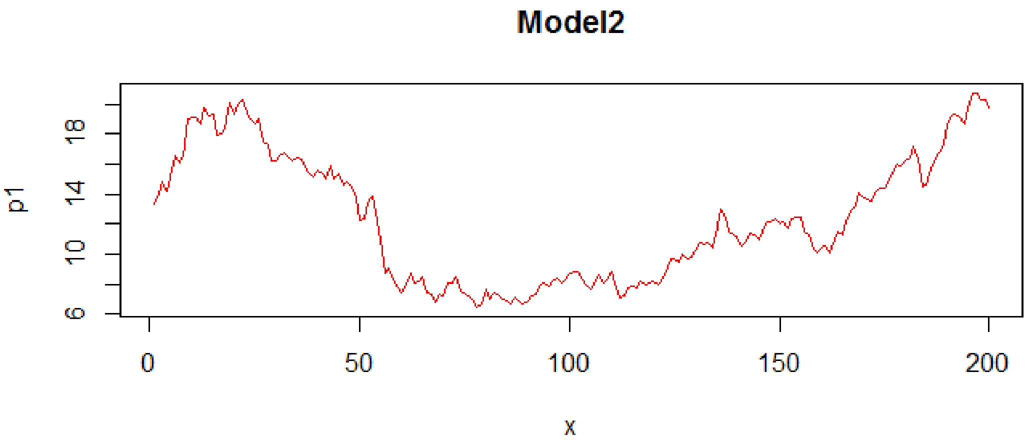


Table 3 Performance Measures

Measures	Model 1	Model 2
Multiple R^2	0.997	0.992
adjR^2	0.997	0.992
AIC	-215.9031	3639.1538
BIC	-183.1108	3667.2615

Model 1 includes all (open,low,high,volume,adjclose) attributes and obtained AIC value as: -215.9031 .Model 2 with the three(open,low,high) attributes and its AIC value is 3639.1538.

Model1 AIC,BIC values are lesser than model2.Hence this work concludes Model 1 is the best model; hence we need to include volume and adjclose attributes for predicting the close price.

5. Conclusion

Model 1 with all features fitted with R^2 value 0.997. This indicates open, high, low, volume and adj close are essential for predicting closing value accurately. Model 2 with open, high and low predict close value fitted with R^2 value 0.992 . This indicates prediction of close value is not affected with adj close. This study reveals with open, high, low and volume itself enough for finding approximate prediction of close value.

References

[1] G. E. P. Box and G. M. Jenkins, Time series analysis: forecasting and control. San Fransisco, CA: Holden- Day, 1976. to Bull and Bear Markets, President, Global Financial Data, Inc.

[2] Halbert White, “Economic prediction using neural networks: the case of IBM daily stock

returns," Department of Economics, University of California, San Diego.

[3] H.White, "Economic prediction using neural networks: the case of IBM daily stock returns," In Proceedings of the second IEEE annual conference on neural networks., II, pp. 451–458,1988.

[4] H.White, Learning in artificial neural networks: a statistical perspective, Neural Computation., Vol. **1** , pp. 425–464,1989.

[5] Henry M. K. Mok, "Causality of interest rate, exchange rate and stock prices at stock market open and close in Hong Kong," Asia Pacific Journal of Management, Vol. **10** (2), pp. 123–143,1993.

[6] W.C. Chiang, T. L. Urban and G. W. Baldridge, "A neural network approach to mutual fund net asset value forecasting," Omega International Journal of Management Science., Vol. 24 (2), pp. 205–215,1996.

[7] K.J. Kim and I. Han, "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index, Expert Systems with Applications," Vol.19,pp. 125–132,2000.

[8] Y. Romahi and Q. Shen, "Dynamic financial forecasting with automatically induced fuzzy associations," In Proceedings of the 9th international conference on fuzzy systems., pp. 493–498,2000.

[9] Lee, S. 2004. Application of likelihood ratio and logistic regression models to landslide susceptibility mapping using GIS. Environmental Management 34(2), 223-232.

[10] Öğüt, Hulisi, et al. 2009, Detecting stock-price manipulation in an emerging market: The case of Turkey, Expert Systems with Applications 36(9), 11944-11949.

[11] Min, Jae H., and Chulwoo Jeong. 2009. A binary classification method for bankruptcy prediction, Expert Systems with Applications 36(3), 5256-5263.

[12] Gharehchopogh, F. S., & Khalifehlou, Z. A. (2012). A New Approach in Software Cost Estimation Using Regression Based Classifier. AWER Procedia Information Technology and Computer Science, Vol: 2, pp. 252-256.

[13] Draper, N. R., Smith, H., & Pownell, E. (1966). *Applied regression analysis* (Vol. 3). New York: Wiley