Stock Analysis Method based on Multiple Linear Regression

Yuheng Liu

Abstract. In order to research the application of multivariate linear regression in stock price analysis, seven independent variables were selected to estimate in this paper. The regression coefficient was calculated by SPSS and the statistical test was carried out. And finally, the model was applied to the closing price prediction, and the prediction error was within acceptable range.

Keywords: stock, stock market, multiple linear regression, investment method

Tob Regul Sci.™ 2021;7(5-1):4222-4228
DOI:doi.org/10.18001/TRS.7.5.1.198

RESEARCH BACKGROUND

The stock was formerly known as the east India company, which was bought and sold by the Dutch in 1602, and the formal stock market originated in the United States. Stock is a kind of negotiable securities, it is the certificate that share company issues to investor when raising capital, the droit that represents holder to share company.

Nowadays, with the continuous development of society, stock has become one of the important ways for citizens to conduct financial management, and its proportion in citizens' income and expenditure is also increasing. It has attracted the attention of many investors with its characteristics of high risk and high return. Therefore, seeking an effective investment method to increase returns and reduce risks has become the primary goal of almost all investors.

The largest of the world's 60 largest exchanges is the New York stock exchange, with a market capitalization of $21.876 trillion, accounting for 26.3% of the global market capitalization of listed companies (2017 data). It certainly occupies the center of global trade. The NASDAQ stock exchange of the United States, the Tokyo stock exchange of Japan and the Shanghai stock exchange of China followed suit. Statistically, there is no doubt that the United States plays a decisive role in the running of the world economy.

Since the birth of the stock market, many economists and statisticians have devoted themselves to the search for mathematical models and analytical methods. In recent years, with the development of computer science and technology, many scholars have applied data mining, machine learning and neural network to stock analysis, which greatly improves the accuracy of stock analysis and has become an important method in the field of financial analysis.

LITERATURE REVIEW

Research Methods of Stock Market in China

Chen Yushan (2007) used ICA (independent component analysis) to analyze stock price returns and came to the conclusion that most stock price returns obey a kind of super gaussian distribution [1]. Wang Dongxiu (2013) proposed an improved Apriori algorithm for the shortcomings of the Apriori algorithm. He used the algorithm to analyze three stocks and concluded that there was a strong correlation between the rise and fall of two of them [2]. Gao Xueying (2017) analyzed the impact of macroeconomic factors on stocks by using the gray correlation method and concluded that the consumer price index had the largest impact on stock price fluctuations [3].

Research Methods Based on Deep Learning or Neural Network

Yao Peifu (2006) analyzed and studied the BP algorithm and put forward the corresponding improvement measures:

Accelerated Iterative Convergence Formula

Modification Function

Adaptive Adjustment of Learning Coefficient

Through the above improvement, the disadvantages of BP algorithm can be alleviated, and the prediction accuracy can be improved [4].

*Yuheng Liu, School of Computer Science and Technology, Xidian University, Xi'an, 710126 China ,*Corresponding author: Yuheng Liu*

The stock price prediction model proposed by Liu Lei (2017) is based on the limited Boltzmann Machine. He sampled and modeled the stock of Gree electric appliances, and the accuracy of the final result is high (63.41%) [5].

Research Methods of Stock Market in Foreign Countries

(1) Traditional data mining or statistical methods

Syeda Shabnam Hasan (2017) analyzed three companies -ACI, Beximco and GP by using SVM (support vector machine) and GPR (gaussian process regression) and concluded that SVM was superior to GPR [6]. Han Lock Siew (2012) used the method of regression analysis and selected six factors affecting the stock market for analysis. They found that the accuracy of the results improved after converting the input data [7].

(2) Research methods based on deep learning or neural network

Raymond S.T.L. ee (2004) adopted the model based on RBF neural network to analyze the Hong Kong stock market from 1990 to 1999, and obtained more accurate results compared with other analysis methods, and he realized different predictions for the long term and short term [8]. Yunnus YETIS (2014) adopted the method based on artificial neural network, selected the stock market data of NASDAQ from 2012 to 2013, trained the model and obtained accurate results. When the index closes above 3, 000, its error rate is less than 2% [9].

RESEARCH MEANING

Stock analysis can be divided into two types. One is the overall evaluation and prediction of the stock market. This kind of analysis is beneficial for the government to make correct policies and also for enterprises to make appropriate development strategies. The other is to predict the rise and fall of a single stock, whose main purpose is to provide investors with a reference to avoid risk.

From the perspective of the government, the stock market' s advantage lies in its ability to concentrate idle social funds widely and contribute to national economic growth. The stock market can empower the role of the market, break geographical restrictions, and improve the resource allocation efficiency. Besides, it can also promote enterprises to constantly improve their business management model, and ultimately promote economic development.

From the perspective of individuals, stock is a method to handle personal investment: the volatility of stock prices allows the potential to make margins. The stock market offers investors with a high-return financial management channel, widens the investment options range, and enhances the investment liquidity. Investors can get their money back by cashing in their shares.

For enterprises, the stock market facilitates funds raising for joint-stock enterprises to meet their operational needs. Also, issuing shares is a way to avoid risks. There will always be operational risks, whatever kind of an enterprise is, especially for the innovative ones. The newly opened up markets, uncertain market prospects, unfamiliar technologies, etc., all these factors would cause the existence of business risks. Issuing shares is a good way to avoid risks. Even the investment fails, the loss is apportioned among each shareholder to an acceptable extent.

However, being uncertain is the nature of the stock market. The misprediction of the falling or rising trend of stocks may result in the inappropriate policies by the government, money loss of shareholders, and disruptions to operations in enterprises. The stock market' s fundamental role is inevitably affected by its instability. At the same time, the stock market stability and macroeconomic stability are closely related, the stock market concussion may have triggered the financial crisis, which prompted financial market liquidity shortage and pressure, higher interest rates and the outbreak of the financial crisis and its influence, is bound to conduct through the financial markets to the real economy, which in turn will lead to economic growth markets such as [10]. Therefore, to accurately predict the result of the stock market is of vital significance.

MODEL CONSTRUCTION

Analysis Mode

In this paper, I use F statistic, R2, modified R2 and correlation coefficient as variables to measure the fitting degree of multivariate linear regression.

① F statistic

F statistic is used to test the joint hypothesis of regression coefficient.

When the joint hypothesis has β_1 = 0 and β_2 = 0 for two constraints:

F statistics use the following formula to make t1 and t2, the two t statistics, combined, i.e.:

$$F = \frac{1}{2}\left(\frac{t_1^{\,2} + \mathrm{t}_2^{\,2} - 2\hat{\rho}_{t1,t2}\, t_1 t_2}{1 - 2\hat{\rho}_{t1,t2}}\right) \qquad (4.1)$$

Where $\hat{\rho}_{t1,t2}$ represents the correlation coefficient estimator of two t statistics.

In order to understand the F statistic, we can first assume that the t statistic is not correlated, then the correlation coefficient of t1 and t2 is 0, then there is

$$F=\frac{1}{2}\,({t_1}^2+{t_2}^2) \qquad (4.2)$$

Under the null hypothesis, t_1 and t_2 are independent standard normal random variables, so F follows $F_{2,\infty}$ distribution. Under the alternative hypothesis that β_1 is not 0 or β_2 is not 0, t_1^2 or t_2^2 is very large, so the test results reject the null hypothesis.

In general, t statistics are correlated, so the F statistic formula in (4.1) modifies this correlation. This makes the F statistic subject to A distribution in $F_{2,\infty}$ large sample whether the t statistic is relevant or not.

When there are q constraints in the joint hypothesis:

The global multivariate regression model is

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\cdots+\beta_kX_{ki}+u_i, i=1,2,\cdot\cdot$$

Define the following vectors and matrices:

$$Y=\begin{bmatrix}Y_1\\Y_2\\\vdots\\Y_n\end{bmatrix} \qquad U=\begin{bmatrix}u_1\\u_2\\\vdots\\u_n\end{bmatrix}$$

$$X=\begin{bmatrix}1 & X_{11} & \cdots & X_{k1}\\1 & X_{12} & \cdots & X_{k2}\\\vdots & \vdots & \ddots & \vdots\\1 & X_{1n} & \cdots & X_{kn}\end{bmatrix}=\begin{bmatrix}X'_1\\X'_2\\\vdots\\X'_n\end{bmatrix} \beta=\begin{bmatrix}\beta_0\\\beta_1\\\vdots\\\beta_k\end{bmatrix}$$

Y represents the n vector of n*1 observations of the dependent variable.

X represents the n*(k+1) dimensional matrix of n observations composed of k+1 regression variables.

(k+1)*1 dimension column vector is the ith observation of k+1 regression variable, and X_i' is the transpose of X_i.

U represents the n*1 dimension vector of n error terms.

β represents the (k+1)* 1 dimension vector composed of k+1 unknown regression coefficients.

In this paper, we consider the joint assumption of q constraints, where q⩽k+1. Each of these q constraints involves one or more regression coefficients. Then the joint primitive hypothesis can be represented by matrix symbols as

$$R\beta = r$$

Where R is a nonrandom matrix of order q* (k 1) with full rank, and r is a nonrandom vector of order q*1. The number of columns of R is q, equal to the number of constraints in the original hypothesis.

this moment

$$F=\frac{(R\hat{\beta}-r)'[R\hat{\sum}_{\hat{\beta}}R']^{-1}(R\hat{\beta}-r)}{q}$$

When the null hypothesis is true, the sampling distribution of the F statistic in a large sample is $F_{q,\infty}$ distribution.

② R2 and modified R2

Regression R2 is the proportion of Yi sample that can be explained (or predicted) by regression variables.

$$R^2=\frac{ESS}{TSS}=1-\frac{SSR}{TSS}$$

The explained sum of squares is ESS= $\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2$, the total sum of squares is TSS= $\sum_{i=1}^{n}(Y_i-\bar{Y})^2$, and the sum of squared residuals is SSR= $\sum_{i=1}^{n}\hat{u}_i^2$.

In multiple regression, unless the estimated value of the regression variable coefficient increases is exactly 0, the number of regression variables R2 will increase as long as the number of regression variables increases. Consider the case where you start with the first regression variable and then add the second regression variable. When OLS was used to estimate a model with two regression variables, OLS found the coefficient value that minimized the sum of squares of residuals. If the coefficient of the new regression variable chosen by OLS happens to be 0, the SSR remains the same whether the second variable is added or not. However, if OLS chooses a non-zero value, this value must reduce SSR. In practice, it is very rare that the estimated value of coefficient is exactly 0, so the introduction of new regression variables will generally reduce SSR. This means that R2 usually increases with the addition of a new regression variable. In this case, R2 overestimates the effect of the regression fitting data. One way to fix it is to reduce or reduce R2 by some factor, and fix R2 is one of them.

Correction R² $\overline{R}^2$ is a correction form of R², that is, it does not necessarily increase after adding new regression variables. Its expression is

$$\overline{R}^2 = 1 - \frac{\text{n-1}}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{s_{\hat{u}}^2}{s_Y^2}$$

The difference between this formula and R squared is that the ratio of SSR to TSS is multiplied by factor $\frac{\text{n-1}}{n-k-1}$. This indicates that $\overline{R}^2$ is 1 minus the ratio of the sample variance of OLS residuals to the sample variance of Y.

About $\overline{R}^2$:

First: $\frac{\text{n-1}}{n-k-1}$ will be more than 1 forever, so $\overline{R}^2$ will be less than R² forever.

Second: Adding a regression variable will make $\overline{R}^2$ have two opposite trends. On the one hand, the decrease of SSR makes $\overline{R}^2$ increase, and on the other hand, factor $\frac{\text{n-1}}{n-k-1}$ increases. So, whether $\overline{R}^2$ is going to go up or down depends on how strong these two things are.

Third, $\overline{R}^2$ may be negative. $\overline{R}^2$ is negative when the sum of squared residuals reduced by all regression variables is too small to offset the factor $\frac{\text{n-1}}{n-k-1}$.

R^2 and $\overline{R}^2$ are not specified:

Whether the variables included in the regression are statistically significant.

Whether the regression variable is the real cause of the change of the dependent variable.

Whether there is a missing variable bias.

Whether the most appropriate set of regression variables has been selected.

③ Correlation coefficient

Before introducing correlation coefficient, introduce covariance first.

Covariance

Covariance is a measure of the degree to which two variables change at the same time. The covariance of X and Y is the expected value of $(X-\mu_X)(Y-\mu_Y)$, where μ_X and μ_Y represent the mean of X and Y, respectively. Let's use $\text{cov}(X,Y)$ or σ_{XY} for the covariance. If X takes one value and Y takes k values, the covariance formula can be expressed as:

$$\text{cov}(X,Y) = \sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)] = \sum_{i=1}^{k}\sum_{j=1}^{l}(x_j - \mu_X$$

Correlation Coefficient

Correlation coefficients include simple correlation coefficient, complex correlation coefficient and partial correlation coefficient.

Simple Correlation Coefficient

The simple correlation coefficient between X and Y is the covariance of X and Y divided by their standard deviation:

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

If $\text{corr}(X,Y) = 0$, then X and Y are unrelated.

DATA SOURCES AND FEATURES

Data Sources

The dependent variables in this paper are from Yahoo Finance. The daily opening price, closing price, the highest price of the day, the lowest price of the day and the trading volume of Facebook from 2012 to May 2018 are collected.

The average monthly active user (MAU) count and net profit data in this paper are derived from Facebook’ s quarterly results.

The standard & poor’ s (S&P 500) index in this paper is from Yahoo Finance.

The conference board consumer confidence index (CBCCI) data in this paper come from Oriental Wealth Network (OWN).

The ISM non-manufacturing index (ISMNMI) data in this paper come from OWN.

The unemployment rate data in this paper come from the Bureau of Labor Statistics (BLS).

The trade account data in this paper are derived from OWN.

Data Features

The data adopted in this research have been published at varying frequencies. For example, the closing price of a stock varies from day to day, yet the unemployment rate is published monthly by the US BLS, while the average monthly number of people living or revenue is released quarterly in Facebook’ s financial report. This main feature of the data has certain impact on the accuracy of the analysis.

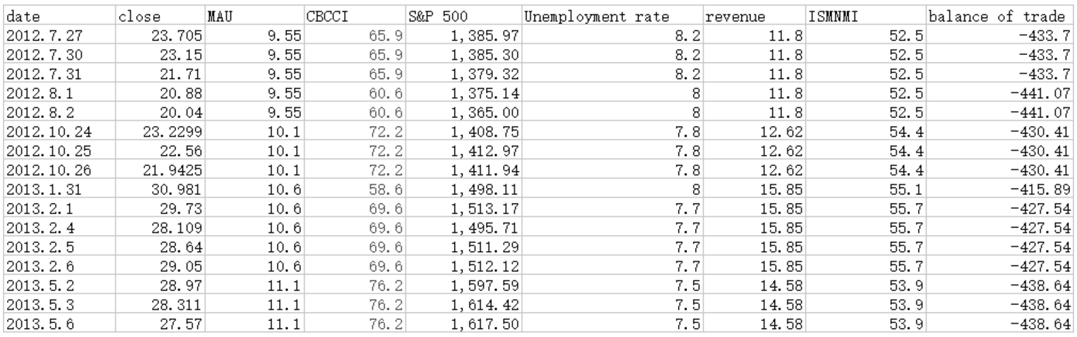
Data Pre-processing

During data collecting, the daily opening price, closing price, the highest price of the day, the lowest price of the day and the trading volume of Facebook were obtained. Then, its daily closing price was selected as the dependent variable.

Next, in order to handle the varying frequencies of data releasing, the raw data were processed as follows: The stock closing price of one week from the date of each financial report of Facebook company was selected as the dependent variable, and the S&P 500 on the corresponding date was selected as the independent variable. Meanwhile, the monthly CBCCI, ISMNMI, unemployment rate, and trade account were used as independent variables for all data in the current month instead of daily data. The average monthly number of people living and the revenue were retained as the independent variables of all the data for each quarter. In that way, problems caused by insufficient data could be alleviated.

The processed sample data are listed as below:

Table 1.



date	close	MAU	CBCCI	S&P 500	Unemployment rate	revenue	ISMNMI	balance of trade
2012.7.27	23.705	9.55	65.9	1,385.97	8.2	11.8	52.5	-433.7
2012.7.30	23.15	9.55	65.9	1,385.30	8.2	11.8	52.5	-433.7
2012.7.31	21.71	9.55	65.9	1,379.32	8.2	11.8	52.5	-433.7
2012.8.1	20.88	9.55	60.6	1,375.14	8	11.8	52.5	-441.07
2012.8.2	20.04	9.55	60.6	1,365.00	8	11.8	52.5	-441.07
2012.10.24	23.2299	10.1	72.2	1,408.75	7.8	12.62	54.4	-430.41
2012.10.25	22.56	10.1	72.2	1,412.97	7.8	12.62	54.4	-430.41
2012.10.26	21.9425	10.1	72.2	1,411.94	7.8	12.62	54.4	-430.41
2013.1.31	30.981	10.6	58.6	1,498.11	8	15.85	55.1	-415.89
2013.2.1	29.73	10.6	69.6	1,513.17	7.7	15.85	55.7	-427.54
2013.2.4	28.109	10.6	69.6	1,495.71	7.7	15.85	55.7	-427.54
2013.2.5	28.64	10.6	69.6	1,511.29	7.7	15.85	55.7	-427.54
2013.2.6	29.05	10.6	69.6	1,512.12	7.7	15.85	55.7	-427.54
2013.5.2	28.97	11.1	76.2	1,597.59	7.5	14.58	53.9	-438.64
2013.5.3	28.311	11.1	76.2	1,614.42	7.5	14.58	53.9	-438.64
2013.5.6	27.57	11.1	76.2	1,617.50	7.5	14.58	53.9	-438.64

Data Analysis

Let the closing price be Y with an average monthly wage of X1, the conference board’ s consumer confidence index be X2, the S&P 500 be X3, the unemployment rate be X4, the revenue be X5, the ISM non-manufacturing index be X6, and the trade book be X7:

$$\mathrm{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7$$

, where β_0 is the constant term.

After calculation $\beta_0 = 80.574$, $\beta_1 = 6.204$, $\beta_2 = -0.295$, $\beta_3 = 0.034$, $\beta_4 = -10.009$, $\beta_5 = 0.282$, $\beta_6 = -1.925$, $\beta_7 = -0.066$, namely:

$$\mathrm{Y} = 80.574 + 6.204X_1 - 0.295X_2 + 0.034X_3 - 10.009X_4 + 0.282X_5 - 1.925X_6 - 0.066X_7$$

In this article, three variables are used to measure the model:

①F statistic

It can be obtained through calculation that the value of F statistic is 1154.841, which is much higher than the F value (2.96) at the significance level of 0.01. Therefore, we can say that only in a few cases, the conclusion obtained from this expression is wrong, that is, the model has passed the F test.

②R2 and modified R2

Through calculation, we can get that the corresponding R2 in this paper is 0.987, and the corrected R2 is 0.986. That is, the change in Y of 0.987(or 0.986) in the equation is caused by the independent variable. This value is very close to 1, which indicates that the regression variable can better predict Y.

③Correlation coefficient

Calculation of Simple Correlation Coefficient:

The simple correlation coefficient of X1 and Y is -0.989. The simple correlation coefficient of X2 and Y is 0.955. The simple correlation coefficient of X3 and Y is 0.961. The simple correlation coefficient of X4 and Y is -0.951. The simple correlation coefficient of X5 and Y is 0.965. The simple correlation coefficient of X6 and Y is 0.409. The simple correlation coefficient of X7 and Y is -0.612. It can be concluded from the analysis that, apart from the weak correlation between X6 and Y, other dependent variables and independent variables are strongly correlated.

Perspective

According to the regression equation, the stock price of Facebook on May 21, 2018 is predicted to be $184.90912 when the relevant data of that day is substituted. In fact, Facebook closed the day at $184.49, within an acceptable margin of error.

Shortcomings

In this analysis, there are the following shortcomings:

In the choice of variables, the independent variable ISM non-manufacturing index and dependent variable stock closing price is not strong correlation. The correlation coefficient is 0.409. This has had an effect on the accuracy of the results.

In the collection of data, some data are not published, such as the daily number of active Facebook. Some data are published once a month, for example, the ratio of unemployment. Therefore, the author can only use the same month or quarter of the data to replace the corresponding place in each piece of data. This will lead to certain errors.

The stock market is a very complex system, and it is difficult to guarantee the practical effect of using multiple linear regression to model the stock market. Although the prediction results are good in this paper, the accuracy of sudden events, such as war, disease, or natural disasters, will be greatly affected.

CONCLUSION AND PROSPECT

Conclusion

The rise and fall of stocks are affected by many factors. In this paper, multiple linear regression is adopted to predict the stock price within a certain range, but the accuracy of prediction is limited. The data selected by the author is limited and the indicators are too few, which is one of the reasons for the inaccurate prediction results.

Prospect

Since the analysis method adopted in this paper is multiple linear regression, the prediction ability of the stock system is limited. In the following research, a novel method, such as neural network, will be used to analyze stock data. At the same time, the analysis results are compared with this method to explore the advantages and disadvantages of the new method and the traditional method in the field of stock analysis.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this paper.

Acknowledgement

First, I’ d like to extend my grateful thanks to my thesis instructor, Mr. Zhao Kaiyang. Mr. Zhao has made some instructive suggestions and recommendations on the research direction of my thesis. During the writing of this paper, he carefully advised on my difficulties, clarified my confusion, and provided many practical suggestions for improvement. Meanwhile, I would like to thank my classmates. We have been learning from each other, together through a perfect unforgettable period. Moreover, I would like to thank my friends, who brought me strong support and inspiration in the preparation of the paper. I would also thank to the authors of the references. It is their works that led me to a good starting point for my research. Thanks to my encouraging and supportive family, teachers, and friends, I was able to finish this paper.

Funding

China National Social Science Foundation Project--Study on the Economic Effect and Optimizing Mechanism of China ’ s Credit Mismatch (15BJY172)

REFERENCE

1. CHEN Yushan, XI Bin. “Application of independent component analysis to analysis of stock market” [J].
2. Computer Engineering and Design, 2007 (06): 1473-1476.
3. WANG Dongxiu, HU Yingchun, LI Hui. “The Application and Research of Improved Apriori Algorithm in stock Analysis” [J]. Bulletin of Science and Technology, 2013, 29(03): 125- 128
4. GAO Xueying, GUO Ronghua. “Analysis of the influence of macroeconomic factors on stock price” [J]. Journal of Hebei Software Institute, 2017, 19(04): 58-61.
5. YAO Peifu, XU Dadan. “Research of BP neural network in stock prediction” [J]. Guangdong Automation & Information Engineering, 2006(01):7-9.
6. LIU Lei. Stock trend prediction based on deep learning[D]. YUNNAN UNIVERSITY OF FINANCE AND ECONOMICS, 2017.
7. Hasan, Syeda Shabnam, et al. “Improved Stock Price Prediction by Integrating Data Mining Algorithms and Technical Indicators: A Case Study on Dhaka Stock Exchange.” (2017).
8. Siew, Han Lock, and M.J.Nordin. “Regression techniques for the prediction of stock price trend.” International Conference on Statistics in Science IEEE, 2012.
9. Lee, Raymond S.T. “iJADE stock advisor: an intelligent agent based stock prediction system using hybrid RBF recurrent network.” Systems Man & Cybernetics Part A Systems &Humans IEEE Transactions on 34.3 (2004): 421-428.
10. Yetis, Yunus, H. Kaplan, and M. Jamshidi. “Stock market prediction by using artificial neural network.” World Automation Congress IEEE, 2014.
11. YANG Xiuchang, LI Guohua. “Stability of Stock Market in China” [J]. Productivity Research, 2009 (21):123-124.