Forecasting Daily Stock Market Return with Multiple Linear Regression

Shengxuan Chen

May 12, 2020

Project Advisor: Dr. Zhong

Abstract

The purpose of this project is to use data mining and big data analytic techniques to forecast daily stock market return with multiple linear regression. Using mathematical and statistical models to analyze the stock market is important and challenging. The accuracy of the final results relies on the quality of the input data and the validity of the methodology. In the report, within 5-year period, the data regarding eleven financial and economical features are observed and recorded on each trading day. After preprocessing the raw data with statistical method, we use the multiple linear regression to predict the daily return of the S&P 500 Index ETF (SPY). A model selection procedure is also completed to find the most parsimonious forecasting model.

Keywords: data mining, daily stock market return, multiple linear regression, model selection

1 Introduction

The application of mathematical and statistical models are used to analyze finances and the stock market is a universal and important field. Because the stock market is affected by numerous factors, it is a difficult task for researchers. The stock markets are complex systems, and many factors act and interact the stock market in nonlinear and dynamic way [2]. The stock market is affected by many interrelated factors, which include: 1) economic variables; 2) industry specific variables; 3) company specific variables; 4) psychological variables of investors; 5) political variables [10]. This paper will use the financial and economical features to be the factors to forecast the daily return of the stock market.

Since the application of mathematical and statistical models involve massive data, we must explore data mining and big data analytic techniques. Data mining is a process to explore and analyze a large quantities of data by computer and get a significant model [3].

Simple linear regression discovers the relationship between two variables which can be presented by a straight line. The variable which actively changes and causes another variable to change is called the independent variable, and another variable which changes with the independent variable is the dependent variable. The independent variable represents certain factor on which the dependent variables relies. Since the stock markets are affected by various factors, the multiple linear regression, which can consider multiple variables influencing the stock market, is a highly established statistical technique in stock market analyses [8].

Since the multiple linear regression has numerous independent variables, the calculation is complex. Thus, statistical software is used in practical applications generally. MATLAB is a useful software to analyze data in mathematics and statistics. In this project, MATLAB is used to analyze big data in order to forecast the daily stock market return after dividing the data to three parts: training part, validation part and testing part [10].

It is also known that the quality of the input data plays a critical role on the efficiency of data mining and big data analysis tools, including the multiple linear regression. Thus, in the project, various statistical methods are used to preprocess the raw data. Therefore, this paper processes these data to make the characteristic of data representing obvious. This paper preprocesses data by adjusting and cleaning the outlier of the data.

Due to some extreme situation or the errors in observation, recording and calculation, one or several values in the data may be significantly different from other values. The non-normal data is called outlier. In the project, the outlier may affect the results, so eliminating the outlier is necessary. In statistics, quartile can define arrange, and any data lying outside these defined bounds can be consider an outlier.

2 Data description

There are so many factors that may affect the daily return of the Standard & Poor's 500 Index Exchange-Traded Fund(SPY), so choosing some representative factors is easier to research. Those factors can be divided into 11 main groups, SPY return in current and three previous days, relative difference in percentage of SPY return, exponential moving averages of SPY return, Treasury bill rates, certificate of deposit rates, financial and economical indicators, the term and default spreads, exchange rate between US Dollar and other currencies, the return of the other seven world major indices, SPY trading volume and the return of the eight big companies in S&P 500 [10]. This paper choose eleven factors from the each part.

The data include the daily return of the S&P 500 Index ETF and the values of those eleven

factors in 1304 trading days from June 1, 2013 to May 31, 2018. The daily return of the S&P 500 Index ETF are used to be dependent variable and the values of those eleven financial and economical features are used to be independent variables. This project research the multiple linear regression between the S&P 500 Index ETF and the eleven factors.

The eleven factors are the return of the SPDR S&P 500 ETF(SPY) in day t, the 5-day relative difference in percentage of the SPY, the 10-day exponential moving average of the SPY, 6-month T-bill rate, secondary market, business days, discount basis, average rate on 6-month negotiable certificates of deposit, quoted on an investment basis, relative change in the gold price, term spread between T6 and T1, relative change in the exchange rate between US dollar and Chinese Yuan, Hang Seng index return in day t, relative change in the trading volume of S&P 500 index and Apple Inc stock return in day t.

Table 1: The 11 financial and economical features of the raw data

Group	Name	Description	Source/Calculation
	Data_SPY	Trading dates considered	finance.yahoo.com
SPY return in current and three previous days	SPYt	The return of the SPDR S&P 500 ETF(SPY) in day t	finance.yahoo.com/(p(t)-p(t-1))/p(t-1)
Relative difference in percentage of the SPY return	RDP5	The 5-day relative difference in percentage of the SPY	(p(t)-p(t-5))/p(t-5)×100
Expoential moving averages of the SPY return	EMA10	The 10-day exponential moving average of the SPY	p(t)×(2/(10+1))+EMA10(t-1)×(1-2/(10+1))
T-bill rates	T6	6-month T-bill rate, secondary market, business days, discount basis	https://fred.stlouisfed.org/series/DTB6
Certificate of deposit rates	CD6	Average rate on 6-month negotiable certificates of deposit(secondary market), quoted on an inverstment basis	H. 15 Release-Federal Reserve Board of Governors
Financial and economical indicators	Gold	Relative change in the price of the gold price	https://www.usagold.com/reference/goldprices/2013.html
The term and default spreads	TE6	Term spread between T6 and T1	TE6=T6-T1
Exchange rate between USD and four other currencies	USD_CNY	Relative change in the exchange rate between US dollar and Chinese Yuan(Renminbi)	https://www.investing.com/currencies/usd-cny-historical-data
The return of the other seven world major indices	HIS	Hang Seng index return in day t	https://finance.yahoo.com/quote/
SPY trading volume	V	Relative change in the trading volumn of S&P 500 index(SPY)	https://finance.yahoo.com/quote/SPY?p=SPY&.tsrc=fin-srch
The return of the eight big companies in S&P 500	AAPL	Apple Inc stock return in day t	https://finance.yahoo.com/quote/AAPL?p=AAPL

3 Data preprocessing

During the process of data collection, recording or calculation, some data may be missing or error. If analyzing the data without preprocessing, the result will be completely different from the expected data. Thus, propressing data is an important step. In this project, the preprocessing include filling the missing data, fixing the outliers and plotting data.

3.1 Missing data

Due to the data include the value of SPY and 11 features in 5 years, which is enormous, there are a lot of missing data in the collected data. This is a very common problem in statistics. Generally, the solution of this problem are delete or interpolation. The delete method does not need extra calculation but the errors are usually large. Thus, this project choose interpolation to fill in the missing data problem.

This project choose spline method, piecewise cubic spline interpolation, to solve this problem.

This method will give a function f defined on a interval $[a, b]$ and a set of nodes such that $a = x_0 < x_1 < ... < x_n = b$. Then a cubic spline interpolant S for f is a function that satisfies the following conditions:

(a) $S_n(x)$ is a cubic spline interpolant on the subinterval $[x_j, x_{j+1}]$, and x_j means the value of SPY in day j;

(b) $S_n(x_j) = f(x_j)$ and $S_n(x_{j+1}) = f(x_{j+1})$;

(c) $S_{n+1}(x_{j+1}) = S_n(x_{j+1})$;

(d) $S'_{n+1}(x_{j+1}) = S'_n(x_{j+1})$;

(e) $S''_{n+1}(x_{j+1}) = S''_n(x_{j+1})$;

(f) One of the following sets of boundary conditions is satisfied: $S''(x_0) = S''(x_n) = 0$ or $S'(x_o) = f'(x_0)$.

This method involves four constants, so there is sufficient flexibility in the cubic spline procedure to ensure that the interpolant not only is continuously differentiable but also has a continuous second derivative [1].

3.2 Outliers

Since the data involved this project include the value of SPY and 11 features in 5 years, it include some outliers. Therefore, clearing those outliers is a important step before data analysis. This project use quartile and interquartile range to find those outliers. The reason which cause outliers may be the extreme values or some error values due to experiment inaccuracy.

To detect the outliers in the whole data, this project will defind a interval and the data which do not lie in the interval are outliers.

There are many method to defind the interval to detect the outlier. Quartile is one type of statistics method. It defines three quartiles. The first quartile Q_1 is the middle number of the first half of the rank-ordered data set, the second quartile Q_2 is the middle number of the all rank-ordered data set and the third quartile Q_3 is the middle number of the second half of the rank-ordered data set.

The interquartile range (IQR) is the difference between the value of 75th percentitles and the value of 25th percentitles. That means $IQR = Q_3 - Q_1$. Then define a interval that the upper boundary is $Q_3 + 1.5IQR$ and the lower boundary is $Q_1 - 1.5IQR$ [9]. Those data which do not lie in the intervals are outliers.

Since the number of those outliers of this data which this project detect are more than 100, this project will replace those outliers rather than delete. Thus, this project will use clip method to replace those outliers. The clip method is that use the lower boundary to replace the the outliers which is smaller than the lower boundary and use the upper boundary to replace the outliers which are larger than upper boundary.

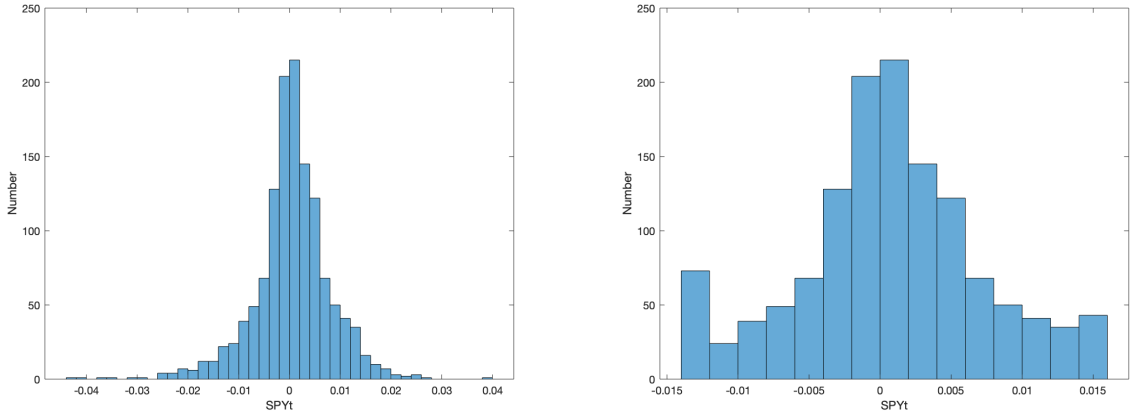
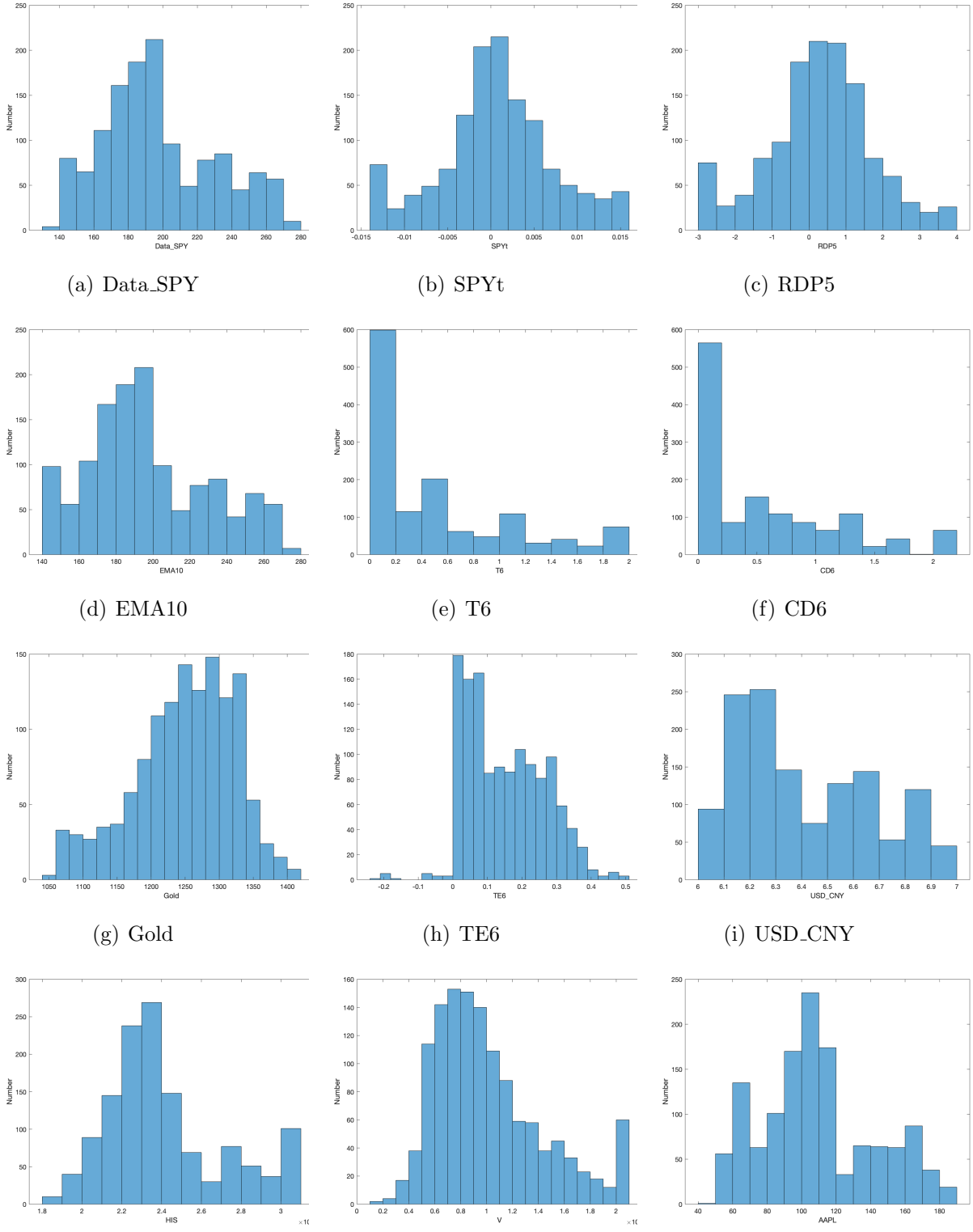


Figure 1: Histogram of SPY current return(on the left); Histogram of adjusted SPY current return(on the right).



(a) Data_SPY (b) SPYt (c) RDP5

(d) EMA10 (e) T6 (f) CD6

(g) Gold (h) TE6 (i) USD_CNY

(j) HIS (k) V (l) AAPL

Figure 2: The histograms of SPY and the eleven features after preprocessing

3.3 Plot data

Histogram, also known as mass distribution chart, is a kind of statistical report chart. The data distribution is represented by a series of vertical stripes or line segments of varying heights. Using histogram can easy represent a large amount of data and intuitively indicate the shape of the distribution.

Thus, After processing the data, this project graph the histogram of daily return of SPY before and after data preprocessing, which can observe the effect of clearing outliers. For example, Figure 1 shows the histograms of factor SPYt, before and after data preprocessing. The horizontal axis of the histogram that after preprocessing is smaller which represents the data after preprocessing are more concentrated. Also, Figure 2 shows the histograms of SPY and the eleven features after preprocessing.

4 Data Analysis

4.1 Data split

In data mining, when performing predictive analysis, the data generally divided into two parts, training data and testing data. In this project, the training data is the first 75% of the data and the testing data is the last 25% of the data. The training data is used to build a mathematical model, and the testing data is used to test the mathematical model.

4.2 Multiple Linear Regression

Linear regression is a linear modeling that reflect the relationship between dependent variable and independent variable. For more than one independent variable, the linear regression is multiple linear regression.

The mathematical modeling of multiple linear regression is

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon \tag{1}$$

where y is the dependent variable; $x_1, x_2, ..., x_n$ are the n independent variables; β_0 is the intercept; $\beta_1, \beta_2, ..., \beta_n$ are the partial regression coefficient and ε is the random error and it distribute N(0,σ^2). [7]

In this project, y is the daily value of SPY in day $t + 1$ and each independent variable is the value in day t.

4.3 Model Selection

When analyzing mathematical models, the different factors has different significant effects. The final model should be the optimal model which is the simplest and most effective, so some insignificant factors need to be removed.

This project use the forward stepwise regression to choose the optimal model. The stepwise regression is a method which choose the independent variable. The method take the all of factors into the linear regression model one by one. When one new factor is chosen into the model, test the accuracy of new model to choose the optimal model. [5] [7]

4.3.1 Forward Selection

P-value is the probability when assuming that null hypothesis is correct. Thus the p-value is smaller, then the factor is more effective. Thus, the variables are added in a empty model one by one from the small p-value to large. Repeat it until there is no significant variable can be added into the model.

The final forward selection model include the return of the SPDR S&P 500 ETF(SPY), the 5-day relative difference in percentage of the SPY, the 10-day exponential moving average of the SPY and relative change in the trading volume of S&P 500 index, which is

$$y = -2.4194 + 66.473x_1 + 0.81059x_2 + 1.0027x_3 - 6.2414^{-9}x_{10} + \varepsilon. \tag{2}$$

Then test the model by the test data and the mean squared error is 4.29654.

4.3.2 Backward Selection

The direction of backward method is inverse of forward method. It starts with the model include all variables and remove the variable which has the largest P-Value, then repeat it until the loss of variable does not cause insignificant.

The final forward selection model include the return of the SPDR S&P 500 ETF(SPY), the 5-day relative difference in percentage of the SPY, the 10-day exponential moving average of the SPY, the 10-day exponential moving average of the SPY, relative change in the price of the gold price, term spread between T6 and T1, relative change in the trading volume of S&P 500 index and Apple Inc stock return in day t, which is

$$\begin{aligned} y = 0.16567 + 64.404x_1 + 0.78415x_2 + 0.96632x_3 + 2.558x_4 + 0.0031115x_6 - 3.353x_7 \\ -6.8475^{-9}x_{10} + 0.030302x_{11}\varepsilon. \end{aligned} \tag{3}$$

Then test the model by the test data and the the mean squared error is 14.17727.

4.3.3 Bidirectional Selection

Bidirectional selection is method which combine forward and backward. It starts with a random model, and add significant variable or remove insignificant variable. Repeat it until there is no insignificant variable in the model and no significant variable be removed from the model. [4]

If the model starts with the average rate on 6-month negotiable certificates of deposit, the bidirectional selection move in the 10-day exponential moving average of the SPY, the 5-day relative difference in percentage of the SPY, the return of the SPDR S&P 500 ETF, the SPY trading volume and term spread between T6 and T1, then move out the average rate on 6-month negotiable certificates of deposit and move in 6-month T-bill rate. Therefore, the final model is

$$y = 1.4115 + 65.788x_1 + 0.80173x_2 + 0.99758x_3 + 1.6054x_4 - 3.5932x_7 - 6.7924^{-9}x_{10} + \varepsilon. \tag{4}$$

Then test the model by the test data and the the mean squared error is 10.62305.

5 Results

The forward selection model has the smallest mean squared error, so the forward selection model is optimal model. Therefore, the significant factors include the return of the SPDR S&P 500 ETF(SPY), the 5-day relative difference in percentage of the SPY, the 10-day exponential moving average of the SPY and relative change in the trading volume of S&P 500 index.

The forecasting multiple linear regression model is

$$y = -2.4194 + 66.473x_1 + 0.81059x_2 + 1.0027x_3 - 6.2414^{-9}x_{10} + \varepsilon. \tag{5}$$

6 Conclusion and Discussions

This project build a multiple linear regression model with some financial and economical features, which can be used to forecast the value of SPY in the second day.

In this project, it use the normal linear regression to find this model. However, in the economic field, time series regression is also a very common regression method. Unlike static data, the time series of a feature comprise values changed with time. Time series data are of interest because of its pervasiveness in various areas ranging from science, engineering, business, finance, economic, health care, to government. Given a set of unlabeled time series,

it is often desirable to determine groups of similar time series. These unlabeled time series could be monitoring data collected during different periods from a particular process or from more than one process. The process could be natural, biological, business, or engineered. Works devoting to the cluster analysis of time series are relatively scant compared with those focusing on static data. However, there seems to be a trend of increased activity. [6] In the future, time series will be added into this project to improve the final model.

References

[1] Richard L Burden and J Douglas Faires. Numerical analysis, brooks. *Cole, Belmont, CA*, 1997.

[2] Andreia Dionisio, Rui Menezes, and Diana A Mendes. An econophysics approach to analyse uncertainty in financial markets: an application to the portuguese stock market. *The European Physical Journal B-Condensed Matter and Complex Systems*, 50(1-2):161–164, 2006.

[3] David Enke and Suraphan Thawornwong. The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with applications*, 29(4):927–940, 2005.

[4] Peter L Flom and David L Cassell. Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. In *NorthEast SAS Users Group Inc 20th Annual Conference*, pages 11–14, 2007.

[5] Mohammad Hossain. A note on model selection in statistics and econometrics. *Journal of Statistical Studies*, 20:59–66, 01 2000.

[6] T Warren Liao. Clustering of time series data—a survey. *Pattern recognition*, 38(11):1857–1874, 2005.

[7] Stanley L Sclove. A review of statistical model selection criteria: Application to prediction in regression, histograms, and finite mixture models. *Histograms, and Finite Mixture Models (June 30, 2011)*, 2011.

[8] Shakhla Shrut, Shah Bhavya, Shah Niket, Unadkat Vyom, and Kanani Pratik. Stock price trend prediction using multiple linear regression. *International Journal of Engineering Science Invention (IJESI)*, 7(1):5, 2018.

[9] Graham Upton and Ian Cook. *Understanding statistics*. Oxford University Press, 1996.

[10] Xiao Zhong and David Enke. A comprehensive cluster and classification mining procedure for daily stock market return forecasting. *Neurocomputing*, 267:152–168, 2017.