***Final Project***

**1.Abstract:**

First, we first observed data, searched for some information, and then made some hypotheses. Then make a simple model, use the backward method to modify, so that the model is closer to the real situation. Finally, draw a conclusion, query the information and make a conclusion based on your own understanding.

**2. Introduction:**

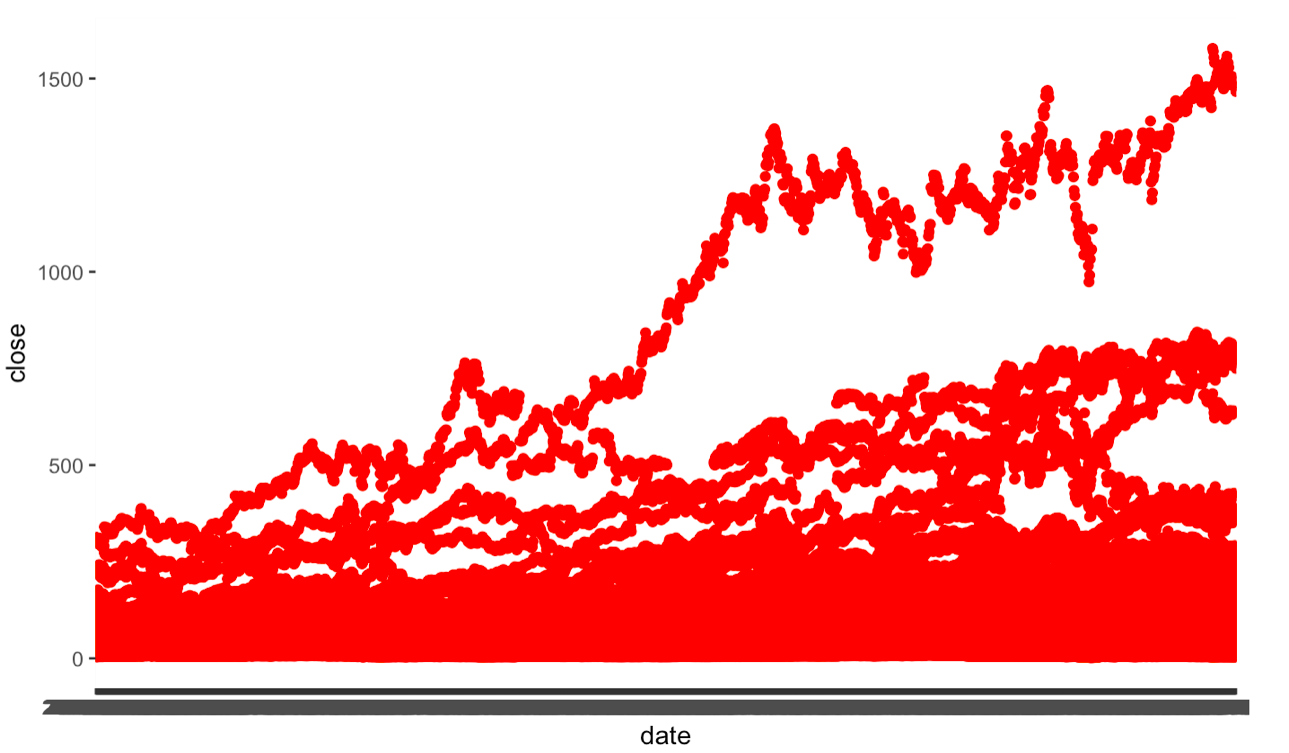
This dataset contains several parts, namely price, securities and fundamentals. These parts of the dataset provide us with S&P 500 companies historical prices with fundamental data in the New York Stock Exchange from 2010 to 2016. Also, fundamental includes Accounts, investments and liabilities, etc.

By studying these parts of the dataset, can help us better understand the impact of securities trading on time. An analysis of this part of the dataset and study the results of analysis can help us understand what may affect the factors better. And may have a positive influence for the future development. Moreover, studying the data from 2010 to 2016 can also help us know what types of listed companies have higher securities prices or what types of companies that people are interested in, and whether there is any alert or development for future listed companies under the influence of time. These above questions have become the purpose of our research on this issue.

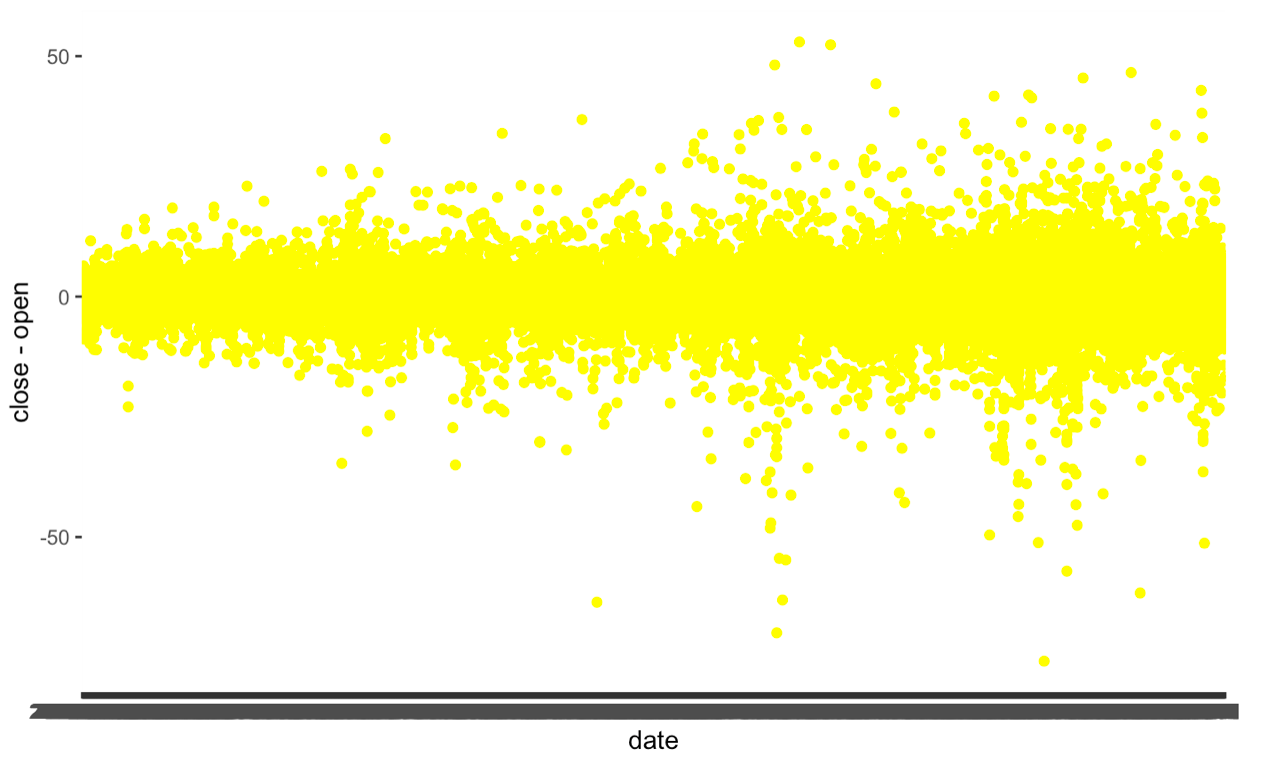
For buyers, it is also the purpose of our research to invest in what company or what type of company that can make us to receive higher returns.

**3. Data Description:**

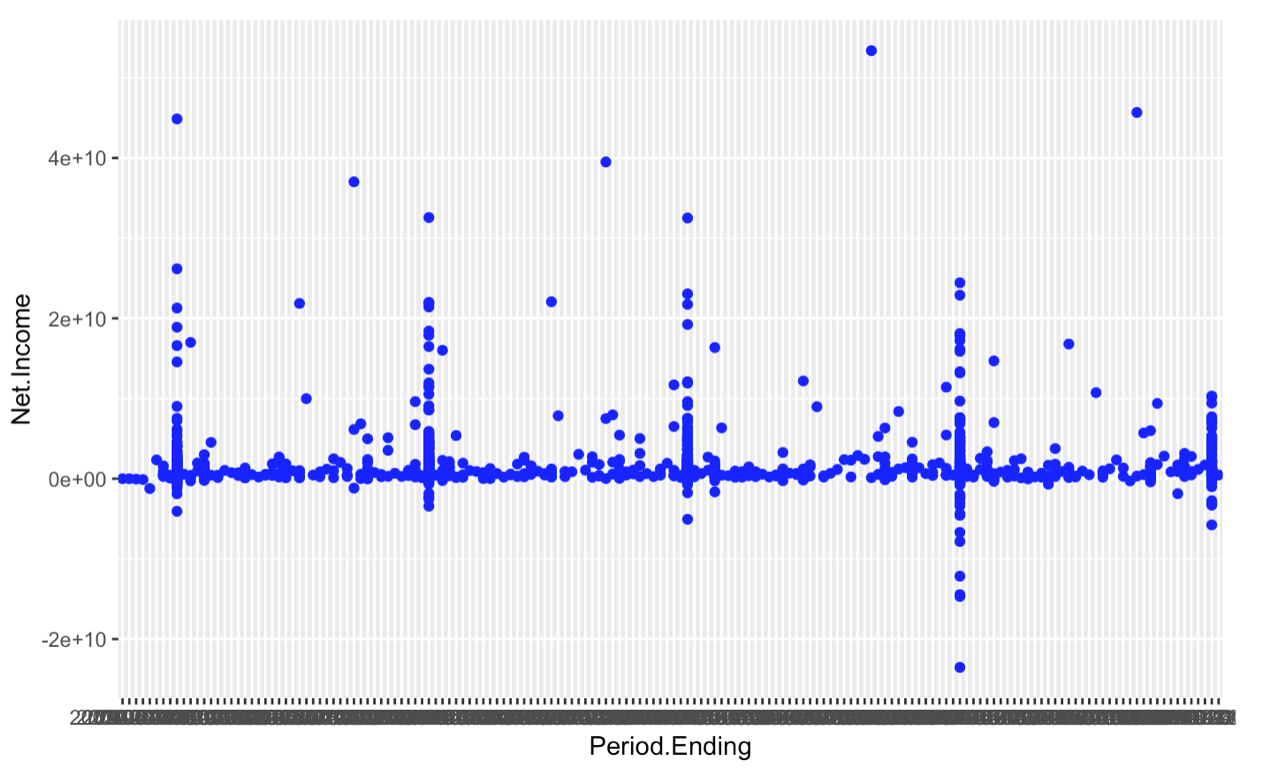
I made three pictures of the two data, fundamental and prices.split.adjusted. The first picture is the stock closing price. We can see that the points are basically rising year by year over time. We can also understand that because the economy has been developing, the price will be higher and higher. But it cannot be said that the growth rate is also increasing year by year.



The second picture is the closing price minus the opening price. Although the points are concentrated in the middle area, it can be seen that the increase in the visual effect is also increasing year by year.



The third picture is the data of Net.Income in the fundamentals data. There will be some very high and very low situations every once in a while. I don't know the reason yet to be studied.



Finally, with regard to the additional data point, we selected the original data in the data and copied a column to make the following regression. One is to make your own regression unique and not make regressions the same. The second is that because it is data itself, it also has certain accuracy.

**4. Method：**

In order to find the relationships between the companies’ financial data and their stock prices, we proposed a regression model.

4.1 Data Preprocessing

Since the goal is to probe the relation between the data of the financial situation of companies and their stock prices, we consider the variables in the file Fundamental.csv as independent variables. And the data for each company are yearly based, therefore, we calculate the average stock price for each company per year from the file prices-split-adjusted.csv.

After combining the two datasets into one, there are some missing values. And missing data are mainly from these variables: Cash Ratio, Current Ratio, Quick Ratio, For Year, Earnings Per Share, Estimated Shares Outstanding. With other information given, we calculated these values and complete the dataset. However, it is possible that denominator is 0 while computing missing values, therefore, some of the blanks are filled with infinite numbers. With no other choice, we delete these values.

Since there are 75 features in total for each company, which is very difficult for computer to process. Also, from the preprocessing part, we know that some of these variables can be derived from other variables, therefore, we first check the correlation coefficient of these variables.

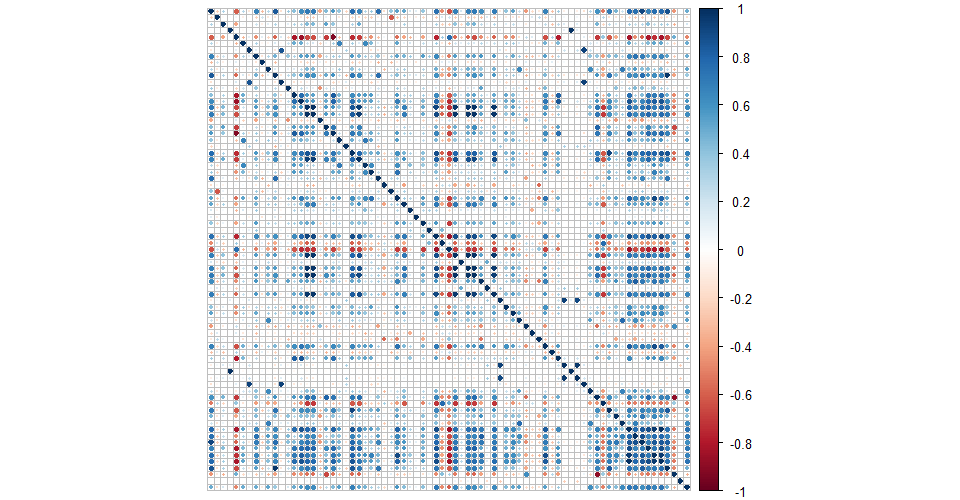
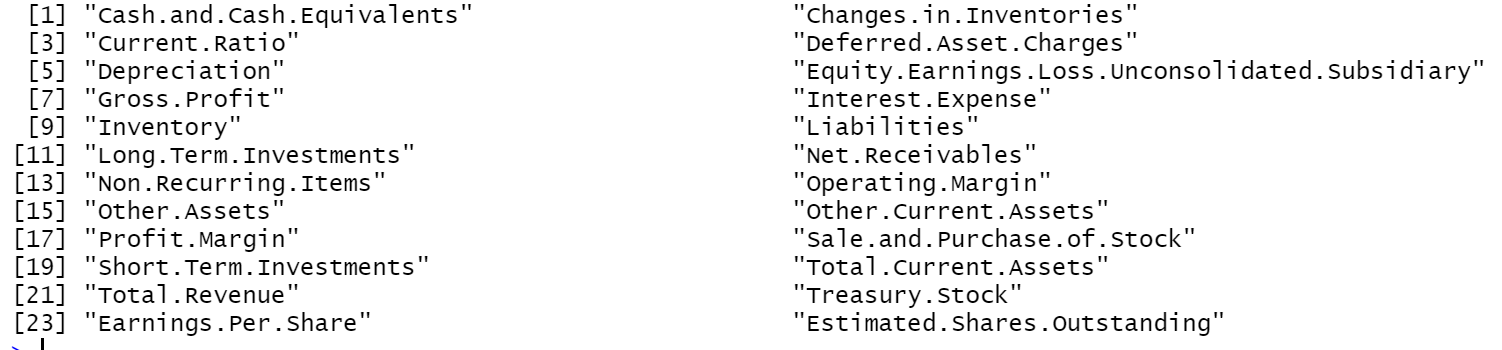
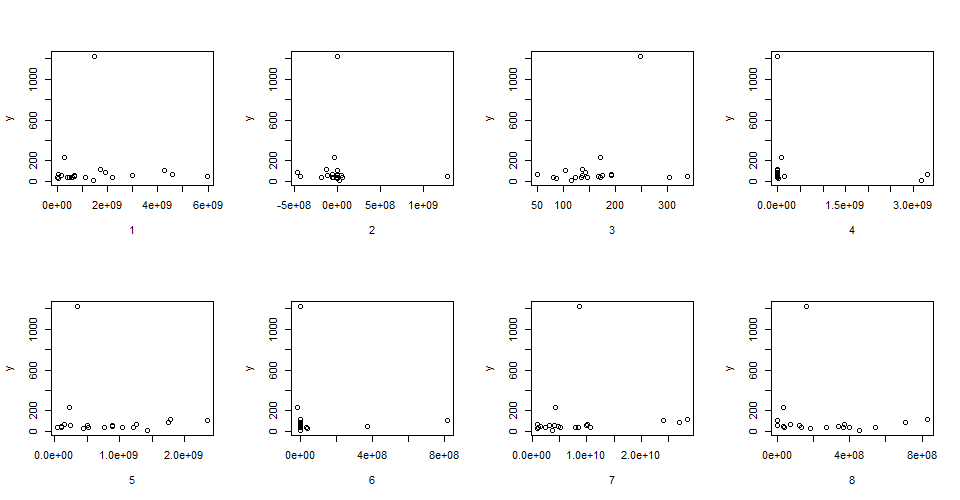


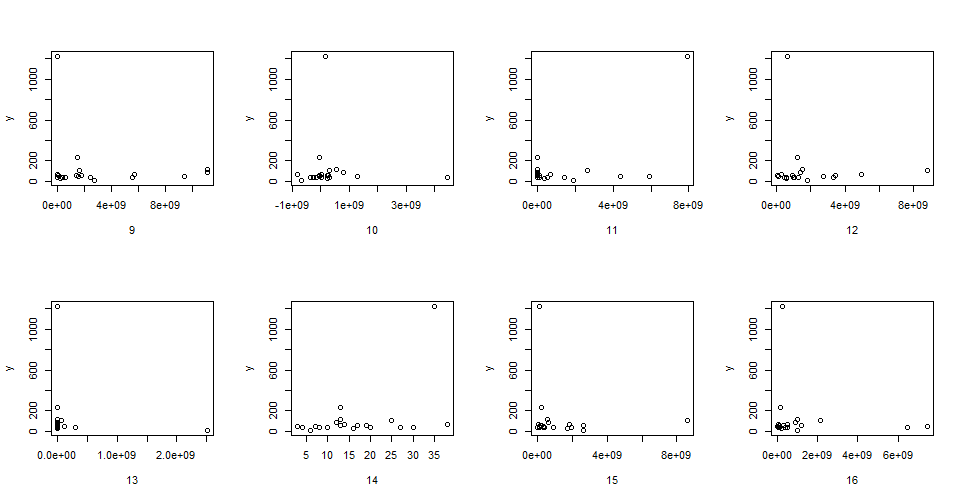
Fig 1. Correlation

From the results above, it is clear that many features are correlated with each other. Therefore, we screen out 24 variables by correlation test.



Since there are more than 1000 data points, we randomly select 20 records and draw the plot for each variable:





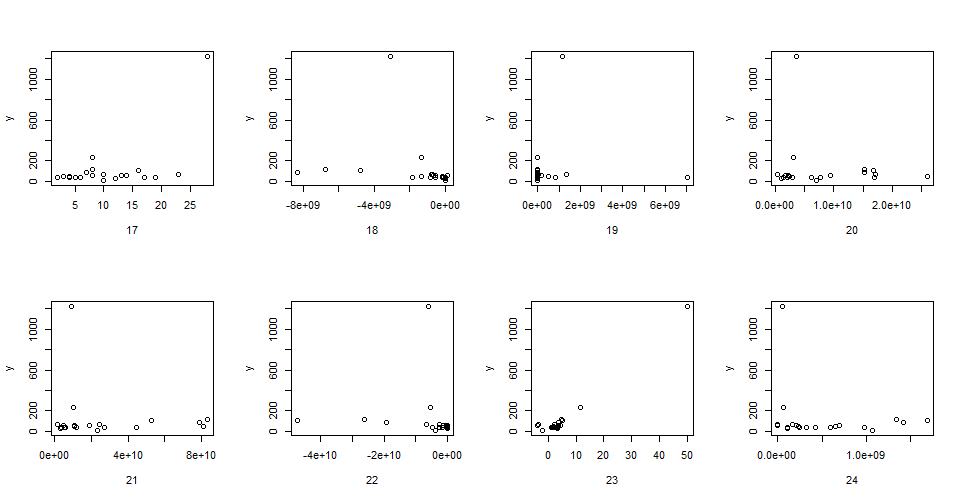


Fig 2. Scatter plot

While trying to transform the independent variable, the result does not improve much and many null and infinite values present. Thus, we use the original independent variables for regression.

4.2 All-subsets regression

After correlation test, there are still 24 variables left, with is too many for regression. To reduce the dimension, we first perform all-subsets regression to find out the best subsets of these variables. In this step, we calculate obtain 15 subsets. By calculating , Mallows Cp and BIC, the best 5 models are selected.

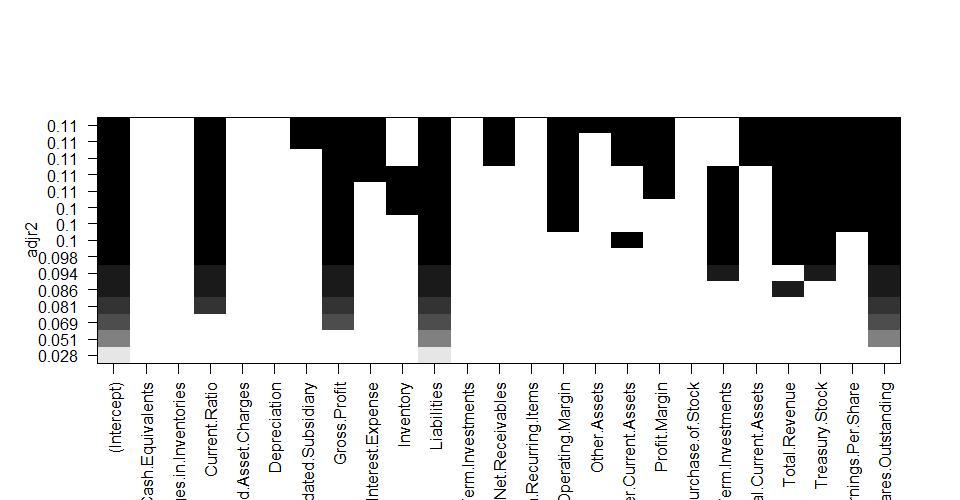


Fig 3.

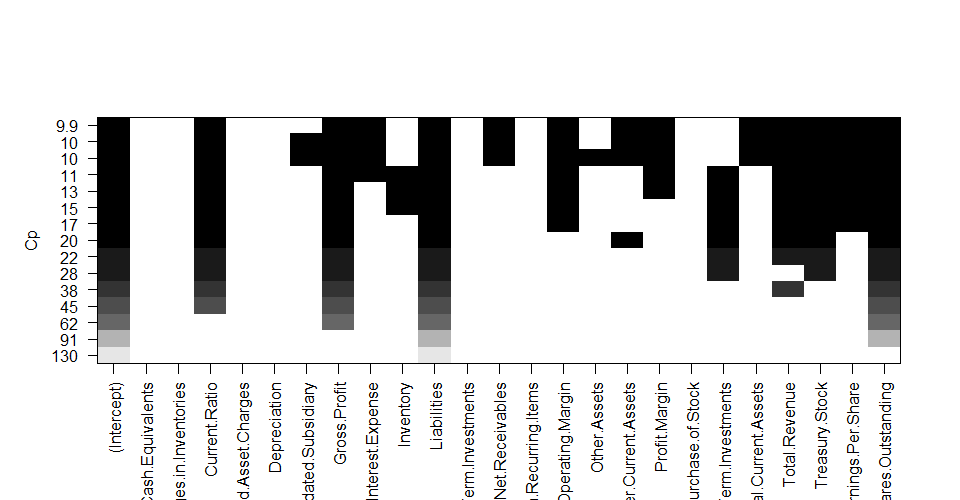


Fig 4. Mallows Cp

And the result of the best model attained by all-subsets regression is shown as follow:

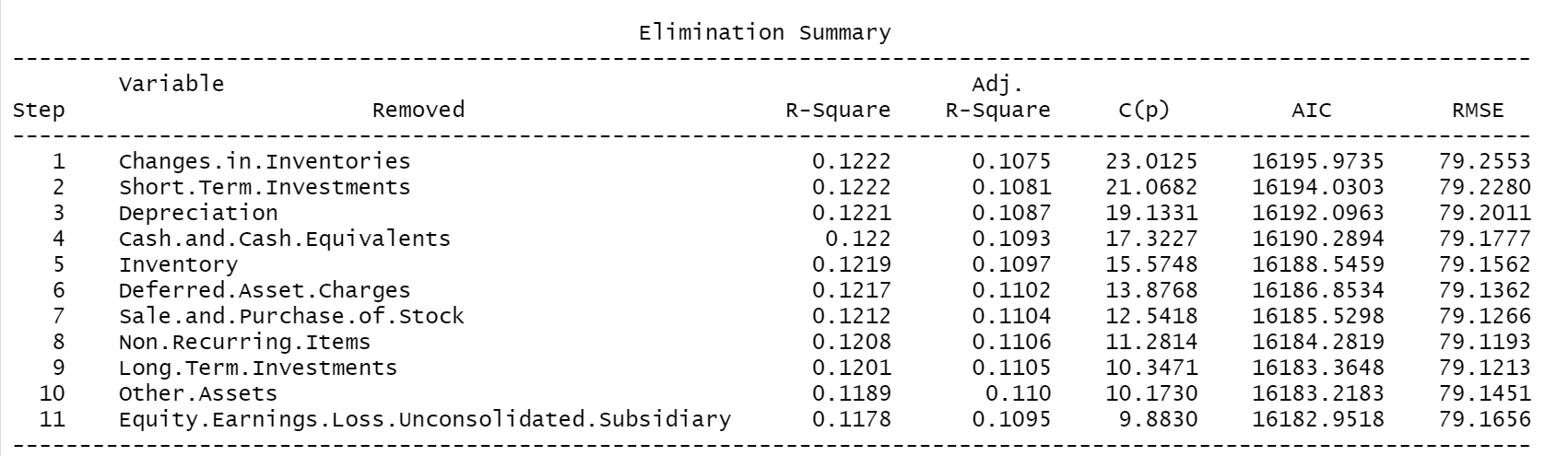
Table 1. The best model attained by all-subsets regression

|  |  |
| --- | --- |
| Variable | Coefficient |
| Current Ratio | 6.292e-02 |
| Gross Profit | 1.948e-09 |
| Liabilities | 1.381e-08 |
| Other Current Assets | -3.755e-09 |
| Short Term Investments | 1.582e-09 |
| Total Revenue | -2.730e-10 |
| Treasury Stock | -6.426e-10 |
| Estimated Shares Outstanding | -3.303e-08 |

4.3 Backward elimination

Also, we use backward elimination to remove the variables which do not contribute much to the target. By setting the p-value as 0.1, there are 11 variables eliminated.

Table 2. Eliminated variables



After this step, the regression model by backward elimination is as follows:

Table 3. The model attained by backward elimination

|  |  |
| --- | --- |
| Variable | Coefficient |
| Current Ratio | 5.539e-02 |
| Gross Profit | 2.144e-09 |
| Interest Expense | -1.159e-08 |
| Liabilities | 1.317e-08 |
| Net Receivables | -1.641e-09 |
| Operating Margin | 7.504e-01 |
| Other Current Assets | -4.936e-09 |
| Profit Margin | -6.549e-01 |
| Total Current Assets | 1.361e-09 |
| Total Revenue | -3.804e-10 |
| Treasury Stock | -5.789e-10 |
| Earnings Per Share | 6.069e-02 |
| Estimated Shares Outstanding | -3.218e-08 |

From the result above, obviously, almost all variables in the model attained by all-subsets regression are covered in the model of backward elimination, while there are 5 variables more in the model above. In order to select the best model, we calculate the AIC for the 5 models obtained in all-subsets regression and the model obtained by backward elimination.

Table 4. AIC for each model

|  |  |  |
| --- | --- | --- |
| Model | Num of variables | AIC |
| Model 1 by all-subsets regression | 8 | 16192.94 | |
| Model 2 by all-subsets regression | 9 | 16190.23 | |
| Model 3 by all-subsets regression | 10 | 16187.78 | |
| Model 4 by all-subsets regression | 11 | 16185.91 | |
| Model 5 by all-subsets regression | 12 | 16183.80 | |
| Model by backward elimination | 13 | 16182.95 | |

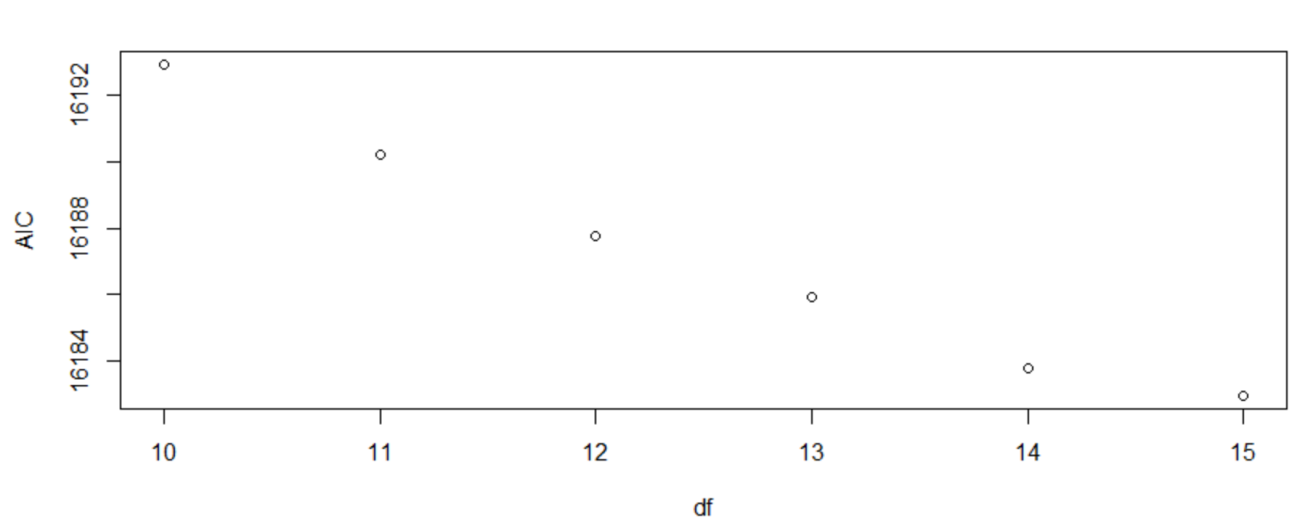


Fig 5. AIC for each model

Although the AIC of the model by backward elimination is the lowest, however, it is only slightly lower than other values. Since the best model attained by all-subsets regression has AIC slightly higher than the lowest one but contain much fewer variables, we choose it as our model.

4.4 Model Evaluation

After determining the regression model, we can evaluate the model based on some indicators.

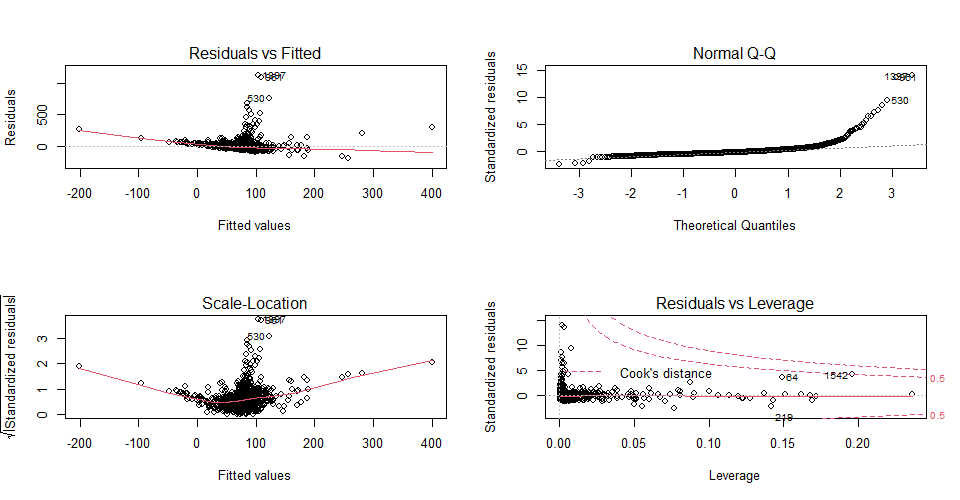


Fig 6. Diagnostic plots

First of all, from the upper left chart, there is no curve indicated, the points are distributed evenly except for several outstanding points. For the top right chart, we can see that the residual follows normal distribution since most points lie around the red line. From the bottom left plot, the result is basically in line with the assumption of equal variance. From the bottom right figure, there are some outliers. And these influential points with large cook distance are found, which are shown in Figure 7.

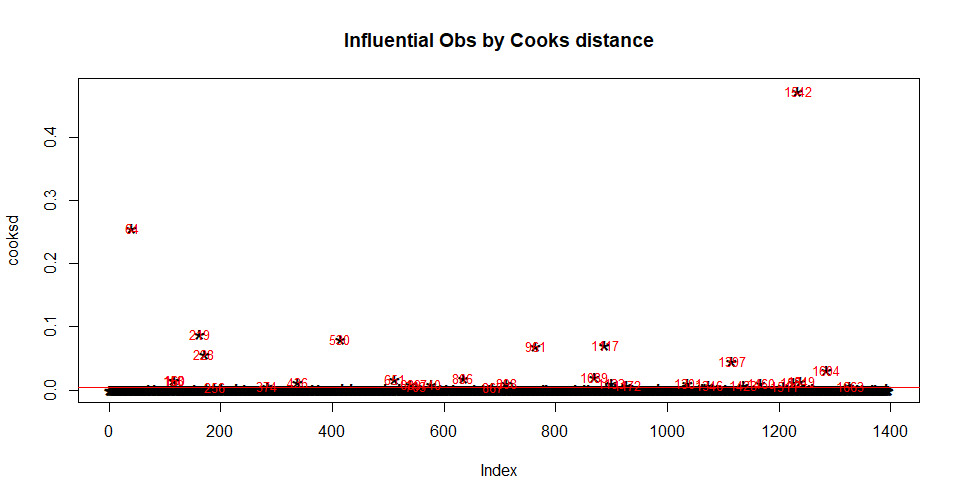


Fig 7. Outliers

After comparing these data points with others, it is found that they have relatively high values, however, for other variables the values are relatively low. And the reason behind this situation might be that the original variables are related, which means that some of them can be derived by other variables, and since we drop off some correlated variables, therefore, it is possible that the data are polarized, which makes them influential. Considering of this reason, we may not delete these data points.

Also, we check the multicollinearity for each variable and the result is shown in below:

Table 5. VIF

|  |  |
| --- | --- |
| Variable | VIF |
| Current Ratio | 1.065 |
| Gross Profit | 6.470 |
| Liabilities | 1.202 |
| Other Current Assets | 1.516 |
| Short Term Investments | 1.893 |
| Total Revenue | 4.167 |
| Treasury Stock | 1.693 |
| Estimated Shares Outstanding | 3.173 |

Since there is no variable with VIF higher than 10, we conclude that there is no multicollinearity in this model.

**5. Result:**

We have some intuitive understanding of the final model, and we have done summary and anova on the final model. From the data, we can see that several data are negatively correlated with our mean, such as Other.Current.Assets, Total.Revenue, Treasury.Stock and Estimated.Shares.Outstanding. The rest are positively correlated. From the summary data, we can see that Current.Ratio has a strong correlation with mean. Secondly, in the t-test, our p-value is also in a qualified state, which is less than or equal to the 0.01 we gave. For the data given by anova, for the F-test, only the p-value of Total Revenue is slightly larger.

**6. Conclusion：**

In the resulting model, I first know that Gross profit margin is a metric analyst use to assess a company's financial health by calculating the amount of money left over from product sales after subtracting the cost of goods sold (COGS). Sometimes referred to as the gross Margin ratio, gross profit margin is frequently expressed as a percentage of sales. So gross profit margin does have a correlation with the average annual stock price. When a company’s corporate finances are in a healthy state, natural stock prices will rise, and people are more likely or more inclined to buy shares of this company. On the other hand, for liabilities (financial accounting), as time grows, the company's debt situation is bound to affect the increase in the debt-to-asset ratio, and also affect the company's stock price. Of course, as we mentioned earlier, economic impact also has an impact on stock prices. Similar to the financial crisis in 2008, the economic downturn over time will also affect stock prices.

**Reference：**

The Relationship Between Accounting Information in the Financial

Statements and the Stock Returns of Listed Firms in Vietnam Stock

Exchange, By Ngoc Hung Dang1, Thi Viet Ha Hoang1 & Manh Dung Tran2

<https://www.researchgate.net/publication/319625568_The_Relationship_Between_Accounting_Information_in_the_Financial_Statements_and_the_Stock_Returns_of_Listed_Firms_in_Vietnam_Stock_Exchange>

<https://www.investopedia.com/terms/g/gross_profit_margin.asp>