Linear Regression Fails to Characterize the Relationship between Log Returns of Stocks

Abstract:

In stock market, there is a potential correlation between companies in the same industry. Using the right model, the stock price of one company can be expressed by the stock prices of other companies. In this report, linear regression model was used to verify such correlation. However, after model selection and outlier removal, the fitting results were still unsatisfactory. The log return of Toyota and Ford cannot be used to interpret the log return of GM.

1. Introduction

The stock market is an extremely complex and unpredictable system. Investors are eager to find a pattern behind the prices, such as the Dead Cat Bounce1 and the Dreaded Vomiting Camel2. Some of these hypotheses have been proven unreliable. But it is an undeniable fact that the stock prices in the same industry usually share similar trends. The economic explanation behind this is these companies have risk exposures to the same factors. For example, traditional auto manufactures are all exposed to the same risks of fuel economy, consumption power and taxes. Changes in these factors would cause the prices of most auto manufactures moving up or down simultaneously.

This report tries to find the correlation of prices between companies within the same industry. After applying linear regression on the log returns of three auto manufactures: Toyota, Ford and GM, result shows that the log return of Ford and GM share similar trends. However, even after model selection and removing all outliers, the correlation is still not strong enough. And the correlation with Toyota is even lower. Thus, the log return of Toyota and Ford cannot be used to interpret the log return of GM.

The rest of this report is organized as follows. Section 2 will introduce some basic concepts of linear regression. Detailed data manipulation and result interpretation will be introduced in Section 3, followed by a brief conclusion in Section 4.

1. Linear Regression

Linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X. 3

The formal definition of linear regression model is as below.

Y = X beta + epsilon

where y denotes the dependent variable, x denotes the explanatory variables, beta denotes coefficients and epsilon denotes errors.

To find the coefficients that minimize square error, the formula for the least square fit is:

1. Experiment and Analysis

After removing outliers1(-311, -334, -402, -644), the basic information of the entire data set is shown in Table 1.

Table1

Toyota Ford GM

Min. :-5.46562 Min. :-9.48100 Min. :-7.92372

1st Qu.:-0.84036 1st Qu.:-0.93526 1st Qu.:-1.18832

Median : 0.06397 Median :-0.02082 Median : 0.04305

Mean : 0.09710 Mean : 0.02336 Mean : 0.03873

3rd Qu.: 1.05164 3rd Qu.: 1.14672 3rd Qu.: 1.16914

Max. : 5.53078 Max. : 7.42877 Max. : 7.44415

There exists a significant difference between Ford’s median and mean, indicating that the data has positive skewness, meaning there are more extreme large values. This could be a potential problem but the difference is in a reasonable range in this case.

Since the linear regression of (GM ~ Toyota + Ford) showed that the coefficient of Toyota is not statistically significant, Toyota is removed from the data set. The linear regression of GM ~ Ford is shown in Table2.

*Residuals:*

*Min 1Q Median 3Q Max*

*-5.7603 -1.0106 -0.0173 0.8664 5.6191*

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) 0.02479 0.05782 0.429 0.668*

*V2 0.59678 0.03023 19.740 <2e-16 \*\*\**

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Residual standard error: 1.535 on 703 degrees of freedom*

*Multiple R-squared: 0.3566, Adjusted R-squared: 0.3557*

*F-statistic: 389.7 on 1 and 703 DF, p-value: < 2.2e-16*

The p-value of V2 is small enough and its T value is large, which means this coefficient does not rely on sampling. However, the P value and T value of intercept are pretty bad, showing the intercept value is not usable. In the meanwhile, an adjusted R-squared value of 0.3557 indicates that only 35% of the variance can be explained with this model. In some situation, it is an acceptable value. But in this case, a more convincing number is needed since volatility is one of the key features of stock price. A number of 35% is not good enough to accurately describe the connection between GM and Ford.

Also, in Fig 1, the outputs are still not satisfactory. This model is not sufficient to characterize the relationship between the log returns.



1. Conclusion

Given the log return of Toyota, Ford and GM, linear regression was performed, trying to find the potential price relationship between the three auto manufactures. However, the results show their linear relations are not clear. Though the log return of GM and Ford are positively correlated, they cannot be expressed in a linear model. Some more advanced non-linear model will be tested in future works.

Reference

1. Investopedia. <http://www.investopedia.com/terms/d/deadcatbounce.asp>
2. Brian Kelly, CNN. <http://www.cnbc.com/id/102147311>
3. Linear Regression. <https://en.wikipedia.org/wiki/Linear_regression>

Appendix