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

ISYE/MATH 6783 - Assignment1

Quan Zhou
Quantitative and Computational Finance
Georgia Institute of Technology
Email: qzhou81@gatech.edu

***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.**

I. 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 Bounce[1] and the Dreaded Vomiting Camel[2]. 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.

II. 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.

$$y = X \times \beta + \epsilon$$

where y denotes the dependent variable, X denotes the explanatory variables, β denotes coefficients and ϵ denotes errors.

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

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

where $\hat{\beta}$ denotes the estimated coefficient, X denotes the explanatory variables and y denotes the dependent variable.

III. Regression and Analysis

After removing outliers[1], the basic information of the entire data set is shown in TABLE I.

TABLE I
Basic infomation of dataset

	Toyota	Ford	GM
Min.	-5.46562	-9.48100	-7.92372
1st Qu.	-0.84036	-0.93526	-1.18832
Median	0.06397	-0.02082	0.04305
Mean	0.09710	0.02336	0.03873
3rd Qu.	1.05164	1.14672	1.16914
Max.	5.53078	7.42877	7.44415

[1]The removed outliers are (-311, -334, -402, -644)

TABLE II
RESULT OF GM $\sim$ FORD

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
Ford	0.59678	0.03023	19.74	<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				

There exists a significant difference between Fords 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 $\sim$ Toyota + Ford) showed that the coefficient of Toyota is not statistically significant, Toyota is removed from the data set. The linear regression of (GM $\sim$ Ford) is shown in TABLE II.

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 itsuation, 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. The residuals vs. fitted plot is not uniformly distributed, indicating the assumption that the relationship is linear may not be reasonable. Plus, there still exists some points that are far from others, implying existance of potential outliers. The normal Q-Q plot does not follow a straight line. Not all values are normally distributed. Similar to the residual plot, the scale location plot are not homogeneous neither. And the cook's distance of some points are still larger than others, which means they are not equally weighted. According to the plots, linear model is not sufficient to characterize the relationship between the log returns.

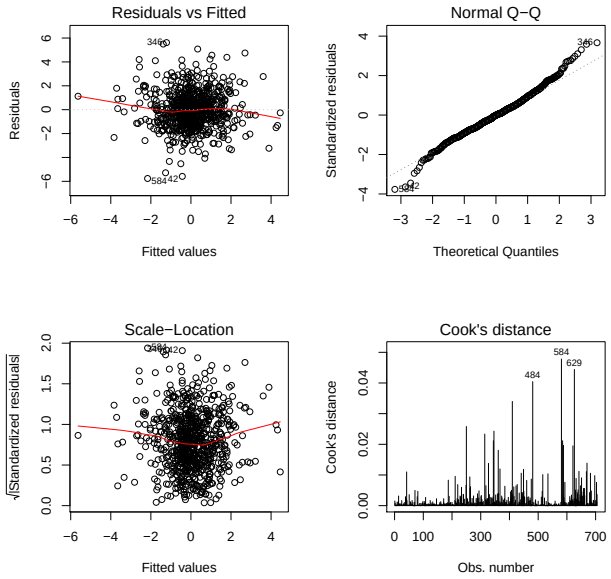

Fig. 1. Result of GM $\sim$ Ford

IV. 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.

REFERENCES

[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 A
LINEAR REGRESSION ON THE ENTIRE DATASET

TABLE I
RESULT OF GM $\sim$ TOYOTA + FORD

Residuals:				
Min	1Q	Median	3Q	Max
-6.2848	-0.9649	-0.0405	0.8977	5.7515
Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	0.007049	0.059138	0.119	0.905
Toyota	0.061321	0.037840	1.621	0.106
Ford	0.614496	0.031322	19.619	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.572 on 706 degrees of freedom
Multiple R-squared: 0.3775, Adjusted R-squared: 0.3757
F-statistic: 214.1 on 1 and 706 DF, p-value: <2.2e-16

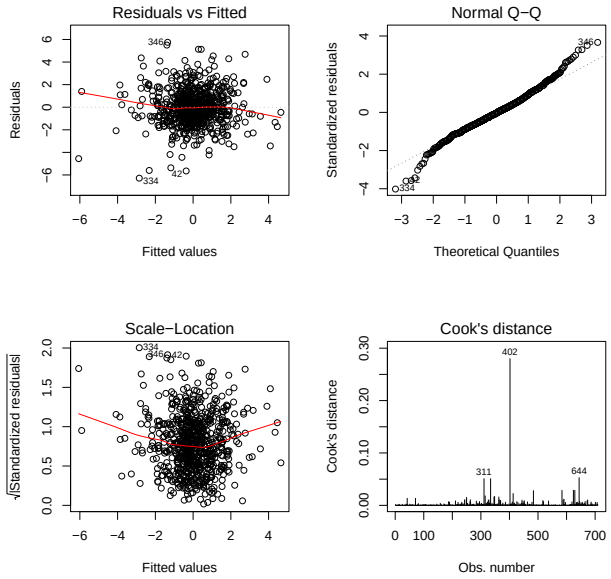

Fig. 1. Result of GM $\sim$ Toyota + Ford

APPENDIX B
R CODE

```
# read data
data.full <- read.csv("E:\\QCF\\2016Spring\\FinancialDataAnalysis\\hw1\\w_logret_3automanu.csv",
header = FALSE)
data.full <- data.full * 100

# summary
summary(data.full)

# linear regression on the entire dataset
lm.full <- lm(V3 V1+V2, data = data.full)
summary(lm.full)
par(mfrow=c(2,2))
plot(lm.full, which=c(1:4))

#remove outliers
data.sub <- data.full[c(-311, -334, -402, -644),]
lm.sub <- lm(V3 V1+V2, data = data.sub)
summary(lm.sub)
par(mfrow=c(2,2))
plot(lm.sub, which=c(1:4))

# linear regression between GM and FORD
lm.ford <- lm(V3 V2, data = data.sub)
summary(lm.ford)
par(mfrow=c(2,2))
plot(lm.ford, which=c(1:4))
```