UNIVERSITY OF CALIFORNIA

Los Angeles

Stock Analysis with Correlation for Gasoline Companies

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Computer Science

by

Yi Xin Sun

2015

ABSTRACT OF THE THESIS

Stock Analysis with Correlation

for Gasoline Companies

by

Yi Xin Sun

Master of Science in Computer Science

University of California, Los Angeles, 2015

Professor Miodrag Potkonjak, Chair

The stock market is a risky and popular attraction for those wishing to add a little extra to their income. Various models have been developed to predict stock movements; they range from complexity to the variables taken into account. Some models might also present performance instability depending on the industries they are simulated in. It then becomes difficult for investors to select and learn a profitable model in a short amount of time.

In this thesis, we develop simple models that aim at maximizing profit for investors in the gasoline industry. Our models focus on correlations between the companies as the bigger players in the industry are often influential enough to affect the stock prices of the smaller companies. Five chosen companies are paired up and several

tests were done to determine the possible occurrence of certain patterns. Three evaluation methods are used to validate and finally, confirm the existence of those patterns.

The thesis of Yi Xin Sun is approved.

Carlo Zaniolo

Milos D. Ercegovac

Miodrag Potkonjak, Committee Chair

University of California, Los Angeles

2015

TABLE OF CONTENTS

Chapter 1

Introduction

The stock market has always been a roller coaster ride for investors; it brings easy profit along with huge risk. Studying a stock's momentum and making investment decisions before possible peaks and dips movements are what investors are trying to do. There are many stock market and they grow and change with different rules. In the recent decade, the number of trading companies has grown tremendously making investment situations trickier and more difficult. Therefore, there exist no concrete sets of measurements or rules that determine the patterns one should be looking for when making investment plans. These patterns may vary from that of a single stock to multiple related stocks. It is a trial and error process to find and test a model that genuinely reflects a stock's movements as momentum varies for different companies. It is equally important to be able to check the models with the right methods. With the many options out there, it becomes a nightmare to pick the correct models and methods.

A majority of the currently available models can be categorized into two kinds: technical historical data and current trends analysis. The problem then becomes: which model to choose to maximize our portfolio profit. In order to differentiate this thesis with the many other papers out there discussing the best analysis methods, we chose to find the correlation between stocks in the same market. Instead of focusing on how just one company acts over time, we focus on finding lag correlation in terms of days to measure the amount of influence a company can have on others. With these relations, we hope to pick out one or two models as a solution to our problem.

In this thesis, we present several different self-developed models to predict future stock action. These models have been trained based on historical prices based on certain patterns seen in the datasets. In order to be more accurate, we restricted the companies studied to major gasoline companies based in United States. The companies are also investigated through the models and patterns in order to better determine their relationship.

Four models were carefully chosen through data analyzing; they are a mixture of different multi data relations. The models are evaluated using two traditional statistical techniques: re-substitution and learning and test. An additional investment strategy is put into test to simulate actual investments and the gain or loss is then recorded. The final part will be a detailed conclusion and review on the results of the various tests the models went through.

Chapter 2 goes over several famous models and books that present similar ideas and methods as our thesis. A detailed explanation about the different self-developed models and information on the data used are given in the third chapter. Statistical results are also included to give a clearer picture of the models functionalities. In Chapter 4, thorough evaluations are performed using three techniques to determine how well and accurate our models are. Finally, a conclusion is formed regarding the patterns and relationships detected with the models.

Chapter 2

Related Work

This chapter talks about the different works that have had an impact on our work. Burton Gordon Malkiel's book on random walk hypothesis provides a great insight into stock market analysis. Other related work include B. Efron's book on the jackknife and bootstrapping methods used to evaluate statistical models. The tests we conduct for evaluation are comparable to the methods Efron discussed in the book. Our models are based on correlation, which is explained in Kendall's work on rank correlation.

Malkiel discusses several strategies for investing in the market under two ideologies: firm foundation theory and the "castle in the air" theory [2]. The technical analysis under the foundation theory states that one should invest based on the movement of historical prices [3] while the "castle in the air" theory says investments should be made based on the popularity of the stock's industry. However, Malkiel writes, "It turns out that the correlation of past price movements with present and future price movements is very close to zero." [4]. Even though Malkiel dismisses the technical analysis saying that it is mostly led by spurious and random events, he admits that it should be respected.

Bootstrapping is a resampling method that "allows estimation of the sampling distribution of almost any statistic using random sampling methods" [5]. This method is derived by Efron after discovering the Jackknife method [6]. Despite the similarities, bootstrap applies to a wider range of problems. Efron states, "The jackknife can be thought of as a linear expression method (a “delta method”) for approximating the bootstrap.” [7]. The bootstrap is perfect for constructing hypothesis tests.

Kendall’s work “is a statistic used to measure the association between two measured quantities” [8]. The Kendall coefficient is defined as follows:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)}$$

Figure 2.1 Kendall coefficient definition

The difference of the number of ordered pairs in two sets is calculated and normalized [9]. By using this coefficient, one is able to calculate the statistical dependency between two variables. The Kendall rank can be easily generalized to for application to other statistical measurements.

Chapter 3

Date and Models

This chapter goes into details behind the ideology of each model and supports the findings with statistical data.

3.1 Data

The five U.S companies chosen are Chevron, Conoco, Eog, Exxon and Occidental. We downloaded around twelve years worth of historical stock data and first graphed them to visualize any sort of similar momentum movement the companies might share. These data were attained from Yahoo as opposed to other online databases such as the Wharton Research Data Services since it was easier to obtain and had the complete set of data required.

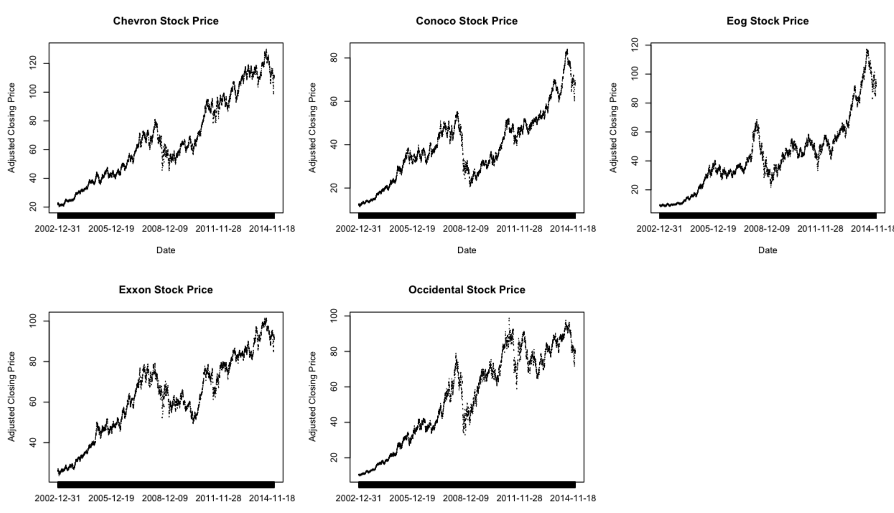

Figure 3.1 Graphs of U.S historical stock price

As expected, the overall shapes of the graphs look alike to each other indicating that these stocks share certain patterns. The standard measurement is set to the daily change of that stock. A negative signal is set if a company's daily closing price was lower than the closing price from the previous day. When comparing multiple companies, the term "inversion" is used when the prices are of different signs on a day. All models are trained using daily stock prices from the year 2013.

3.2 Models

All the models involve multiple company comparisons. They are structured based on the understanding that bigger companies are more influential therefore causing them to "foresee" changes in the market. By exploring the different patterns and options available, we dip into a deeper analysis of the relationship between stocks of similar nature.

3.2.1 Company A Leads Company B (Two Companies Inversion)

This is a model that involves two companies: A and B. If there is a decrease in A's price on day T ($Close_T - Close_{T-1}$) and B's price increases within the next n days ($Close_{T+n} - Close_T$), then A leads B with a lag of n days. The American companies were tested with each other for lags ranging from one to five days.

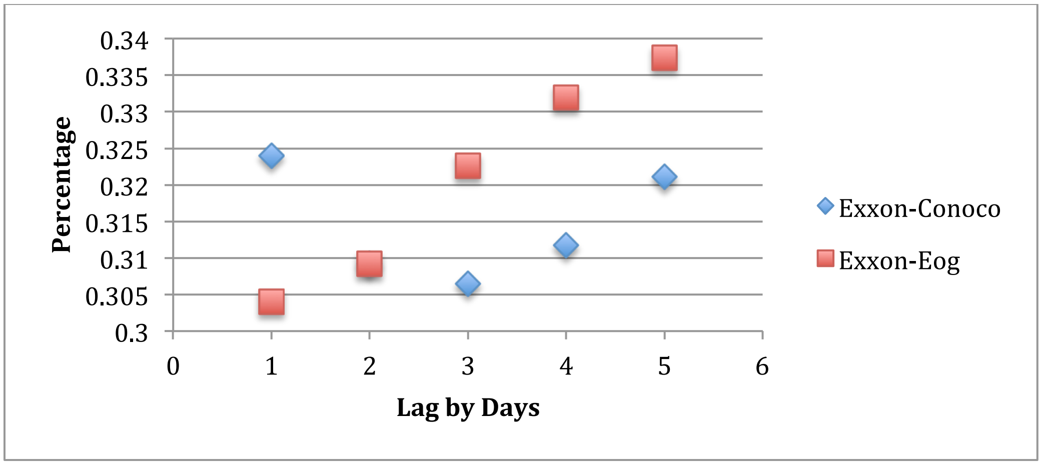

Figure 3.2 Different lag percentage of correlation between the top pairs by days for Two Companies Inversion model.

With these data, we form the following table that points out the top correlated pairs with different lag days.

A-B	Lag (Days)	Percentage
Exxon-Eog	5	0.34
Exxon-Eog	4	0.33
Exxon-Conoco	1	0.32

Table 3.3 Top three most correlated companies for Two Companies Inversion model.

The average of all the results was around 0.27. The top three results are decently higher than the average, which is what we are looking for in this case. Another observation is that Exxon tends to be the leader of the group, leading every company with an average of 0.31. Eog, on the other hand, is an underachiever with an average of 0.24.

3.2.2 Company A Leads Company B (Two Companies Non-Inversion)

This is very similar to the inversion-leading model except we are looking for non-inversions this time. If there is an increase in A's price on day T ($Close_T - Close_{T-1}$) and B's price increases as well within the next n days ($Close_{T+n} - Close_T$), then A leads B with a lag of n days with non-inversion.

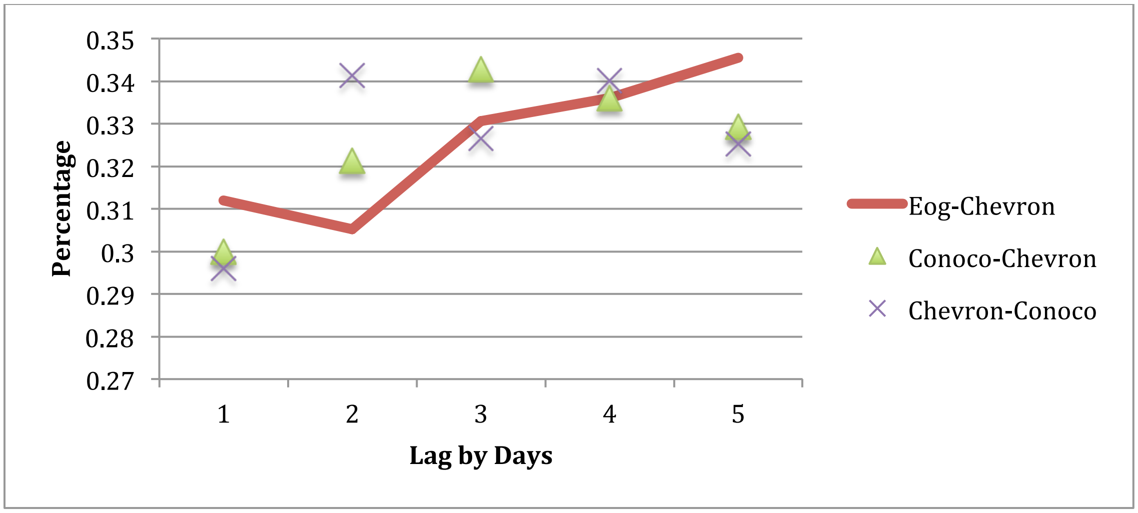

Figure 3.4 Different lag percentage of correlation between the top pairs by days for Two Companies Non-Inversion model.

With these data, we form the following table that points out the top correlated pairs with different lag days.

A-B	**Lag (Days)**	**Percentage**
Eog-Chevron	5	0.35
Conoco-Chevron	3	0.35
Chevron-Conoco	2	0.34

Table 3.5 Top three most correlated companies for Two Companies Non-Inversion model.

The average was around 0.29. Interestingly, Exxon exhibits the lowest average of a mere 0.26 while Eog reaches an average of 0.31.

3.2.3 Companies AB Leads Company C (Three Companies Inversion)

Three companies are involved in this model: A, B and C. Company C is dependent on the other two companies to lead it. If there is decrease in both A's and B's prices on day T ($Close_T - Close_{T-1}$) and C's price increases within the next n days ($Close_{T+n} - Close_T$), then A and B leads C by n days.

A&B-C	Lag (Days)	Percentage
Conoco&Exxon-Eog	5	0.24
Conoco&Exxon-Occidental	4	0.23
Conoco&Exxon-Occidental	5	0.23

Table 3.6 Top three most correlated companies for Three Companies Inversion Model.

The percentages are relatively low compared to the previous two models. This implies that it is less likely for two companies to lead the market. Exxon and Conoco seem to be the prominent candidates that drive the market. Exxon is also the leading company in Two Companies Inversion model.

3.2.4 Two Straight Days (Straight Days Inversion)

The relation between companies is sometimes longer than just a day and this model looks into a two days relation. If there is a decrease in A's price on days T and T+1 ($Close_{T+1} - Close_T$) and B's price increases as within the next n days ($Close_{T+1+n} - Close_{T+1}$), then A leads B for two days with a lag of n days with inversion.

A-B	Lag (Days)	Percentage
Exxon-Eog	4	0.17
Exxon-Eog	5	0.16
Exxon-Eog	3	0.16

Table 3.7 Top three most correlated companies for Straight Days Inversion model.

Interestingly, Exxon and Eog seem to have the tightest connections as seen also in Two Companies Inversion model. The percentage recorded for other connections barely reached 0.10.

Chapter 4

Evaluation

This chapter looks at three evaluation methods used to test the accuracy of the models we developed.

4.1 Learning and Test

The learning and test method is very straightforward; the trained models are tested with historical prices from year 2014. This is an easy and useful way to confirm if the patterns our models are based on are legitimate and will reappear in the near future. In the following, the results are analyzed to evaluate a model's reliability.

4.1.1 Company A Leads Company B (Two Companies Inversion)

A-B	**Lag**	**Percentage**		**A-B**	**Lag**	**Percentage**
Exxon-Eog	4	0.32		Exxon-Eog	5	0.31
Exxon-Eog	5	0.31		Exxon-Eog	4	0.32
Exxon-Conoco	5	0.30		Exxon-Conoco	1	0.28

Table 4.1 A comparison between the 2013 highest three company percentages and the 2014 new percentages for those companies for Two Companies Inversion model.

From the table above, it shows that the inversion-leading scheme is valid and does repeat itself. The results on the left side shows the new top three percentages while the right side depicts the new percentages for the top three companies from the previous year. Two pairs made it to the new top three. This also indicates the degree of influence Exxon has on other companies, especially Eog. Even though the statistics for the Exxon-Conoco pair with lag of one day took a dip, it was still above the average of around 0.25; the replacement Exxon-Conoco pair with five-day lag hints at a long-term effect instead.

4.1.2 Company A Leads Company B (Two Companies Non-Inversion)

A-B	**Lag**	**Percentage**		**A-B**	**Lag**	**Percentage**
Conoco-Eog	4	0.32		Eog-Chevron	5	0.27
Conoco-Exxon	3	0.31		Conoco-Chevron	3	0.27
Conoco-Eog	5	0.30		Chevron-Conoco	2	0.25

Table 4.2 A comparison between the 2013 highest three company percentages and the 2014 new percentages for those companies for Two Companies Non-Inversion model.

The old top three pairs did fairly poorly for the new-year and barely achieved scores around the average. Despite so, Conoco seems to be highly influential, as it had appeared in the old patterns as well. A possible explanation is Conoco becoming more powerful and starting to have more effect on the market. Chevron gains independence and is not led easily by other companies.

4.1.3 Companies AB Leads Company C (Three Companies Inversion)

A&B-C	**Lag**	**Percentage**
Chevron&Exxon-Conoco	5	0.24
Chevron&Exxon-Eog	4	0.23
Exxon&Occidental-Eog	4	0.22
A&B-C	**Lag**	**Percentage**
Conoco&Exxon-Eog	5	0.21
Conoco&Exxon-Occidental	5	0.21
Conoco&Exxon-Occidental	4	0.20

Table 4.3 A comparison between the 2013 highest three company percentages and the 2014 new percentages for those companies for Three Companies Inversion model.

The three-way pattern is a long-term relationship where two companies react within a reasonable amount of time to changes in the market. Exxon, as seen from Two Companies Inversion model, has always been dominant for inversion patterns while Eog is on the weaker side. The new statistics confirms this occurrence as Eog replaces Occidental. Chevron is overtaking Conoco in becoming a closer partner with Exxon.

4.1.4 Two Straight Days (Straight Days Inversion)

A-B	Lag	Percentage		A-B	Lag	Percentage
Chevron-Eog	4	0.14		Exxon-Eog	4	0.12
Chevron-Eog	5	0.14		Exxon-Eog	5	0.12
Chevron-Conoco	5	0.13		Exxon-Eog	3	0.09

Table 4.4 A comparison between the 2013 highest three company percentages and the 2014 new percentages for those companies for Straight Days Inversion model.

The leading company is changed from Exxon to Chevron, which is another symbol of power gaining. Eog continues to remain lower down in the food chain with Conoco. The Exxon-Eog pairs are doing fairly well for the new-year. Overall, the pattern statistics has fallen from 0.17 to 0.14. This pattern occurs less frequently compared to that of the other models but still provides us with valuable information.

4.2 Re-substitution

Re-substitution is an evaluation method that randomly partitions the data based on a random number generator. Random days are picked from the 2013 dataset that the models are trained with. In this situation, 50% is the threshold. The dates picked out are the leads and further calculations are done for lags. The same random set of days is used for the three re-calculations for each model. The idea behind this technique is to verify the patterns by including randomness, which removes aggregation possibilities. If a random segmentation of data displays the pattern then it is extremely likely that the whole set of data, instead of a portion, follows that general pattern.

4.2.1 Company A Leads Company B (Two Companies Inversion)

A-B	Lag (Days)	Percentage
Exxon-Eog	5	0.33
Exxon-Eog	4	0.32
Exxon-Conoco	1	0.33

Table 4.5 Re-substitution statistics on the top three pairs for Two Companies Inversion model.

The resulting statistics is similar to the original: they are considerably higher than the average. The slightly decreased percentage for the second pair might be indicative that the actual chance of the pattern occurring is a number slightly lower than what we previously calculated. This confirms our suggestion that an inversion-leading pattern exists for these three pairs of companies. Their tight correlation is something that will be helpful to stock investors.

4.2.2 Company A Leads Company B (Two Company Non-Inversion)

A-B	Lag (Days)	Percentage
Eog-Chevron	5	0.33
Conoco-Chevron	3	0.31
Chevron-Conoco	2	0.35

Table 4.6 Re-substitution statistics on the top three pairs for Two Companies Non-Inversion model.

The result deviation is slightly more obvious in this model. The percentages are staying reasonably well above the average indicating the presence of a pattern. The re-calculated statistics for the first and second pairs of company are lower than what we expected. This indicates some aggregation pattern in the data, but overall, the results are determinative regarding the pattern behind this model.

4.2.3 Companies AB Leads Company C (Three Companies Inversion)

A&B-C	Lag (Days)	Percentage
Conoco&Exxon-Eog	5	0.24
Conoco&Exxon-Occidental	4	0.21
Conoco&Exxon-Occidental	5	0.25

Table 4.7 Re-substitution statistics on the top three pairs for Three Companies Inversion model.

The second pair of company displays deviation while the numbers remain almost unchanged for the rest. The lowered percentage hints at pattern instability and should be noted if this model is to be used. Since this pattern involves two companies to lead at the same time, its statistics is more susceptible to aggregation and chance. However, the re-substitution method proves that this relationship between the companies exists.

4.2.4 Two Straight Days (Straight Days Inversion)

A-B	Lag (Days)	Percentage
Exxon-Eog	4	0.17
Exxon-Eog	5	0.17
Exxon-Eog	3	0.16

Table 4.8 Re-substitution statistics on the top three pairs for Straight Days Inversion model.

The new statistics turns out to be better than the original; all three pairs resulted in increased percentages. The random dataset might have included a range of dates where the pattern was prominent. By now, it has become very obvious that the company Exxon has a special presence. In reality, Exxon does have the largest market capitalization out of the five selected companies.

4.3 Investment Strategy

The investment strategy is an evaluation technique that simulates real-world investment situations. In this strategy, all trained models are put into use with the 2014 data. The start investment amount is $1,000,000. Five separate investment plans are tested out for each scenario. The different plans vary by the portion of investment that is put into the market every time a pattern signal is detected. For comparison, a plan where the full amount is invested for a whole year is included.

This method allows us to observe how well our models serve in reality. It also visually displays the correlation between risk and the percentage invested. Is it true that the more we invest, the more we lose or gain? With these questions in mind, we can take a look at the numerical data and understand the market better to our benefits. Investors have their own mindsets and habits and thus we need different plans to fully reflect investor choices.

4.3.1 Company A Leads Company B (Two Companies Inversion)

Investment	Stock	Lag	Profit	Percentage
	Exxon-Eog	NA	121985.6	12.20
100%		5	53599	5.36
50%		5	58642	5.86
25%		5	97118	9.71
10%		5	71926	7.19
	Exxon-Eog	NA	121985.6	12.20
100%		4	85609	8.56
50%		4	89033	8.90
25%		4	83807	8.38
10%		4	56750	5.68
	Exxon-Conoco	NA	27765.25	2.78
100%		1	84912	8.49
50%		1	22630	2.26
25%		1	1021	0.10
10%		1	-2486.6	-0.25

Table 4.9 Investment Strategy results for the top three pairs of companies for Two Companies Inversion model.

This model proves itself to be effective as all scenarios except one provided gains. From the overall return, Eog's price increased significantly through the year; our model did not do exceptionally well as the highest return is 3% lower. For the first pair, it appears that going for a less risker option rewards more gains. Raising the risk for the other two pairs works better for a nicer return. More data and analysis will be needed if we were to determine the level of risk for each pair. However, through this test, the model displays reliability in terms of returns.

4.3.2 Company A Leads Company B (Two Companies Non-Inversion)

Investment	Stock	Lag	Profit	Percentage
	Eog-Chevron	NA	-64115.12	-6.41
100%		5	-157228.1	-15.72
50%		5	-59678.7	-5.97
25%		5	-24163.4	-2.42
10%		5	-5397.6	-0.54
	Conoco-Chevron	NA	-64115.12	-6.41
100%		3	75506	7.55
50%		3	49093	4.91
25%		3	44067	4.41
10%		3	25156	2.52
	Chevron-Conoco	NA	27765.25	2.78
100%		2	181108	18.11
50%		2	77963	7.80
25%		2	37626	3.76
10%		2	15877	1.59

Table 4.10 Investment Strategy results for the top three pairs of companies for Two Companies Non-Inversion model.

The model performed extremely well for the two pairs involving Chevron and Conoco. The returns were higher for the risker investment plans surprisingly. This is an indication that there were days with extremely high returns while bad days produced manageable loss. The first scenario however, produced negative returns for all five investment cases. A simple and reasonable explanation would be a drastic change in the relation between the two companies.

4.3.3 Companies AB Leads Company C (Three Companies Inversion)

Investment	Stock	Lag	Profit	Percentage
	Conoco&Exxon-Eog	NA	121985.6	12.20
100%		5	76741	7.67
50%		5	-4521.8	-0.45
25%		5	45809	4.58
10%		5	37640	3.76
	Conoco&Exxon-Occidental	NA	-77160.98	-7.72
100%		4	-69767.3	-6.98
50%		4	-92354.76	-9.24
25%		4	-47777.87	-4.78
10%		4	-15898.23	-1.59
	Conoco&Exxon-Occidental	NA	-77160.98	-7.72
100%		5	-75161.7	-7.52
50%		5	-92122.5	-9.21
25%		5	-32028	-3.20
10%		5	-5366.5	-0.54

Table 4.11 Investment Strategy results for the top three pairs of companies for Three Companies Inversion model.

This model performed rather disappointingly. There was no positive return except for certain cases in the first scenario. The first scenario contained a relation between Exxon and Eog, which has been appearing frequently in other models. The gains are largely due to this special connection between the two companies. On the other hand, Occidental has rather unstable affiliations with Conoco and Exxon. This might be due to a change in one of the companies.

4.3.4 Two Straight Days (Straight Days Inversion)

Investment	Stock	Lag	Profit	Percentage
	Exxon-Eog	NA	121985.6	12.20
100%		4	-299177.2	-29.92
50%		4	-146647.2	-14.66
25%		4	-58959.42	-5.90
10%		4	-18063.52	-1.81
	Exxon-Eog	NA	121985.6	12.20
100%		5	-180257.8	-18.03
50%		5	-45685.28	-4.57
25%		5	-9414.119	-0.94
10%		5	1403.368	0.14
	Exxon-Eog	NA	121985.6	12.20
100%		3	-337122.6	-33.71
50%		3	-222144.3	-22.21
25%		3	-113167.3	-11.32
10%		3	-42899.08	-4.29

Table 4.12 Investment Strategy results for the top three pairs of companies for Straight Days Inversion model.

The results only contain one positive return by 0.15%. Despite the pleasing results from the re-substitution test, the model does not seem to work well under actual investments. Not only were there almost no gains, the losses were extremely high. The loss was as high as 33.7% for one of the cases. One observation is that the risk rises with the invest percentages. Compared to the 12.2% gain from solely leaving the fund in the market for a year, the results are very discouraging.

Chapter 5

Conclusion

We studied the problem of predicting the momentum of gasoline companies by measuring the degree of correlation via various models. The principle behind the four self-developed models is that there exist strong correlations between companies in the same economic segment. In these studies, we used historical stock data from Yahoo Finance and wrote scripts in R language. The models were selected out of a variety of possible correlation patterns. The patterns are then validated by evaluation methods. Finally, we decide to either accept or reject each model resulting in a set of useable tools to maximize an investor's portfolio profit.

A model is accepted if it passes two out of the three tests. The result concluded in two accepted models: two companies inversion and two companies non-inversion. The rejected patterns were too volatile for us to conclude that the relationships were solid. Exxon, having the largest market capitalization in U.S leads all three pairs in the inversion model. Chevron, another major player in the team, is largely involved in the non-inversion relationships with other companies, particularly Conoco. It can be concluded that certain companies share similar trends that can be used to maximize our investment profit.

References

[1] Abdi, Herve. "Kendall Rank Correlation Coefficient." *SpringerReference* (2011): n. pag. Web. 17 Oct. 2015.

[2] "Bootstrapping (Statistics)." *Wikipedia.* Wikimedia Foundation, n.d. Web. 17 Oct. 2015.

[3] Efron, B. "Project Euclid." *Efron : Bootstrap Methods: Another Look at the Jackknife.* N.p., n.d. Web. 17 Oct. 2015.

[4] "Jackknife Resampling." *Wikipedia.* Wikimedia Foundation, n.d. Web. 17 Oct. 2015.

[5] "Kendall Rank Correlation Coefficient." *Wikipedia.* Wikimedia Foundation, n.d. Web. 17 Oct. 2015.

[6] Malkiel, Burton Gordon. *A Random Walk down Wall Street.* New York: Norton, 1973. Print.

[7] "Review: A Random Walk Down Wall Street - The Simple Dollar." *The Simple Dollar Review A Random Walk Down Wall Street Comments.* N.p., 13 Apr. 2007. Web. 10 Sept. 2015.

[8] Udland, Myles. "Obsessing Over Stock Charts Puts Holes In Investors' Shoes And Yachts In Brokers' Docks." *Business Insider*. Business Insider, Inc, 07 Sept. 2014. Web. 15 Oct. 2015.

[9] "Why Is Summer Fuel More Expensive than Winter Fuel?" *HowStuffWorks.* N.p., 05 June 2008. Web. 27 Sept. 2015.