NYU Courant MATH-GA 2708

Assignment2 Written Report

Yuanzhe Yang yy3460

Chien-Yueh Shih cs6380

Zedi Qiu zq2015

Problem 1

Readme:
In this prompt, we will build an impact model using public data from 2007-06-20 to 2007-09-20.

The functionalities include:

1) TAQTradeReader: import the trade data
2) TAQQuotesReader: import the quote data
3) TAQMatrix_test: perform unit test for our script
4) TAQAdjust: filter out only S&P 500 listed tickers and adjust the stock price and size for stock splitting/ stock buyback.
5) TAQProcess: process ticker by ticker with multi-threading to save the stocks feature values in the matrix form.
6) TAQRegression: perform non-linear regression to compute market impact parameters and perform residual analysis for the regression model
7) TAQFilter: extract S&P 500 stocks list and stock splitting information.
8) TAQMatrices: perform tools for matrix construction in the TAQProcess steps.

The working flow in this project will be from TAQTrade/QuoteReader → TAQFilter → TAQ_Adjust → TAQ_Process → TAQ_Regression. In the meanwhile, we could run the TAQMatrix-test to perform unit test for our matrix computation
By noticing that the tickers data are independent with each other, we implement a parallel processing mechanism using Joblib which enables us to run multiple tasks at the same time. The maximum tasks are confined by how many CPU you have with your device.

Since Cleaning and Computing matrix is very time consuming, we only show the regression output in the window between 20070907 and 20070920. If you are interested in exploring more data in different date scope, feel free to check it out.

2） Model Explanation
In this model, we will leverage the stock's value matrix includes

1) 2-minute mid-quote returns
2) Total daily volume
3) Arrive price
4) Terminal price
5) Imbalance share between 9:30 to 3:30
6) Volume-weighted average price between 9:30 to 3:30
7) Volume-weighted average price between 9:30 to 4:00

and calculate the 10-day look-back on average daily and the 10-day standard deviation of mid-quote returns for regression.

In this case, our non-linear regression model is defined as:

$$h = \eta\sigma\left(\frac{X}{(6/6.5)V}\right)^{\beta} + \langle noise \rangle$$

where

h: is the temporary impact: $(\tilde{S} - S_{930})$ – g with g measures $\frac{S_{400}-S_{300}}{2}$

X: is the imbalance value between 9:30 AM and 3:30PM
V: is the average daily volume
σ: is the standard deviation of 2-min returns scaled to 1 day
η: is the temporary market impact
β: is the temporary market impact

We create a class named TAQMatrix to help us compute all the needed value matrices. To be more specific, we will compute Imbalance share, total daily volume, and vwaps from trades data and 2-minute mid-quote, Arrival price, terminal price from quotes data.

After we collected all the needed data matrices, we implemented the TAQRegression which can help us easily run the None Linear Regression on input data and generate analysis on factor parameters (eta/ beta).

Before proceeds into the model, we performed NA value checking and drop the rows with large proportions of NA values. Our aim is to leave the dates and tickers with few NA values in their scope. To be more specific we set the default value of max_null_in_day to be 40 and max_null_in_ticker to be 3. This filter drops ten tickers' and one day's data. After that, we fill in the rest NA values with the mean values. In our case, we use Scipy curve_fit optimization methods to help us find the fitted value for the non-linear regression. We also implemented an alternative Scipy least_squares optimization method to help us validate the result.

In order to eliminate the sample bias of our parameters and sort out a way to optimize parameters from limited input data, we utilized bootstrapping technique in this project. The following is two distinctive ways of implementing bootstrapping. We run 40 trials of random bootstrapping by default.

Bootstrapping Methods

In this case, we tried two different methods for bootstrap. The first one is the residual bootstrap.

1. Compute $\hat{\beta} = (X^T X)^{-1} X^T y$ and $\hat{\varepsilon} \coloneqq (\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)^T = y - X\hat{\beta}$
2. With replacement, we randomly draw n samples from $\{\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n\}$ by assuming a probability $\frac{1}{n}$ from each $\hat{\varepsilon}_i$ Denote the bootstrapped residuals by $\varepsilon^* = (\varepsilon_1^*, \dots, \varepsilon_n^*)^T$
3. Compute $\hat{\beta}^* = (X^T X)^{-1} X^T y^*$
4. Repeat steps 2 through 3 multiple times

The second one is the paired bootstrap.

1. With replacement, randomly draw n pairs from the set $\{(y_i, {x_i}^T)\}_{i=1}^n$, by assuming a probability of $\frac{1}{n}$ for each pair. Denote the bootstrapped data by $X^*= (x_1^*, \dots, x_n^*)^T$ and $y^*= (y_1^*, \dots, y_n^*)^T$
2. Compute $\hat{\beta}^* = (X^T X)^{-1} X^T y^*$
3. Repeat steps 1 through 3 multiple times.

3) Result Analysis

1. With two different bootstraps methods, we find that
   For residual bootstrapping methods:

$\eta = 0.16459$

t-eta = 0.6654

$\boldsymbol{\beta}$ = 0.93338

t-beta = 2.60465

For paired bootstrapping methods:

$\eta = 0.0206$

t-eta = 0.16617

$\boldsymbol{\beta}$ = 0.98015

t-beta = 1.8835

We also performed t statistics analysis for the parameters. By computing the standard error of parameters from bootstrapping result, we can compute our t-statistic. From the results we can see, since the t statistics value is larger than 1, we can conclude that the parameters are significant under 90% significant level. Note that the result might vary greatly according to different choice on data-source, dates, cleaning threshold, and so on. More work has to be done in order to optimize the result.

2. In this part, we will perform normality checking for the residual values. From the statistics characteristics we can see that for the residual values:

   Mean: -0.09

   Variance: 1.823

   Skewness: -1.2098

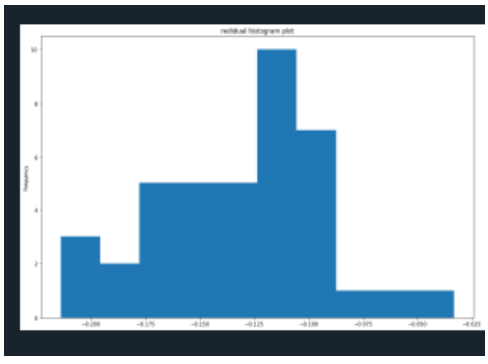
   Kurtosis: 9.245

   Jarque_bera P value: 0.0001

   From the statistics value we can see that the residual value follows a general pattern of normal distribution with a thin 'bell' with a high peak.

   Notice that from Jarque_bera P-value, the residual value seems not perform perfect normality. One of the potential reasons maybe that the sample size is small for valuation. In this case, we also plotted the histogram to show the results

   From the histogram, we can also see that the residual value loosely follow the normal distribution which is aligned with our assumption.



3. In this section, we break the stocks into two halves based on the liquidity level of the stocks.

From the data, we can see that in both cases, the RMSE and R-squared are on the same level. It indicates that the curve fits similar well for both cases.

For illiquid stocks:

$$\eta = 0.12798$$
$$\boldsymbol{\beta} = 0.93758$$

For liquid stocks:

$$\eta = 0.03006$$
$$\boldsymbol{\beta} = \text{-}0.44649$$

We can also notice that our parameters for illiquid stocks are higher than those for liquid stocks for stocks in between 20070907 and 20070920. This result might be biased because it is not very representative for all-time data. However, we can infer that less liquid stock will have higher temporary impact since higher η and $\boldsymbol{\beta}$ compared with those more liquid stocks. Our guess is that higher liquid stocks will recover more quickly from the temporary impact compared with illiquid stocks and they will trigger less momentum effect when people are selling off or buying up liquid stocks simply because they are more liquid.

```
less active stocks have following eta and beta:
(0.5, 0.5)
optimized parameters through scipy curve_fit is [0.12798503 0.93675774]
RMSE: 2.9843907161967462
R-squared: 0.001047102970382241

more active stocks have following eta and beta:
(0.5, 0.5)
optimized parameters through scipy curve_fit is [ 0.03005847 -0.44649536]
RMSE: 0.28328774319890254
R-squared: 0.0011292208090041278
```

4. In this part, we perform White's general test for heteroskedasticity in order to see if the variance of the residual is constant.

   The idea of the White's general test is based on the estimation of the relationship between residual value and the exogenous variables.

   $$\hat{\varepsilon}^2 = \alpha_0 X_1 + \cdots + \alpha_p X_p + \gamma_1 X_1^2 + \cdots \gamma_p X_p^2$$

   The function is in the form of:

   $$\hat{u}^2 = b_0 + b_1 \hat{y} + b_2 \hat{y}^2$$

   Null Hypothesis: $b_1 = b_2 = 0$

   In this case, we input the η and $\boldsymbol{\beta}$ with

   $$m = \sigma \eta sgn(X) \left| \frac{X}{V(6.5/6)} \right|^{\beta}$$

   From the results we LM and F-test is not statistically significant at 95% level. We failed to reject the null hypothesis. We can conclude that the residual is homoscedastic.

```
residual analysis
heterokasicity check
{'Test Statistics': -8.444469887976638, 'Test
Statistics P-Value': 1.0, 'F-Statistics':
-2.1008651965410188, 'F-test p-value': 1.0}
```

Problem 3(a)

(1) Pairs Trading

Summary: A trading strategy that is built up with holding a long position and a short position in two stocks with high positive correlation.

Mechanism: It is a market-neutral strategy that involves long and short positions in two different securities with high positive correlation. The two off setting positions form a basis that benefit from either a positive or negative trend.

Advantages and Disadvantages: The investors profit if the strategy performs as expected; the investors are also able to mitigate potential loss that might occur in the process. However, it is hard to identify two high positive correlation securities. Second, the correlation is calculated from historical data, but past data can't indicate future trends.

Comments: High frequency trading firm usually hold positions in a very short period of time, usually minutes or less, so, they need computers to find out potential pairs trading opportunities and execute for them.

References:

Pairs Trade

Getting up to Speed on High Frequency Trading

Pairs Trading – A Real-World Profitable Strategy

(2) **Latency Arbitrage**

Summary: A strategy through which high-frequency traders target differences between a stock's price on various markets.

Mechanism: The same asset might have different prices throughout different exchanges, leaving an arbitrage chance. This strategy takes advantage of this scenario in which buy asset at a lower price in one exchange and sell the same asset at a higher price in the other exchange.

Advantages and Disadvantages: HFT traders earn little profit per share with this strategy, but it could easily accumulate to a significant number. It also incurs little risk as the transactions are usually done in a very short time.

Comment: This strategy requires cutting-edge computers and algorithms to discover the arbitrage opportunities and to execute it before they disappear. It requires very little financial model as traders don't need to understand how prices change through time neither do they need to understand how it is priced. Instead, they just need to build less latency algorithms to identify these opportunities and act faster than other market participants.

Reference:

Latency Arbitrage

HIGH FREQUENCY TRADING: LATENCY ARBITRAGE ACCOUNTS FOR 33% OF SPREAD, SAYS BIS

What is Arbitrage? 3 Strategies to know

(3) Order Anticipation Strategy

Summary: A directional strategy that is designed to anticipate other's large orders and then to make a profit with this foreknowledge.

Mechanism: First high frequency traders detect the order patterns. Due to the fact that a large order is usually split into multiple small orders to prevent from leaking too much information to the market and causing market impact, it is hard to tell whether there is a huge liquidity demand on the market. One possible approach to solve this is to send numerous small orders. Once they are all filled, we could have

some confidence that there is a huge demand on the market. Knowing the existence of the demand, HFT traders can then exploit this opportunity with best conditions, like co-location and faster computers, to act faster than other traders. Then, they act as the counterparties as the other traders on the market. In this way, they can make a profit.

Comments: Some concerns of SEC are that whether this strategy could provide liquidity to the market and whether they are fair. Anticipation strategy doesn't provide extra liquidity to the market; instead, they prey on other's orders. It is fair, since traders take risks, and they don't trade with any non-public information. Thus, unlike front-running, SEC doesn't prohibit this strategy.

Reference:

Order anticipation

Market Data Patterns, Order Anticipation, and An Example Trading Strategy

(4) **ETF Arbitrage**

Summary: A strategy to earn profit with the mispricing between ETF and its underlying.

Mechanism: Thanks to the inefficiency of the market, there exists mispricing between assets. For example, if ETF value doesn't equal to the underlying values, HFT traders could buy the undervalued assets and sell the overpriced at the same time to make profit. Although this chance doesn't happen often and only exist for seconds if any, HFT firms have fasters computers and closer location to exchanges giving them better conditions to exploit this chance.

Comments: Some people might think latency arbitrage and ETF arbitrage are the same, but we argue that they are fundamentally different. The former one doesn't involve financial knowledge and pricing models. It only relies on technology. The latter one, however, requires traders to consider the price of assets. It bets on the probability there exists a difference and will converge eventually.

Reference:

How ETF Arbitrage Works

Problem 3(b)

Without a doubt these high frequency traders earn a lot. We can have a glimpse of how much they earn from Glassdoor, a social media focusing on company's reviews and statistics. Take Citadel securities, a renowned high frequency trading firm, a trader in the firm can make $200k/year in average and an additional payment $150k/year. Citadel securities is not alone, other HFT firms also pay extremely high salary to hire these talented traders.

According to Bloomberg, Citadel has a revenue approximately of $7 billion and an EBIT of 4.1 billion. According to Citadel securities website, they platform trades nearly 25 % of US equities trading volume and all HFTs account for approximately 50 - 60% of US equity trading volume. We assume that all HFTs earn nearly the same amount of money per trading volume, so the profit of HFTs should be around 8.2 billion to 9.6 billion.

A hedge fund using leverage will typically invest investors' capital and the borrowed money to make investments to increase the potential return of the fund. Leverage will magnify both the potential gain and potential loss from an investment. Citadel has leverage of about 8.8 times investment capital.

Reference:

Citadel securities trader salary

SEC Investor Bulletin from Office of Investor Education

Problem 3(c)

High Frequency Trading emerges when people starting to introduce algorithms into financial industry. It is believed to have some characteristics: (1) automated process for trading, (2) high speed in the submission of orders

and in the process of incoming information, (3) submission of high numbers of orders and quotes. Some may argue that HFT potentially provide more liquidity to the market and contribute to the reduction of bid-ask spread. However, we believe that it also imposes systemic risk to the market as discussed in the reference. We provide 3 systemic risks potentially caused by the speed, depth and frequency of HFT firms.

Adverse Selection in Orders and Market-Making

The first risk refers to adverse selection in orders describing a situation where market participants are affected by information asymmetric with HTF traders acting as "informed" traders. In other words, HFT market makers have the ability to process price information faster than non-HFT market makers. Non-HFT market makers can't keep up with the incoming information flow, so they will charge higher for the information risk they are facing. As the result, slower non-HFT market makers are crowded out from the markets. The ultimate dysfunctional and self-induced process of price formation, like the one observed on 6 May 2010.

Correlation and Herd Behavior

According to research, HFTs have common statistical patterns in their quotes and 78% of the trades made by them can be explained with only 3 common strategies, suggesting that HFTs have high correlation in their transactions. Additionally, the optimization of algorithms in HFT would remove the human interference in daily trading. These facts lead to two-fold systemic impacts. First, if a negative shock affects several HFT, it will lead to widespread failures of HFT given their interconnectedness and contagion. Second, the fact that HFT dominate the "active" trading could lead to over-reaction to new information. Jiang et al. (2014) find that HFT increase volatility in the US Treasury market during, shortly before and after the announcement of macroeconomic news. The over-reaction to new information could lead to herd behavior and higher probabilities of tail events or crashes.

Market Power and Barriers to Entry

This vulnerability stemming from the expansion of HFT in financial market refers to market power and barriers to entry. From the empirical observations, HFT usually account for a very large proportion of the orders posted in exchanges. Even though the orders are usually cancelled, they give HFT a privilege to determine the prices of securities. In other words, HFT could manipulate the price and cancel orders other than ones that are executed with the ideal price. The enormous investment in technology and complicated algorithms gives HFTs the advantages which slower market makers don't have. Moreover, the fictions in financial markets and the costly technology used by HFT create barriers to entry for new participants, leading to wider gap between HFT and non-HFT.

In sum, we do believe that HFT potentially impose a systemic risk to the market.

Problem 3(d)

We would like to propose an intraday Dynamic Pairs Trading use correlation and co-integration. It's based on market-neutral statistical arbitrage strategy, i.e., it doesn't matter whether the market is trending upward or downwards, the two positions for each stock hedge against each other.

Idea

The pairs trading strategy uses trading signals based on the regression residuals and were modeled as a mean-revering process.

Methodology

(1) correlation test

We should start testing the correlation (ρ) between each of two stocks as below. The higher the ρ is, the more positive relation of two stocks are.

$$\rho = \text{Correlation}(X, Y) = \frac{Cov(X, Y)}{SD(X) * SD(Y)}$$

If the correlation is high, say 0.8, traders may choose the pairs for pairs trading. So, if A stock goes up, the chances of B going up are also high. Based on this, a market neutral strategy is played where A is bought with 1 share and B is sold with n shares. Let's say the pairs trading strategy based on the spread between the prices of two stocks.

$$Spread = \log(A) - nlog(B)$$

Since we require our pairs to be mean reverting, we have to test co-integration of pairs.

(2) Co-integration test

In short, co-integration is a statistical measure of two or more time series variables used to indicate if a linear combination of the variables is stationary. Therefore, if A and B are cointegrated then it implies that the equation above is stationary. Given the property of stationary time series, expected spread = 0, whenever any deviation from this expectation is a case of statistical abnormality, hence a case for pairs trading.

(3) Choose pairs

For any pair of stocks, define the spread as above. Then, calculate 'n' using regression so that the spread is as close to 0 as possible. After that, we study the residuals in our regression model and if it has a trend, that means the spread is not stationary. We can also run Augmented Dicky Fuller test to make sure our variables are stationary or not.

(4) Entry Points and Exit Points

We define upper and lower thresholds as a time to enter and exit pairs trading. When we cross upper threshold, we go short: Sell A and Buy B; when we cross lower threshold, we go long: Buy A and Sell B. There is a chance that the spread doesn't revert back to mean, then we will set up a boundary to prevent from this which create further loss.

(5) Performance measurement

We may examine the performance of our strategy with Sharpe ratio for the period of our training dataset and compare it to market performance, say S&P 500. Also, we can test how different parameters (thresholds, confidence intervals) will affect the performance. In addition, we would use sectors from different time periods to stress test the performance of the strategy in different market situations.

Reference

Pairs Trading Basics