Empirical Methods in Finance

Project #1: Cointegration and Pair Trading

Due Friday April 14, 2025, at the beginning of the class

The goal of this project is to design a statistical arbitrage strategy by implementing pair trading. First, we investigate the characteristics of the available stocks by computing descriptive statistics. Second, we test the stationarity of log prices for each asset. Third, we test cointegration between each pair of assets in order to select the best pair for our pair-trading strategy. Finally, we implement the strategy in-sample and out-of-sample.

Please write a report to answer questions below, i.e., do not comment your code and do not use screenshots from your code. For each question, answer as **concisely and precisely** as possible. Your results must be clearly presented and commented.

Please follow the question numbering.

The only accepted format is a PDF file for your report, and py-files or m-files for your code. The code should be self-sufficient, i.e., the grader should be able to run your code without modification (except for the data path). Import the raw data file in your code (do no change the data file in excel). If you use external libraries, it is your responsibility to understand their commands. Projects provided after the deadline will not be considered.

Personal advice: Do not wait until the week before the deadline to start working on your project. As you can see, there is quite a lot of work to do. If you wait until the deadline, you will not have time to hand in a proper assignment.

Data

You have been assigned a dataset on Moodle. You can download the corresponding csv-file and the list of symbols on Moodle. For each asset, you are given adjusted prices (open and close), volumes, and highs and lows.

1 Descriptive Statistics

Compute the daily and weekly simple return and the daily and weekly continuously compounded (or log) return using the adj. close price. For both definitions of returns, compute the annualized sample mean, annualized variance, skewness, kurtosis, minimum, and maximum. Compute the weekly returns from Monday to Monday.

Q1.1 Compare simple return and log-returns in daily frequency. (2.5 points)

Q1.2 Compare simple return and log-returns in weekly frequency. (2 points)

Q1.3 How are the above descriptive statistics changed when you change the frequency from daily to weekly? (Consider log-returns only). (1 point) **(Total: 5.5 points)**

2 Stationarity

To test for stationarity, we use the Dickey-Fuller test. To do so, we run the following regression for each series:

$$p_t = \mu + \phi p_{t-1} + \varepsilon_t, \tag{1}$$

where $p_t = \log(P_t)$ denotes the log-price.

Q2.1 Define the null and alternative hypotheses and explain the intuition between the null hypothesis. (1 point)

Q2.2 Explain the decision rule, i.e., how you compute the test-statistic and decide to reject or accept the null hypothesis. (1 point)

Q2.3 If we ran the following regression $r_t = \mu + \phi p_{t-1} + \varepsilon_t$, with $r_t = p_t - p_{t-1}$, how would we test the stationarity of log-prices? Show that this is equivalent to the previously defined null hypothesis. (1 point) **(Total: 3 points)**

2.1 Critical Values

The critical values provided by Fuller (1976) are -2.58 at 10%, -2.89 at 5% and -3.51 at 1%.[1] We would like to compute our own critical values for our sample size using Monte-Carlo simulations. The idea is to perform a large number of replications (take $N = 10,000$) of the following experiment i:

1. Simulate a time series of T error terms $\varepsilon_t^{(i)}$, $t = 1, \ldots, T$ distributed as $N(0, 1)$. T is the length of your series.

2. Compute a time series of prices, assuming that they are driven by a random walk $p_t^{(i)} = p_{t-1}^{(i)} + \varepsilon_t^{(i)}$.

3. Estimate the AR(1) model $p_t^{(i)} = \mu^{(i)} + \phi^{(i)} p_{t-1}^{(i)} + \varepsilon_t^{(i)}$.

4. Compute the test statistic $t(\phi^{(i)} - 1)$ for the null hypothesis of the DF test.

Repeat this experiment N times to obtain the distribution of $t(\phi^{(i)} - 1)$. The critical values correspond to the quantiles at 10%, 5%, and 1% of the distribution $t(\phi^{(i)} - 1)$.

Q2.4 Why do we simulate random walks? (0.5 point)

[1] In Fuller (1976), the sample size is $T = 100$.

Q2.5 Plot an histogram for the N values of $t(\phi^{(i)} - 1)$, $i =, 1 \ldots, N$. What do you observe? What does the distribution of $t(\phi^{(i)} - 1)$ represent? (2 points)

Q2.6 Compute the critical values of the DF test. (0.5 point)

Q2.7 Assume we redo the simulations, but this time we simulate the following AR(1) process $p_t^{(i)} = 0.2\ p_{t-1}^{(i)} + \varepsilon_t^{(i)}$. What would be the distribution of $t(\phi^{(i)} - 1)$? (1 point) **(Total: 4 points)**

Redo the simulation for $T = 500$, as we will need the corresponding critical values later in the project.

2.2 Testing Non-stationarity

For each asset, test the null that the log-price has a unit root. That is, you have to run the regression given by Equation (1).

Q2.8 Compute the test statistic and carry out the DF test (DF statistic, critical value, p-value, reject or not reject the null hypothesis). To compute the p-value, use the distribution $t(\phi^{(i)} - 1)$ plotted in Q2.5. What is your conclusion? (2 points)

Q2.9 What do your results imply regarding cointegration? (0.5 point) **(Total: 2.5 points)**

3 Cointegration

For the pair-trading strategy, we need to find pairs of assets that are cointegrated. Therefore, we would like to test for cointegration for all the possible pairs in our dataset.

To test for cointegration between for the pair (A-B), we proceed as follows:

First, we estimate their relationship by running a regression between their contemporaneous log-prices:

$$p_t^A = \alpha + \beta p_t^B + z_t$$

Then, we use the Dickey-Fuller test for testing the null of unit root in z_t. So we estimate the regression

$$\Delta \hat{z}_t = \mu + \phi \hat{z}_{t-1} + \varepsilon_t$$

3.1 Critical Values

The critical values for this test with $T = 100$ are -3.07 at 10%, -3.37 at 5%, and -3.96 at 1% provided by Phillips and Ouliaris (1988). As before, we would like to compute our own critical values for our sample size. We run the following simulation exercise:

1. Simulate two time series of independent random walks: $p_t^{A(i)} = p_{t-1}^{A(i)} + \varepsilon_t^{A(i)}$ and $p_t^{B(i)} = p_{t-1}^{B(i)} + \varepsilon_t^{B(i)}$ for $t = 1, ..., T$. Error terms $\varepsilon_t^{A(i)}$ and $\varepsilon_t^{B(i)}$, $t = 1, \ldots, T$, are distributed as independent $N(0, 1)$.
2. Estimate the linear relationship between the two time series (p_t^A, p_t^B):

   $$p_t^{A(i)} = \alpha + \beta p_t^{B(i)} + z_t^{(i)}$$

   Under the null of no cointegration, the residual series $\hat{z}_t^{(i)}$ should be non-stationary. We compute the DF test statistic on $\hat{z}_t^{(i)}$.
3. Estimate the AR(1) model for the residuals, under the alternative hypothesis, i.e., $\Delta\hat{z}_t^{(i)} = \mu^{(i)} + \phi^{(i)}\hat{z}_{t-1}^{(i)} + \varepsilon_t^{(i)}$. Compute the t-stat for $\phi^{(i)}$, denoted by $t(\phi^{(i)})$.

Repeat this experiment N times to obtain the distribution of $t(\phi^{(i)})$. Since we simulated independent random walks, we obtain the distribution under no-cointegration. The critical values correspond to the quantiles at 10%, 5%, and 1% of the distribution $t(\phi^{(i)})$. Redo the simulation for $T = 500$ to compute critical values, which we will use later in the project.

Q3.1 Plot an histogram of the distribution of $t(\phi^{(i)})$ and report the critical values. What do you observe? (2 points) **(Total: 2 points)**

3.2 Testing for Cointegration

Test for cointegration on each pair of assets (in both directions, i.e., $A \rightarrow B$ and $B \rightarrow A$).

Q3.2 Report the results of the cointegration tests. Report the test statistic, p-values, and your conclusion. Clearly report for each pair whether you found cointegration or not. (2 points)

Q3.3 Report the parameters estimates $\hat{\alpha}$ and $\hat{\beta}$. Comment. (1 point)

Q3.4 Which pair is the most *strongly* cointegrated and why? In one sentence, explain the economic intuition behind this result. You will use this pair for our pair-trading strategy. (1 point)

Q3.5 Plot the time series of the prices of this pair of assets on the same graph. (0.5 point) **(Total: 4.5 points)**

4 Pair Trading

From the cointegration test we selected the pair assets for our pair-trading strategy. To make money, we aim at exploiting statistical arbitrage underlying their cointegration relationship with a pair-trading strategy. The first step is to define the spread between the prices that we will use as signal for trading.

4.1 Trading Signal

Let the pair of assets (A, B) be cointegrated :

$$P_t^A = \alpha + \beta P_t^B + z_t, \tag{2}$$

where P_t^A is the price of asset A and P_t^A the price of asset B. We use prices instead of log-prices to simplify the construction of the pair-trading strategy.

We define the spread z_t as:

$$z_t = P_t^A - \alpha - \beta P_t^B \tag{3}$$

The spread (or signal) is normalized as $\tilde{z} = z_t/\sigma(z_t)$.

Q4.1 Given that (A, B) are cointegrated, what is the main property of z_t? If $z_t >> 0$ what does it tell you about the price of the asset A with respect to the price of the asset B? Elaborate on statistical arbitrage. (1.5 points)

Q4.2 Compute the sample $\tilde{z}_t$ using the adjusted prices at close and plot it. (2 points)

Q4.3 We are interested in the auto-correlation of $\tilde{z}_t$. Plot the auto-correlogram up to 10 lags with the confidence interval and run a Ljung-Box test with 10 lags. What do you observe? What does it imply for the pair-trading strategy? (2 points) **(Total: 5.5 points)**

We opt for the following algorithmic trading strategy:

- **Signal 1**: if $\tilde{z}_t > \tilde{z}^{in}$ → short position for A and long position for B. Close positions when $\tilde{z}_t \leq 0$. We take position $Q_1 = \begin{bmatrix} -1 \\ \beta \end{bmatrix}$.
- **Signal 2**: if $\tilde{z}_t < -\tilde{z}^{in}$ → short position for B and long position for A. Close positions when $\tilde{z}_t \geq 0$. We take position $Q_2 = \begin{bmatrix} 1 \\ -\beta \end{bmatrix}$.

Q4.4 Show that this is not a self-financing strategy. (1 point)

Q4.5 Assume α and β are known and we can trade without delay and frictions. When $\tilde{z}_t = \tilde{z}^{in}$, we short-sell 1 unit of asset A and buy β unit of asset B. Show that the profit is equal to $\tilde{z}^{in}\sigma(z_t)$ when the position is closed at $\tilde{z} = 0$. (1.5 points)

Q4.6 To implement this strategy, we have to choose a $\tilde{z}^{in}$. Explain the trade-off underlying this choice. (1 point) **(Total: 3.5 points)**

4.2 Pair-trading Strategy

As an MScF student, you now want to implement this strategy. You start with an initial wealth $W_0 = \$1000$ that you deposit on a trading platform. At the end of each day, you compute the signal $\tilde{z}_t$ using the ajdusted prices at close. If the observed signal requires a trade, you execute it the next morning at the adjusted price at open.[2] Your trading platform requires you to deposit an initial margin of 50% on your short position. That is, if you short-sell 10 units of asset A with $P^A = \$100$, the platform keeps in custody \$500. So, with your initial wealth, you can short-sell a maximum of $Q^A = 20$ units of asset A. In other words, the maximum leverage is 2, with leverage as $L = P^A \times Q^A / W$.

4.2.1 Direct Strategy

Remember that our strategy imposes that for one unit short (long) of asset A, we long (short) β unit of asset B. Therefore, if $\alpha < 0$ and we short 1 unit of asset A, we need to invest α of our money to buy β units of asset B. Because of the leverage constraint on the short position, the allocation changes if α is positive or negative.

As result, for a given wealth W, the number of units $Q = (Q^A, Q^B)$ we buy is as follows:

- if $\alpha > 0$
  - For Signal 1, the allocation is

$$Q_1 = \frac{L \times W}{P_t^A} \begin{bmatrix} -1 \\ \beta \end{bmatrix}$$

  - For Signal 2, the allocation is

$$Q_2 = \frac{L \times W}{\beta P_t^B + L \times (P_t^A - \beta P_t^B)} \begin{bmatrix} 1 \\ -\beta \end{bmatrix}$$

- if $\alpha < 0$
  - For Signal 1, the allocation is

$$Q_1 = \frac{L \times W}{P_t^A - L(P_t^A - \beta P_t^B)} \begin{bmatrix} -1 \\ \beta \end{bmatrix}$$

  - For Signal 2, the allocation is

$$Q_2 = \frac{L \times W}{\beta P_t^B} \begin{bmatrix} 1 \\ -\beta \end{bmatrix}$$

Q4.7 Implement the strategy of the investor described above with $\tilde{z}^{in} = 1.5$. What is the profit? Report the final wealth, the largest and the lowest wealth level, and the number of trades. Plot the signal $\tilde{z}_t$, the evolution of wealth, the positions, and the leverage. (5 points)

Q4.8 Redo the last point with a maximum leverage, L, equal to 20. What do you observe? (1.5 points) **(Total: 6.5 points)**

[2] To keep things simple, we assume you can trade at the adjusted prices. In reality, you trade at the market price and you should keep track of the dividends.

4.2.2 Stop Loss

With this levered trading strategy, we are exposed to losses (temporary or not). Therefore, we want to implement a stop loss, i.e., we want to define a rule such that we close the positions if we are losing too much money:

- For Signal 1, close the positions if $\tilde{z}_t > \tilde{z}^{stop}$
- For Signal 2, close the positions if $\tilde{z}_t < -\tilde{z}^{stop}$

Q4.9 Explain in one sentence the logic behind the stop-loss rule. Assume we use $\tilde{z}^{in} = 1.5$ and $\tilde{z}^{stop} = 1.75$. We are interested in measuring the probability of hitting the $\tilde{z}^{stop}$ the day after opening the position at $\tilde{z}^{in} = 1.5$. Since $\tilde{z}_t$ is autocorrelated, we use an AR(1) model. Estimate an AR(1) model and use the conditional distribution to compute $\Pr(\tilde{z}_{t+1} > \tilde{z}^{stop})$. (3 point)

Q4.10 Implement the pair-trading strategy with a $\tilde{z}^{stop} = 2.75$. Report the same metrics as in Q4.7. Compare the two strategies. What is your conclusion? (4 points) **(Total: 7 points)**

4.3 Out-of-sample Pair-trading Strategy

So far, we estimated the parameters of the relationship $P_t^A = \alpha + \beta P_t^B$ over the full sample. This raises two issues. First, this in-sample approach cannot be implemented in real time. Second, the parameters can be in fact time-varying, i.e., α_t and β_t.

To address these issues, we use a rolling-window to estimate the parameters:

1. Estimate the parameters α and β over the first two years of data (use 500 observations).
2. Compute the signal $\tilde{z}_t$ for the following next 20 days with the estimated parameters. To normalize z_t, use the standard deviation estimated on the estimation period. This is the out-of-sample spread.
3. Roll the window by 20 days (e.g., observations 21 to 521), estimate the parameters.
4. Repeat the last two points until you have covered the whole sample.

Q4.11 Using this rolling window, compute the rolling correlation between the prices of the pair of assets. Also compute the rolling correlation between the returns of the pair of assets. Plot both correlation series and comment. (1 point)

Q4.12 Using this rolling window, estimate the parameters $\hat{\alpha}_t$ and $\hat{\beta}_t$ on each subsample. Plot the dynamics of the parameters estimates, and compare them to the estimates based on the full sample. Plot the in-sample spread and the out-of-sample spread together. What do you observe? (2.5 points)

Q4.13 Using the out-of-sample spread, apply the pair-trading strategy (use the same values of z^{in} and z^{stop} as before). As before, report the performance of the strategy. Comment on the use of a rolling window. (2 points)

Q4.14 Using this rolling window, test for cointegration for each subsample (the sample size is 500). Plot the p-values with a stem graph. What do you observe? Do you find cointegration over all the subsamples? What does it imply for our pair-trading strategy? (2 points)

Q4.15 To take into account that the cointegration relationship might break, we adjust our strategy by closing and not trading when there is no cointegration. That is, if we find no cointegration on a subsample, we close our position if open and do not trade until we find cointegration again. Implement this strategy and as usual report the performance. Comment. (2.5 points)

Q4.16 Write a **short** conclusion about statistical arbitrage and pair-trading. Is there anything else we should take into account before opening our hedge fund? (2 points) **(Total: 12 points)**