Empirical Methods in Finance

Project #2: Cointegration and Pair Trading

Due Friday April 21, 2023, at the beginning of the class

The goal of this project is to design a statistical arbitrage strategy by implementing pair trading. First, you test the stationarity of log prices for each asset. Second, you test cointegration between each pair of assets in order to select the best pair for your pair-trading strategy. Finally, you implement the strategy in-sample and out-of-sample.

For each question, answer as **concisely and precisely** as possible. Your results must be clearly presented (no screenshots from your code!) 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. ot submit an mlx or jupyter notebook with comments as a report, write a proper report file. 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.

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 Stationarity

To test for stationarity, you use the Dickey-Fuller test. To do so, you 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.

Q1.1 Define the null and alternative hypotheses and explain the intuition between the null hypothesis.

Q1.2 Explain the decision rule, i.e., how you compute the test-statistic and decide to reject or accept the null hypothesis.

1.1 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 of the DF test correspond to the quantiles at 10%, 5%, and 1% of the distribution $t(\phi^{(i)} - 1)$.

Q1.3 Why do you simulate a random walk?

Q1.4 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?

Q1.5 Compute the critical values of the DF test.

Q1.6 Redo the simulation for $T = 500$, as you will need the corresponding critical values later in the project.

1.2 Testing Non-stationarity

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

Q1.7 Compute the test statistic and carry out the DF test (DF statistic, critical value, p-value, reject or not reject the null hypothesis) for daily log-prices. To compute the p-value, use the distribution $t(\phi^{(i)} - 1)$ plotted in Q1.4. What is your conclusion?

Q1.8 What do your results imply regarding cointegration?

---

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

2 Cointegration

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

To test for cointegration between for the pair (A-B), you 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, you use the Dickey-Fuller test for testing the null of unit root in z_t. So you estimate the regression

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

2.1 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, you compute your own critical values for our sample size. You 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. You 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., $\Delta \hat{z}_t^{(i)} = \mu^{(i)} + (\phi^{(i)} - 1)\hat{z}_{t-1}^{(i)} + \varepsilon_t^{(i)}$. Compute the t-stat for $(\phi^{(i)} - 1)$, denoted by $t(\phi^{(i)} - 1)$.

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

Q2.1 Plot an histogram of the distribution of $t(\phi^{(i)} - 1)$ and report the critical values. What do you observe?

2.2 Testing for Cointegration

You now test for cointegration on each pair of assets (in both directions, i.e., $A \to B$ and $B \to A$).

Q2.2 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.

Q2.3 Report the parameters estimates $\hat{\alpha}$ and $\hat{\beta}$. Comment.

Q2.4 Which pair is the most *strongly* cointegrated and why? In one sentence, explain the econonomic intuition behind this result. You will use this pair for our pair-trading strategy.

Q2.5 Plot the time series of the prices of this pair of assets on the same graph (display p_t^A and $\alpha + \beta p_t^B$).

---

3 Pair Trading

From the cointegration test, you selected the pair of assets for our pair-trading strategy. You aim at exploiting the statistical arbitrage opportunity underlying the cointegration relationship with a pair-trading strategy.

To implement a pair-trading strategy, you need to consider how a typical hedge fund works. Assume that you start with an initial wealth $W_0 = \$1000$ that you deposit on a trading platform (or on your prime broker account). Assume that you do not use debt to finance your long positions. Then, the financing of the strategy works as follows:

1. The fund short-sells an amount $S = W \times L/(1+L)$ of asset A by deposit an amount $W/(1+L)$ of its equity on its broker account. The proceeds $W/(1+L)+W\times L/(1+L) = W$ are segregated on the broker account (Equity short account) and cannot be used for other purposes.

2. The fund invests the equity left (Equity long account), $W \times L/(1+L)$ to buy a long position in asset B.

3. In the end, the fund has invested the same amount of short in asset A and long in asset B. Its overall gross leverage is $2L/(1+L)$.

The fund's balance sheet looks as follows:

Assets		**Liabilities and Equity**	
Cash proceeds	W	Short securities	$Short = W \times L/(1+L)$
Long securities	$Long = W \times L/(1+L)$	Equity	N_t^H
		- Long account	$W \times L/(1+L)$
		- Short account	$W/(1+L)$

To fix ideas, assume $W = 1000$ and $L = 2$.

3.1 Trading Signal

The first step is to define the spread between the prices that you will use as signal for trading. 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^B the price of asset B. You use prices instead of log-prices to simplify the construction of the pair-trading strategy. Of course, the logic of the cointegration remains the same.

The spread z_t is defined as:

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

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

Q3.1 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.

Q3.2 Compute the sample $\tilde{z}_t$ using the adjusted prices at close and plot it.

Q3.3 Plot the auto-correlogram of $\tilde{z}_t$ 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?

3.2 In-sample Pair-trading Strategy

3.2.1 Direct Strategy

You first opt for the following algorithmic trading strategy:

- **Signal 1**: if $\tilde{z}_t > \tilde{z}^{in}$ → you enter short position for asset A and a long position for asset B. The number of shares of asset A that you sell (N_t^A) is determined by $P_t^A N_t^A = W_t \times L/(1+L)$. The number of shares of asset B that you buy (N_t^B) is determined by $P_t^B N_t^B = W_t \times L/(1+L)$. You close positions when $\tilde{z}_t \leq 0$.
- **Signal 2**: if $\tilde{z}_t < -\tilde{z}^{in}$ → you enter a short position for asset B and a long position for asset A. You close positions when $\tilde{z}_t \geq 0$.

Q3.4 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.

Q3.5 Redo the last point with a maximum leverage, L, equal to 20. What do you observe?

3.2.2 Stop Loss

With this levered trading strategy, you are exposed to losses (temporary or not). Therefore, you want to implement a stop loss, i.e., you want to define a rule such that you close the positions if you 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 $\tilde{z}_t < -\tilde{z}^{stop}$

Q3.6 Explain in one sentence the logic behind the stop-loss rule. Assume $\tilde{z}^{in} = 1.5$ and $\tilde{z}^{stop} = 1.75$. You 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 $\Pr(\tilde{z}_{t+1} > \tilde{z}^{stop})$.

Q3.7 Implement the pair-trading strategy with a $\tilde{z}^{stop} = 2.75$. Report the same metrics as in Q3.4. Compare the two strategies. What is your conclusion?

3.3 Out-of-sample Pair-trading Strategy

So far, you 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, you 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.

Q3.8 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.

Q3.9 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?

Q3.10 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.

Q3.11 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?

Q3.12 To take into account that the cointegration relationship might break, you adjust our strategy by closing and not trading when there is no cointegration. That is, if you find no cointegration on a subsample, you close our position if open and do not trade until you find cointegration again. Implement this strategy and as usual report the performance. Comment.

Q3.13 Write a **short** conclusion about statistical arbitrage and pair-trading. Is there anything else you should take into account before opening your hedge fund?
