MATH 60210: Homework #3

Professor Anthony Sanford

Due: April 8, 2024 by 11:55pm on ZoneCours

Instructions: For this assignment, you are to work in groups of **three**. I will **not** make the groups. It is your responsibility to find team members for your homework. If you have trouble finding a team, you can reach out to me via email and I will try to match you with other classmates. This assignment includes two deliverables: 1) your write-up and 2) your code. Failure to submit one of these **two** components will be considered an incomplete submission. Your code must by written in Python. Your code should run smoothly and be clearly annotated. If you used resources outside of what is assigned for this course, you **must** give credit to the source. For example, if you googled code and used whatever you found, reference the source. Otherwise, you are cheating. Your write-up is meant to analyze, discuss, and interpret your results. Simply submitting a table without clearly discussing that table is not an answer to the question. Consider what I do in class – I present something and discuss it. Think of your write-up in this exact way. You are interpreting your findings for someone who is reading your report. Recall that HEC Montréal has rules regarding plagiarism apply to both your written answers and your computer code.

There are 3 multi-part questions, points for each question/part is written below. For each part, there are four possible grades:

- E: This is the grade if you literally write nothing. Worth 0%
- C: This is the grade if you really don't understand what you're doing...but you wrote something. Worth 50%.
- B: This is the grade if you got the answer mostly correct. Worth 80%. This will likely be the most common grade.
- A: This is the grade if your answer is as good (or better!) than mine. Worth 100%.

Note: Comments in your code that allow me to easily understand what you were trying to do may improve your grade. Particularly ugly or inelegant python code may reduce your grade by 10% (e.g. if you're cutting and pasting a lot because you don't know how to use loops or functions or array operations effectively). You may also lose points if your code is not running correctly or does not provide the same solution as your write-up answer.

Problem 1 [40 Pts], Simulating Forecast Uncertainty

Use the PJM.csv dataset for this question.

a. Write a function that accepts [10 pts]:

- a $p \times 1$ array $s = [s_{t-p+1}, \ldots, s_t]$
- a $(1+p) \times 1$ array $phi = [phi_0, \ldots, phi_p]$

and returns:

- a real scalar containing the 1-period-ahead forecast $\hat{s}_{t+1}$ from the AR(p) model with parameters phi.

Report and interpret the results.

b. Write a function that accepts [10 pts]:

- a a $T \times 1$ array $s = [s_1, \ldots, s_T]$
- an integer p

and returns:

- a $(p+1) \times 1$ array phi containing the estimated coefficients $[\phi_0, \ldots, \phi_p]$ of an ARp model for s
- a $(T-p) \times 1$ array of residuals $[\hat{e}_t]$ for the fitted ARp model.

Report and interpret the results.

c. Write a function that accepts [10 pts]:

- a BitGenerator bg
- a $p \times 1$ array $s = [s_{t-p+1}, \ldots, s_t]$
- a $(1+p) \times 1$ array $phi = [phi_0, \ldots, phi_p]$
- a $T \times 1$ array e
- an integer $h > 0$

and returns:

- $MySim$, an $h \times 1$ array of simulated values for $[s_{t+1}, \ldots, s_{t+h}]$ based on the ARp parameters phi and bootstrapped values from e.

Report and interpret the results.

d. Write a function that accepts [5 pts]:

- a BitGenerator bg

- a $T \times 1$ array $s = [s_1, \ldots, s_T]$
- three integers $p, h, N > 0$

and returns:

- an $N \times h$ array of simulated values for $s_{T+1}, \ldots, s_{T+h}$ based on an estimated AR(p) model for s and bootstrapped forecast errors.

Report and interpret the results.

e. Using your answer to part (d), plot the 90% confidence interval for electricity prices for the next 20 hours based on an AR(24) model and 1,000 simulations. [5 pts]:

Problem 2 [60 Pts], Unit roots & Cointegration

For this problem, you will useUK_bonds.parquet containing daily yields on UK Treasury bonds. You are also given a data set S&P 500 Historical Data.csv containing values of the S&P500 Index. In addition to the ADF t statistic that we talked about in class, Dickey and Fuller also proposed a simpler 'normalized bias' statistic $\delta \equiv T \cdot \hat{\beta}$ where

- T is the number of observations used to estimate $\hat{\beta}$
- $\hat{\beta}$ is the estimated coefficient β from one of the ADF regressions that we saw in class

$$\Delta y_t = \alpha + \beta \cdot y_{t-1} + \sum_{i=1}^{p} \phi_i \cdot \Delta y_{t-i} + v_t or \Delta y_t = \alpha + \delta \cdot t + \beta \cdot y_{t-1} + \sum_{i=1}^{p} \phi_i \cdot \Delta y_{t-i} + v_t$$

In this question, you'll explore the behaviour of both statistics.

a. Start this question by providing descriptive statistics and time-series graphs for your data. Make sure to label your graphs correctly and discuss the results (both the descriptive statistics and the graphs).[10 pts]

b. Write a function that accepts [10 pts]:

- a series s
- an integer p

and returns:

- the ADF t statistic with p lags for H_A : stationary
- the ADF t statistic with p lags for H_A : trend-stationary
- the ADF δ statistic with p lags for H_A : stationary
- the ADF δ statistic with p lags for H_A : trend-stationary

Report the results in this order and interpret the results.

c. Write a function that accepts [10 pts]:

- a series s
- integers N , T, p

and then

- simulates N random walks, each of length T by bootstrapping from Δs
- runs your 'ADFuller(s,p)' function on each of the N random walks

and then returns:

- a _DataFrame_ whose columns are the 4 ADF test statistics *inorder* from part a and whose rows are the 90%, 95% and 99% critical values for the $H_0 : s$ is $I(1)$.

Report and interpret the results.

d. For this question, do the following [10 pts]:

- Using data on the last 400 observations of the 'SP500' index for s , use your function from part b to estimate critical values for the 4 test statistics. $Usep = 3, T = 400, N = 500$. Show your results.
- Compare them to the results from the autotest above for part b. Do any of the distributions of the test statistics seem to have shifted? If so, what might cause this?

e. Write a function that accepts [10 pts]:

- a $T1$ series s
- integers p and $N < T$

and runs your 'ADFuller(s, p)' function on a rolling window of N observations of s. It should then return:

- a 3×4 dataframe whose
  - columns are the 4 tests statistics (in order)
  - first row is the most significant value found for each of your 4 tests statistics
  - second and third rows are start and end date of window for which those most significant statistics were found

Report and interpret the results.

f. For this question, Use your 'rolling_ADF(s, p, N)' function to test the last 1500 observation of the log of the 'S&P500' index using $p = 3$ and $N = 250$ to answer the following questions [10 pts]:

- Where do you find the strongest rejection of the unit root null hypothesis?
- Do you think that the US stock market is sometimes I0?

Problem 3 [10 Pts], Practice for final exam; True/False/Unknown

For this question, you will answer **ONE** of the following: true/false/unable to determine. For these questions, to get full credit, you must 1) say whether the statement is true, false, or you cannot determine based on the information provided and 2) provide justification for your answer. Under no circumstances should you answer a mixture of these answers. If the answer is true, it must always be true. Otherwise, it is false. If you do not have enough information to make a conclusion, state why you cannot determine if the statement is true or false and what information you would need in order determine if the statement is true or false.

1. In order to determine the efficacy of our forecast model, we can estimate a test statistic of the forecast errors for a simple hypothesis test. [2.5 pts]

2. Your boss, who attended graduate school, wants you to conduct a forecast and validate your forecast using the k-fold cross-validation methodology. He claims that this is the best method for balancing the bias-variance tradeoff. [2.5 pts]

3. Nonparametric regressions have the advantage of decreasing bias while parametric regressions decrease variance. [2.5 pts]

4. Your new boss states that, after conducting a hypothesis test, she can conclude that a time-series is I(1) because she was able to reject the null in the Kwiatkowski–Phillips–Schmidt–Shin test and she also failed to reject the null in the Phillips-Perron test. [2.5 pts]