Week 2
Assignment 2 (Algo)

Gurnoor Sohal & Shreshtha Gupta

December 16, 2025

Submission Requirements

Although, we prefer Jupyter notebook submissions (use markdown for theoretical answers), you can also submit your Python (or MATLAB, etc.) code with screenshots (or in any other form of your choice) of the console output and answers to your theory parts as a separate pdf. Once again, when you are done, upload your submissions to your fork.

Exercise 1

Suppose we have a data set with five predictors, $X_1 = \text{GPA}$, $X_2 = \text{IQ}$, $X_3 = \text{Gender}$ (1 for Female, 0 for Male), $X_4 = \text{Interaction between GPA and IQ}$, and $X_5 = \text{Interaction between GPA and Gender}$. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$, $\hat{\beta}_5 = -10$.

1. Which answer is correct, and why?
   (a) For a fixed value of IQ and GPA, males earn more on average than females.
   (b) For a fixed value of IQ and GPA, females earn more on average than males.
   (c) For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.
   (d) For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.
2. Predict the salary of a female with IQ of 110 and a GPA of 4.0.
3. True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

Exercise 2

The Auto dataset can be downloaded from the Week2 folder of the main project repository or directly from the statsmodels library.

1. Use the `lm()` function to perform a simple linear regression with `mpg` as the response and `horsepower` as the predictor. Use the `summary()` function to print the results. Comment on the output. For example:
   (a) Is there a relationship between the predictor and the response? Comment.
   (b) What is the predicted `mpg` associated with a `horsepower` of 98? What are the associated 95% confidence and prediction intervals?
2. Plot the response and the predictor. Use the `abline()` function to display the least squares regression line.
3. Plot the diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

Exercise 3

Before attempting complex pairs trading, it is essential to understand the behavior of individual assets. You are analyzing the currency pair **USD.CAD** (United States Dollar v/s Canadian Dollar). You have to determine if the raw price series of USD.CAD is stationary enough to be traded with a simple mean-reversion model *(data is available in .csv format in the Week2 folder of the main project repository).*

Tasks

1. Plot the daily closing price series of USD.CAD. Does the plot look like a flat noise signal or a random walk? Does it have a clear mean?
2. Perform an Augmented Dickey-Fuller (ADF) test on the log-prices of USD.CAD (use a lag of 1). Report the t-statistic, the p-value, and the critical value at 90% confidence. Is the series stationary?
3. Regardless of the ADF test result, calculate the half-life (in days) of mean reversion using linear regression on the daily price changes:

$$\Delta y(t) = \lambda y(t-1) + \mu + \epsilon(t)$$

   Even if the strategy were theoretically stationary, would a half-life of this magnitude be practical for a short-term trader? Why/ why not?

Exercise 4

In this exercise, you will be investigating a potential mean-reversion strategy between two country-specific ETFs, **EWA (Auz) and EWC (Canada)** *(data is available in .csv format in the Week2 folder of the main project repository).* Perform a statistical analysis to determine if these two assets form a stationary, tradeable pair. You are expected to implement the mathematical models for cointegration and mean reversion either from scratch or using standard libraries.

Tasks

1. Generate a plot containing the daily closing prices of both ETFs. Calculate the correlation coefficient of their price series.
2. Run an Ordinary Least Squares (OLS) linear regression to model the relationship between the two assets. What was the resulting Hedge Ratio (β) for your choice of variables?
3. Using the β derived in the previous step, calculate the daily value of the spread portfolio
4. Perform an ADF test on the spread to check for stationarity.
    - Use a lag of 1 and assume zero drift
    - Use $\alpha = 0.95$ to state your conclusions
5. Assuming the series is mean-reverting, calculate the half-life of the mean reversion