Assignment 1: Linear Regression and FDR

Language: Python

Due: Thursday 11:59 **AM**

Data: This dataset is a playground for fundamental and technical analysis. It is said that 30% of traffic on stocks is already generated by machines, can trading be fully automated? If not, there is still a lot to learn from historical data.

fundamentals.csv: metrics extracted from annual SEC 10K fillings (2012-2016), should be enough to derive most of popular fundamental indicators. The data has 1781 observations and 78 attributes.

1. Data Exploration and Visualization: Explore the "fundamentals.csv". Include any other plots you find interesting. (10 pts)
2. Linear Regression Model Development: Create linear regression to predict Estimated Shares Outstanding. Explain your model. (15 pts)
3. Multicollinearity in Linear Regression: Explain how multicollinearity can affect the interpretation of a linear regression model's coefficients. (Written) (10 pts)
4. P-Value Analysis and Histogram: Create a histogram of the p-values. Is there any skewedness? Provide your explanation. (10 pts)
5. False Discovery Rate Control with BH Procedure: Given the p values you find, use the BH procedure to control the FDR with a q of 0.1. How many "true" discoveries do you estimate? (15 pts)
6. Sensitivity Analysis of FDR Control: If you apply the BH procedure at different q values, how do the results change? What does this tell you about the robustness of your significant variables? (10 pts)

7. Exploring Interaction Terms: (10pts, 5 pts for each)

   a. Expand your linear regression model by adding interaction terms. Create interaction terms between pairs of predictors (up to quadratic terms, i.e., terms of power two). You should include both original predictors and their interaction terms in your model.

   b. Briefly explain why interaction terms might be important in the context of predicting Estimated Shares Outstanding using fundamental financial metrics.

8. Model Evaluation with Interaction Terms: (10pts, 5 pts for each)

   c. Evaluate the performance of this new model with interaction terms. Compare it with the performance of the original model without interaction terms using appropriate metrics.

   d. Discuss any significant changes in the model's performance or the coefficients of the predictors.

9. FDR Analysis with Interaction Terms: (10pts)

   a. Create a histogram of the p-values for the new model including interaction terms. Discuss any noticeable differences from the histogram you created for the original model. (3pts)

   b. Apply the Benjamini-Hochberg (BH) procedure to control the False Discovery Rate (FDR) with a q-value of 0.1. How many significant predictors are identified now, including both main effects and interaction effects? (5pts)

   c. Compare these results with those obtained from the original model. Discuss the impact of including interaction terms on the number of discoveries and the control of the FDR. (2pts)