Laboratory Exercise 3

Correlation Analysis

The assigned material to complete this lab includes:

- Required R tutorials and readings:
  - Learn the `data.table` package and its usage. Available at https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html and https://cran.r-project.org/web/packages/data.table/data.table.pdf.
  - Learn the `PerformanceAnalytics` package and its usage. Available at https://cran.r-project.org/web/packages/PerformanceAnalytics/index.html and https://www.rdocumentation.org/packages/PerformanceAnalytics/versions/1.4.3541.
  - Learn the `psych` package and its usage. Available at https://cran.r-project.org/web/packages/psych/index.html and https://cran.r-project.org/web/packages/psych/vignettes/overview.pdf.
- Optional R tutorials and readings:
  - Learn the `apply` family in R as alternatives to loops. Available at https://www.datacamp.com/community/tutorials/r-tutorial-apply-family#gs.=FpFgUc.
  - Learn more about `data.frame`. Avaialble at https://www.datacamp.com/community/tutorials/15-easy-solutions-data-frame-problems-r#gs.PkWlrA8.
- Excel data file: lab3_data.xlsx.
  - Cross-sectional data of a universe comprised of all NYSE stocks available in Morningstar Direct®.
  - The first element in each column is a text string that identifies the data contained in the columns. Columns (A:C) include the stocks name, ticker, and exchange; while columns (D:F) contain the book value of equity (BE), closing price, and market value of equity (or market capitalization) for the stocks at the end of year 2004. Columns (G:K) include the explanatory variables: cash flow-to-price (C/P) ratio, earnings-to-price (E/P) ratio, earnings before interest depreciation and amortization-to-price (EBITDA/P) ratio, free cash flow-to-price (FCF/P) ratio, and sales-to-price (S/P) ratio for the stocks at the end of year 2004. The last column, column L, contains the yearly return at the end of year 2005, which is the observed or explained variable.

Upload the Excel data file 'lab3_data.xlsx' in R and run a correlation analysis for the stock returns during 2005 on the stock characteristic (factor) exposures available at the end of year 2004.

Problem 3.1 Clean the data for any missing or erroneous values in order to include only the stocks that have complete data available on the above variables at the end of 2004. Create the corresponding variables.

Problem 3.2 Since the existence of a single outlying value can markedly influence the results of correlation coefficients, remove any outliers present in the variables, which we have previously defined as data values that are 3 standard deviations away from their means.

Problem 3.3 Regression models intended to explain the cross-section of average stock returns tend to make certain assumptions regarding the collected data in order to avoid distortion in the interpretation of the results. Given the difficulty in interpreting stocks with negative book values and negative earnings, we will exclude them from our data sets. However it is important to note that stocks with negative book values and earnings, behave like stocks with low book value of equity-to-market value of equity (BE/ME) and low earnings-to-price (E/P) ratios, which tend to have lower average returns. Also to avoid that the data sets are dominated by small cap stocks, we will exclude stocks with a closing price lower than $5 and with a market capitalization lower than $100m. Filter the data to avoid stocks with negative book values and earnings, with a closing price lower than $5, and with a market capitalization lower than $100m (any of these three).

Problem 3.4 Compute the correlation coefficients to quantify the direction and strength (if any) of linear association between each pair of the variables. You may use the package `psych` to create a correlation matrix.

(a) Use the function `chart.Correlation` in the package `PerformanceAnalytics` to visualize the relationship between each pair of the variables.

(b) Compute the correlation coefficients to quantify the direction and strength (if any) of linear association between each pair of the variables. You may use the package `psych` to create a correlation matrix.

(c) Create a matrix of p-values in order to test the hypothesis of no correlation between each pair of the variables against the alternative of significant correlation between each pair of the variables. Interpret the results.

Problem 3.5 Run a correlation analysis to evaluate the presence of linear associations between the explained variable and the explanatory variables.

(a) Use the function `chart.Correlation` in the package `PerformanceAnalytics` to visualize the relationship between the explained variables and each of the explanatory variables.

(b) Compute the correlation coefficients. You may use the package `psych` to create a correlation matrix.

(c) Create a matrix of p-values in order to test the hypothesis of no correlation between each pair of the variables against the alternative of significant correlation between each pair of the variables. Interpret the results.

Problem 3.6 Compute the z-scores for each of the explanatory variables and compute the correlation coefficients between each pair of the explanatory variables. Do these results change your answers in problems 3.4? Why? Aggregate the z-scores for the explanatory variables and compute the correlation coefficient between the aggregated explanatory variables and the explained variable. Interpret the results.

Problem 3.7 Based on the correlation coefficients and their significance tests in problem 3.4, is there any variable that may introduce multicollinearity to a regression that combines the 5 fundamental factors so as to explain the average returns for the next year? Identify the variable(s) and explain how we can correct for multicollinearity.

Problem 3.8 Analyzing the correlation coefficients and their significance tests of problem 3.5 and the results in problem 3.7, would we run the risk of overfitting a regression model that incorporates all of the 5 fundamental factors to explain average returns for the next year? Why? Is there any way to prevent overfitting a regression model?