Laboratory Exercise 4

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 `FinTS` package and its usage. Available at https://www.rdocumentation.org/packages/FinTS/versions/0.4-5.
  - Learn the `car` package and its usage. Available at https://cran.r-project.org/web/packages/car/index.html and https://www.rdocumentation.org/packages/car/versions/2.1-5.
- 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: lab4_data.xlsx.
  - Cross-sectional data of a universe comprised of all NASDAQ, NYSE, and NYSE Amex Equities 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:E) contain the earnings-to-price (E/P) ratio and closing price for the stocks at the end of year 2004. Columns (F:H) include the explanatory variables: market value of equity (or market capitalization), book-to-market equity (BE/ME), and 1-year price momentum. The last column, column I, contains the explained variable: yearly return at the end of year 2005.

Upload the Excel data file 'lab4_data.xlsx' in R and run a multivariate, cross-sectional regression of stock returns during 2005 on the stock characteristic (factor) exposures available at the end of year 2004.

Problem 4.1 Clean the data for any missing or erroneous values in order to include only the stocks that have complete data available on the above factor at the end of 2004. Create the corresponding variables.

Problem 4.2 Remove any outliers present in the variables, i.e., those data values that are 3 standard deviations away from their means. Also, filter the data further to avoid stocks with negative book values and earnings, with a closing price lower than \$5, and with a market capitalization lower than \$100m.

Problem 4.3 Create a new variable within the dataset called lnSize, which is the natural logarithm of the size factor exposure. One reason for taking logarithm is to reduce the influence of extreme values.

Problem 4.4 Run a bivariate, cross-sectional test on each of the 3 stock factors to examine how they are related to future returns.

(a) Create a scatter plot to visualize the relationship between the explained variable and each of the explanatory variables.

(b) Compute the correlation coefficient between the variables and calculate the p-values to test the hypothesis of no correlation between the variables against the alternative of significant correlation between the variables. Interpret the results.

Problem 4.5 Express the stocks factor exposures and returns in terms of **within industry z-scores** (calculate z-scores with only stocks in the same industry) in order to build the characteristic matrix. Then run a multivariate, cross-sectional regression of stock returns during 2005 on the normalized value of the factor exposures. To run this regression use the function `lm`. Show the regression results. You may also use the function `summary` to summarize the regression output.

Problem 4.6 Using the analysis of the variance of the explained variable (ANOVA) section in the regression output assess the statistical meaning of the coefficients in terms of:

(a) Whether this regression model fits the data well.

(b) Whether there is a statistically significant (or valid) fit in this multiple regression model.

Problem 4.7 Are **zLnSize** and **zBeMe** jointly significant in our regression model? Conduct a statitical test for the null hypothesis that these two coefficients are jointly zero with the following procedure:

(a) Get the unrestricted sum of the squared residuals or errors (USSR) of the original (unrestricted) multiple regression model calculated in problem 4.5.

(b) Impose the linear restriction on the model, where we omit 2 of the regressors, and run the resulting restricted regression.

(c) Obtain the restricted sum of the squared residuals (RSSR).

(d) Calculate the relevant F-statistic for this test on a subset of linear restrictions with the following formula,

$$F_{0,R} = (\frac{RSSR - USSR}{q}) \div (\frac{USSR}{n-(k+1)})$$

where $q =$ the number of linear restrictions, which it is 2 in this case, $n =$ the number of observations and $k =$ the number of regressors in the unrestricted (original) regression.

(e) Find the critical value of this F-statistic for a significance level of 5% and its related p-value. Note these values must be calculated using q and $n-(k+1)$ degrees of freedom.

(f) Based on the results, state whether you will reject or not the null hypothesis that **zLnSize** and **zBeMe** are not statistically significant in our model.

Problem 4.8 Using the slope coefficients, standard errors, and confidence intervals from the regression output, determine whether there is a valid relationship for the individual coefficients of this model. Use these results along with those of the t-statistics and related p-values to test the null hypothesis that the population slope parameter of the model is different from zero.

Problem 4.9 What is the resulting estimated factor premia? Interpret the economic meaning of this estimate.

Problem 4.10 Test the regression diagnostics for any sign of heteroskedasticity. How would you adjust the regression for heteroskedasticity?

Problem 4.11 Test the regression diagnostics for the presence of serial correlation. How would you adjust the regression for serial correlation?