Data Assignment 3: Constructing Additional Factors and Return Correlations

Financial Data Analytics

It's understood that all assignments are individual work. Failure to observe this may result in immediate failure of the course.

This assignment accounts for 9% of the course mark.

In this assignment we will construct three additional stock return factors. This assignment is built on top of Assignments 1 and 2. The three factors are: idiosyncratic volatility, betting against beta, and momentum. We will calculate these factors for Nasdaq 100 stocks (see Assignment 2 for the list of the stocks and data downloaded from there, as well data downloaded in Assignment 1). Please calculate the monthly factors for these stocks for the period of January, 2000 to 2021, as instructed below. Summary statistics below refer to: *N* (number of observations), mean, standard deviation, median, minimum, 1st percentile, 99th percentile, and maximum.

1. (10 pts) Estimating beta using the CAPM model. Use Fama-French's factors for market return and riskfree rate in CAPM. In month *t*, each stock (permno)'s beta is estimated using the CAPM model from an estimation window of the past 36 months [*t*-35, *t*]. Ensure that there are at least 12 observations for the estimation (i.e., set the estimated beta to missing if there are fewer than 12 observations in the estimation window). Name this variable *beta*. Note that you need data going back before January 2000 in order for you to be able calculate *beta* of your sample period. Report the summary statistics of *beta*.

2. (15 pts) Estimating idiosyncratic volatility, named *ivol*. Following Ang et al. (2006), idiosyncratic volatility is calculated as the standard deviation of regression residuals from estimating an Fama-French three-factor model (market, size, and value) using daily stock returns within the month. To accomplish this,

    a) Download daily returns of the stocks from CRSP. You only need daily stock returns in this step (that is, the following variables from CRSP daily returns data: date, permno, ret);
    b) Merge your downloaded daily returns with Fama-French daily factors (which you also need to download);
    c) Run a regression of the Fama-French three-factor model for **every stock and every month** using all daily returns within the month. Save the regression residuals. Drop the regression if there're fewer than 10 observations in the month; and
    d) Calculate a stock's month-*t* idiosyncratic volatility as the standard deviation of the residuals of daily returns *within* the month.

    Report the summary statistics of *ivol*.

3. (10 pts) Estimating the momentum characteristic of the stock, named *mom*. A stock's month-*t* momentum is defined as its cumulative (compound) returns at months [*t*-11, *t*]. Ensure that

there are at least 10 observations for the estimation (i.e., set the estimated momentum to missing if there are fewer than 10 observations in the estimation window).

Report the summary statistics of *mom*.

4. (5 pts) After Questions 1-3, winsorize the three factors, for every month, at the top and bottom 3 standard deviations. Report again the summary statistics of winsorized *beta*, *ivol*, and *mom*.

5. (10 pts) Combine your three winsorized variables in Assignment 2 (*lnSize*, *bk2mkt*, and *eP*) with your variables after Question 4. Now add month t+1 stock return to your data. Run a correlation analysis between month t+1 stock return (the explained variable) with each of your winsorized factors (6 in total here, the explanatory variables) to evaluate the presence of linear associations.

    a. Use the Python function *pairplot* of the *seanborn* package or R 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 between the explained variables and each of the explanatory variables.
    c. Create a matrix of correlations and p-values in order to test the hypothesis of no correlation between each pair of the explanatory variables against the alternative hypothesis of significant correlation between each pair of the variables. Interpret the results.

Please submit to Dropbox "Data Assignment 3", by 11:59 pm, Thursday Feb. 16:

1. One dataset including all variables up to Question 4. To aid with grading, please output all your datasets in one excel file, with the first two columns being "Permno" and "Date (or yyyymm)".

2. Your codes (can be Python, SAS, Stata, R, SPSS, Matlab, etc.) in Dropbox.

3. A final output report. Please make sure your output report is easy to read. Coefficient estimates do not exceed 4 decimal places, and t-statistics do not exceed 2 decimal places. Any submitted work with output that is *only* embedded into codes will automatically get at least 25% off the entire mark. We grade your work on your final output "report" and only recourse to your codes and data if needed.

4. Any notes if you wish to identify problems and any thoughts in the entire process. As it goes, the key to data analytics rests on good data cleaning work (called "data curation" if you want a fancy big-data word). A good note that has good understanding of data issues may have 5 bonus points.