MIT Sloan School of Management

Analytics of Finance — 15.450
Hui Chen — Spring 2023

Problem Set 6

(Due: 1:00 PM, Thursday, May 11)

1. **Interview questions**:

   (a) True or False (and explain): Since the OLS estimator is BLUE, it is the best way to train a linear prediction model.

   (b) Explain how to apply cross validation to time series data.

2. **Predicting stock returns.** You are looking for signals that can predict returns for individual stocks in order to design a quantitative trading strategy.

   The file Data_HW6.zip contains monthly returns and various characteristics for a selected list of stocks for the period of 1989 to 2016. The characteristics include stock price (Price), market equity value (MV), market-to-book equity ratio (M2B), sales-to-asset ratio (S2A), short-term-debt-to-asset ratio (SD2A), long-term-debt-to-asset ratio (LD2A), price-to-earnings ratio (PE), and quarterly sales (Sales). Notice that some of these variables are only updated once a quarter.

   (a) The first step is to construct a set of features for the prediction model. Recall that we should not directly use non-stationary variables as predictors (why?). Let's use the following features:

      i. M2B, S2A, SD2A, LD2A, PE

      ii. Sales growth (SG): use yearly growth rates (i.e., growth rate of sales in quarter t relative to sales in quarter $t-4$).

      iii. Past 1-month return (R1) and 3-month returns (R3). You can ignore dividends when computing returns.

      Altogether, we will have 8 features.

   (b) Next, we will turn the realized return r_t^i into a binary variable. Define $y_t^i = 1\{r_t^i > 0\}$ (i.e., simply the sign of return). Predict y_{t+1}^i using a logistic regression with the 8 features constructed earlier. Estimate the model using data from 1989 to 2011 (this is your **training sample**).

   (c) Construct the confusion matrix in the **test sample** (post 2011) with the cutoff $\bar{p} = 0.5$. Compute the Type I/II error rates and overall error rates.

   (d) Briefly explain how you would choose the cutoff $\bar{p}$ for the trading strategy.

   (e) **Bonus question:** Use best subset selection to decide which of the 8 features we should include in the model. Remember to do this **using only the training sample**. Redo part 2c. Does the new model outperform the old model in part 2b?