ECON6008 Homework #1

Due: Feb. 11, 2025, 2:00pm

You are allowed to work in a group of no more than 5 (including 5) students and submit one copy of your assignments. You can also work alone. If you work in a group, you must state all the group members' names clearly on the **cover page**. All group members will receive equal marks. You need to submit an **electronic version** of your assignment to `tianxie@smu.edu.sg`. You can use **any software** to complete the assignment. For the coding related questions, you must **present your codes** with necessary comments and put them in the assignment as appendix.

1. (10 marks) Consider the following regression model

$$\boldsymbol{y} = \beta_0 \cdot \boldsymbol{\iota} + \boldsymbol{\epsilon},$$

   where $\boldsymbol{y}$ is a $n \times 1$ response variable, $\boldsymbol{\iota} = [1, 1, ..., 1]^\top$ is the $n \times 1$ constant term with β_0 being the associate coefficient, and $\boldsymbol{\epsilon}$ is a vector of error terms. Prove that the OLS estimate of β_0 is simply the mean of $\boldsymbol{y}$.

2. (30 marks) Describe how we determine the predictor importance in regression tree and bagging tree. Discuss their similarities and differences carefully. Explain why one is more reliable than the other one.

3. (60 marks) This question is about the larger VIX data set `vixlarge.csv` that contains the VIX data and the associated dates.

   (a) (10 marks) Plot the VIX data against date in `line`. Clearly label the horizontal and vertical axises.

   (b) (10 marks) Pick 5 nodes and fit the data using one of the regression splines. State the method you choose clearly and show the plot.

   (c) (20 marks) Let the dependent variable $\boldsymbol{y}$ be the VIX and the first and the second columns of the independent variable $\boldsymbol{X}$ be the intercept term and the lag of VIX (set $x_0 = 0$). Conduct a one-step-ahead **rolling window** exercise.

      i. Set the window length at 3000 and make forecast on the next period y_{t+1}.

      ii. Start from the beginning and roll until the end.

      iii. For each roll, we make forecast using ridge and lasso methods with tuning parameter $\lambda = 1, 10$ for each method. In total, we compare 4 methods.

      iv. Comparing the forecasts with the actual true values of y_{t+1}. Compute the mean squared forecast errors and the mean absolute forecast errors for the four methods and report them in a table.

v. Which method has the best performance and which one has the worst? Provide your understanding and explanation of the results.

vi. Come up with an algorithm that can beat the best performing method stated in question v. **Clearly describe** your motivation, the details of the algorithm, and the results.

(d) (20 marks) We now consider a more general forecasting exercise with model

$$y_{t+h} = f(x_t) + u_{t+h}, \qquad \text{for } t = 1, ..., n-h$$

where h is the forecasting horizon. Note that Q3(c) is the special case with $h = 1$ and $f(\cdot)$ being the ridge or LASSO estimator. We now replicate Q3(c) with $h = [1, 5, 10, 22]$ using LASSO and the regression tree. Choose your own tuning parameters this time, state them clearly, and report your forecasting results in a table. What do you observe?