Homework 1

Notation:

- N: number of assets or stocks
- P: number of predictors
- $\mathcal{S}_{\text{train}}$: Sample used for estimating the model. For example, Y_{train} is the vector of N monthly stock returns at the end of December 2024, and X_{train} is the $N \times P$ matrix of predictors observed at the end of November 2024.
- $\mathcal{S}_{\text{test}}$: Sample used for conducting forecasts. For example, Y_{test} is the vector of N monthly stock returns at the end of January 2025, and X_{test} is the $N \times P$ matrix of predictors observed at the end of December 2024.

Exercise:

1. Use $\mathcal{S}_{\text{train}}$ to regress Y on X, and then compute the sample square root of MSE $RMSE_{\text{train}}$, leverage scores h_{ii} and $|\hat{y}_i - \hat{y}_{(-i)}|$, where $\hat{y}_{(-i)}$ is the leave-one-out prediction.
$$RMSE_{\text{train}} = \left(\sum_{i \in \mathcal{S}_{\text{train}}} (y_i - x_i' \hat{\beta})^2 \right)^{1/2}$$
2. Report the minimum, 1%-percentile, 25%-percentile, median, 75%-percentile, 99%-percentile, and maximum of h_{ii} and $|\hat{y}_i - \hat{y}_{(-i)}|$. Do you observe any influential observations?
3. Use $\mathcal{S}_{\text{test}}$ and the model estimated in Step (1) to compute the out-of-sample RMSE. That is,
$$RMSE_{\text{test}} = \left(\sum_{i \in \mathcal{S}_{\text{test}}} (y_i - x_i' \hat{\beta})^2 \right)^{1/2}$$
4. Remove the observations with $h_{ii} > 2\frac{P}{N}$ and repeat Step (1) and (3). Do you get a lower or greater RMSE?
5. Remove three observations with the largest $|\hat{y}_i - \hat{y}_{(-i)}|$ and repeat Step (1) and (3). Do you get a lower or greater RMSE?