Data Science / Data Analysis - January 30, 2025

Master's Degree in Computer Engineering - University of Salerno

Analyze the dataset `RegressionDSDA250130.csv`, which contains $n = 100$ observations of a dependent variable Y and $p = 25$ predictors X_j $(j = 1, 2, \ldots, p)$, all potentially useful for predicting Y. To this end, it is required to use the R environment.

1. Determine the linear models that minimize the Mean Squared Error (MSE), identifying the significant predictors for the prediction of Y and estimating their coefficients β_j, using the following strategies:

   i) **Best subset selection (BSS)** based on the **Bayesian Information Criterion (BIC)**,

   ii) **Backward stepwise** using **5-fold cross-validation**,

   iii) **Ridge regression**,

   iv) **LASSO regression**.

2. Evaluate the test MSE of the linear models identified in the previous point and select the regression strategy that allows building the empirical linear model that minimizes, among the ones examined, the test MSE.

70% of the observations in the dataset must be used for model training and parameter selection, while the test set must consist of the remaining 30% of the given data.