Data Science & Data Analysis Homework

Master's Degree in Computer Engineering – University of Salerno

Question on Linear Regression

Analyze in R the dataset *Regression2024.csv*, which consists of $n = 60$ observations of a dependent variable Y and $p = 30$ predictors X_j $(j = 1, 2, \ldots, p)$, potentially useful for predicting Y. The objective of the analysis is, after comparing different regression techniques presented during the course, to determine the empirical linear model that minimizes the prediction error on a test set. You are asked to:

a) evaluate the correlation and *multicollinearity* among the predictors;

b) estimate the parameters β_j $(j = 0, 1, \ldots, p)$ by implementing the least squares estimator for multiple regression **without using the** `lm()` **function**, and calculate the *p-values* for the parameter tests **without using the** `summary()` **function**;

c) estimate the parameters of the multiple regression model using the `lm()` function, calculate the *p-values* using the `summary()` function, and compare the results with those obtained in point b);

d) select the strategy that allows you to build the regression model that minimizes the test error on the variable Y, choosing among:

i) stepwise, using all approaches (*forward*, *backward*, and *hybrid*), testing metrics such as *AIC* and *BIC*, and approaches based on *cross-validation*;

ii) ridge regression and

iii) LASSO;

e) identify the significant predictors for predicting Y and provide the estimated values of their coefficients β_j using the strategy determined in point d).

It is required that 80% of the dataset observations be used for training the models and selecting the best strategy, while the remaining 20% should be used for final testing.