Assignment #2: Linear Regression

Problem 1

(1) Fit a linear model for the Boston dataset in MASS library using median value of owner-occupied homes (*medv*) as response and average number of rooms per dwelling (*rm*) as the predictor (use the basic syntax *lm(y~x,data=dataname)*. What are the coefficients? What does it suggest about the fitness? Show the scatter plot as well as the linear model fit in one figure.

(2) Fit a linear model using the same input and output in Question (1), but replace the predictor with log(rm) (i.e., use the basic syntax *lm(y~log(x),data=dataname)*). What are the coefficients? Is this model a better fit compared to the one in Question (1)? Justify your answer. Show the data and the linear model fit in one figure.

(3) Fit a linear model using the same output (medv) in Question (1), but regress it against the lstat variable. What are the coefficients? How does this model fit compared to the one in Question (1)?

Problem 2

Fit a regression model of medv on lstat and lstat^2 (syntax *lm(y~x+I(x^2), data=dataname)*). Provide a summary of the model. Suppose that we have another linear model which simply fits medv with predictor lstat (used in Question 1(3)), which model has better fitness? Justify your answer.

Problem 3

(1) Except lstat and rm, there are other predictors in the Boston dataset. You can check the whole dataset using syntax *?Boston* and *summary(Boston)*. Fit a multiple linear regression model of medv on all the predictors (syntax: *lm(y~., data=dataname)*). What are the coefficients? What does it suggest about fitness?

(2) Do you think this model is some sort of cumbersome? Improve this model by reducing the inputs based on the summary of the model in Question 3(1)). Explain the methodology used for variable selection and provide a summary of the final model.
(syntax: *lm(y~predictor1+predictor2+…+predictorN, data=dataname)*)

Submit through link: eCampus -> Assignments:

Deadline: Feb 22 (Mon) @11:59pm