Home Work 5-1

1. McDonald and Ayers [1978] present data from an early study that examined the possible link between air pollution and mortality. (data_table_B15.xlsx file summarizes the data). The data description is as follows:

   MORT = total age-adjusted mortality from all causes, in deaths per 100,000 population.

   PRECIP = mean annual precipitation (in inches),

   EDUC = median number of school years completed for persons of age 25 years or older,

   NONWHITE = percentage of the 1960 population that is nonwhite

   NOX = relative pollution potential of oxides of nitrogen

   SO2 = relative pollution potential of sulfur dioxide. Relative pollution potential is the product of the tons emitted per day per square kilometer and a factor correcting the SMSA dimensions and exposure.

   a. Fit a multiple linear regression model relating the mortality rate to these regressors.

   b. Test for significance of regression. What conclusions can you draw?

   c. Use t tests to assess the contribution of each regressor to the model. Discuss your findings.

   d. Find a 95% CI for the regression coefficient for SO2.

   e. Run all possible models and choose the best one with justifications. (You may not consider PRESS statistic)

   f. Run forward, backward and stepwise regression on the data.

   g. Do all 3 procedures picked the same model? If yes: Should it happen all the time, If NO: Why don't they pick the same?

   h. Perform the residual analysis of your final model and provide the final estimated model (you must describe the rule you applied).

2. (a) To judge all possible subset models, are the two criterion "Adj-R^2" and "MSE" same? Prove what ever your response is.

   (b) One person decided to adopt forward selection procedure but at every stage he wanted to choose that variable which provides the highest increase in adj-R^2 and also significant at level α. Will he get the same result as our usual current procedure? Justify.

   (c) In backward selection procedure, once you drop an insignificant variable do you think standard error will of the model will increase or decrease? Justify.