CSCI E-106:Assignment 8

Due Date: November 16, 2020 at 7:20 pm EST

Instructions

Students should submit their reports on Canvas. The report needs to clearly state what question is being solved, step-by-step walk-through solutions, and final answers clearly indicated. Please solve by hand where appropriate.

Please submit two files: (1) a R Markdown file (.Rmd extension) and (2) a PDF document, word, or html generated using knitr for the .Rmd file submitted in (1) where appropriate. Please, use RStudio Cloud for your solutions.

Problem 1

Refer to the the Efficacy of Nosocomial Infection Control (SENIC) data set. The primary objective of the Study on was to determine whether infection surveillance and control programs have reduced the rates of nosocomial (hospitalacquired) infection in United States hospitals. This data set consists of a random sample of 113 hospitals selected from the original 338 hospitals surveyed. Each line of the dataset has an identification number and provides information on 11 variables for a single hospital. The data presented here are for the 1975-76 study period. (15 points, 5 points each)

a-) Second-order regression model is to be fitted for relating number of nurses (Y) to available facilities and services (X).

b-) Fit the second-order regression model. Plot the residuals against the fitted values. How well does the second-order model appear to fit the data?

c-) Test whether the quadratic term can be dropped from the regression model; use α=0.1, State the alternatives, decision rule, and conclusion.

Problem 2

Use the fortune data under the faraway r library, data(prostate,package="faraway"). Use the prostate data with lpsa as the response and the other variables as predictors.

Implement the following variable selection methods to determine the "best" model: (40 points, 10 points each)

a-) Backward elimination

b-) AIC

c-) Adjusted R_a^2

d-) Mallows C_p

Problem 3

Refer to the SENIC data set in problem 1. Length of stay (Y) is to be predicted, and the pool of potential predictor variables includes all other variables in the data set except medical school affiliation and region. It is believed that a model with $log(Y)$ as the response variable and the predictor variables in first-order terms with no interaction terms will be appropriate. Consider cases 57-113 to constitute the model-building data set to be used for the following analyses.(45 points, 9 points each)

a-) Prepare separate dot plots for each of the predictor variables. Are there any noteworthy features in these plots? Comment.

b-) Obtain the scatter plot matrix. Also obtain the correlation matrix of the X variables. Is thereevidence of strong linear pairwise associations among the predictor variables here?

c-) Obtain the three best subsets according to the C_p criterion, Which of these subset models appears to have the smallest bias?

d-) The regression model identified as best in part c is to be validated by means of the validation data set consisting of cases 1-56. Fit the regression model identified in part c as best to the validation data set. Compare the estimated regression coefficients and their estimated standard deviations with those obtained in Part C.

e-) Also compare the error mean squares and coefficients of multiple determination. Does the model fitted to the validation data set yield similar estimates as the model fitted to the model-building data set?