CSCI E-106:Assignment 9

Due Date: November 23, 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 Brand preference data, build a model with all independent variables (45 pts, 5 points each)

a-) Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with α= 0.10. State the decision rule and conclusion.

b-) Obtain the diagonal elements of the hat matrix, and provide an explanation for the pattern in these elements.

c-) Are any of the observations outlying with regard to their X values according?

d-) Management wishes to estimate the mean degree of brand liking for moisture content $X_1 = 10$ and sweetness $X_2 = 3$. Construct a scatter plot of X_2 against X_1 and determine visually whether this prediction involves an extrapolation beyond the range of the data. Also, determine whether an extrapolation is involved. Do your conclusions from the two methods agree?

e-) The largest absolute studentized deleted residual is for case 14. Obtain the DFFITS, DFBETAS, and Cook's distance values for this case to assess the influence of this case. What do you conclude?

f-) Calculate the average absolute percent difference in the fitted values with and without case 14. What does this measure indicate about the influence of case 14?

g-) Calculate Cook's distance D; for each case and prepare an index plot. Are any cases influential according to this measure?

h-) Find the two variance inflation factors. Why are they both equal to 1?

Problem 2

Refer to the Lung pressure Data. Increased arterial blood pressure in the lungs frequently leads to the development of heart failure in patients with chronic obstructive pulmonary disease (COPD). The standard method for determining arterial lung pressure is invasive, technically difficult, and involves some risk to the patient. Radionuclide imaging is a noninvasive, less risky method for estimating arterial pressure in the lungs. To investigate the predictive ability of this method, a cardiologist collected data on 19 mild-to-moderate COPD patients. The data includes the invasive measure of systolic pulmonary arterial pressure (Y) and three potential noninvasive predictor variables. Two were obtained by using radionuclide imaging emptying rate of blood into the pumping chamber or the heart (X_1) and ejection rate of blood pumped out of the heart into the lungs (X_2) and the third predictor variable measures blood gas (X_3). (35 points, 5 points each)

a-) Find the best regression model by using first-order terms and the cross-product term. Ensure that all variables in the model are significant at 5%.

b-) Obtain the residuals and plot them separately against Y and each of the three predictor variables. On the basis of these plots. should any further modification of the regression model be attempted?

c-) Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Does the normality assumption appear to be reasonable here?

d-) Obtain the variance inflation factors. Are there any indications that serious multicollinearity problems are present? Explain.

e-) Obtain the studentized deleted residuals and identify outlying Y observations. Use the Bonferroni outlier test procedure with α= .05. State the decision rule and conclusion.

f-) Obtain the diagonal elements of the hat matrix. Are there any outlying X observations? Discuss.

g-) Cases 3, 8, and 15 are moderately far outlying with respect to their X values, and case 7 is relatively far outlying with respect to its Y value. Obtain DFFITS, DFBETAS, and Cook's distance values for these cases to assess their influence. What do you conclude?

Problem 3

Refer to the Prostate cancer data set. Serum prostate-specific antigen (PSA) was determined in 97 men with advanced prostate cancer. PSA is a well-established screening test for prostate cancer and the oncologists wanted to examine the correlation between level of PSA and a number of clinical measures for men who were about to undergo radical prostatectomy. The measures are cancer volume, prostate weight, patient age, the amount of benign prostatic hyperplasia, seminal vesicle invasion, capsular penetration, and Gleason score. (20 points, 5 points each)

a-) Select a random sample of 65 observations to use as the model-building data set. Develop a best subset model for predicting PSA. Justify your choice of model. Assess your model's ability to predict and discuss its usefulness to the oncologists.

b-) Perform appropriate diagnostic checks to evaluate outliers and assess their influence.

c-) Fit the regression model identified in part a to the validation data set. Compare the estimated regression coefficients and their estimated standard errors with those obtained in part a. Also compare the error mean square 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?

d-) Calculate the mean squared prediction error and compare it to MSE obtained from the model-building data set. Is there evidence of a substantial bias problem in MSE here?