CSCI E-106:Assignment 6

Due Date: November 3, 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 Commercial properties data set. A commercial real estate company evaluates vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions.The variables in the data set are the age (X_1), operating expenses and taxes (X_2), vacancy rates (X_3), total square footage (X_4),and rental rates (Y). (35 points, 5 points each)

a-) Obtain the scatter plot matrix and the correlation matrix. Interpret these and state your principal findings.

b-) Fit regression model for four predictor variables to the data. State the estimated regression function.

c-) Obtain the residuals and prepare a box plot of the residuals. Does the distribution appear to be fairly symmetrical?

d-)Conduct the Breusch-Pagan test for constancy of the elTor varhmce, assuming log σ_i^2=γ_0+$\gamma_1 X_1$+$\gamma_2 X_2$+$\gamma_3 X_3$+$\gamma_4 X_4$ use $\alpha = .01$. State the alternatives, decision rule, and conclusion

e-) Obtain QQ plot and error vs. fitted values, and comment on the graphs.

f-) Estimate β_0, β_1, β_2, β_3 and β_4 jointly by the Bonferroni procedure, using a 95 percent family confidence coefficient. Interpret your results.

g-) X_1=5, X_2=8.25, X_3=0 and X_4=250000, calculate the predicted rental rate and 95% confidence interval

Problem 2

Refer to the CDI data set. You have been asked to evaluate two alternative models for predicting the number of active ve physicians (Y) in a CDI. Proposed model I includes as predictor variables total population (X_1), land area (X_2), and total personal income (X_3). Proposed model II includes as predictor variables population density (X_1, total population divided by land area), percent of population greater than 64 years old (X_2), and total personal income (X_3).(40 points, 10 points each) a-) Obtain the scatter plot matrix and the correlation matrix for each proposed model. Summarize the information provided.

b-) For each proposed model, fit the first-order regression model with three predictor variables.

c-) Calculate R^2 for each model. Is one model clearly preferable in terms of this measure?

d-) For each model, obtain the residuals and plot them against $\hat{Y}$, each of the three predictor variables. Also prepare a normal probability plot for each of the two fitted models. Interpret your plots and state your findings. Is one model clearly preferable in terms of appropriateness?

Problem 3

Refer to Grocery retailer data set.A large, national grocery retailer tracks productivity and costs of its facilities closely. Each data point for each variable represents one week of activity. The variables included are the number of cases shipped (X_1),the indirect costs of the total labor hours as a percentage (X_2), a qualitative predictor called holiday that is coded 1 if the week has a holiday and 0 otherwise (X_3), and the total labor hours (Y). (25 points, 5 points each)

a-) Fit regression model to the data for three predictor variables. State the estimated regression function. How are b_1, b_2 , and b_3 interpreted here? (5 points)

b-)Prepare a time plot of the residuals. Is there any indication of that the error terms are correlated?

c-) Obtain the analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X_1; with X_3 , given X_1; and with X_2 , given X_1, and X_3.

d-) Test whether X_2 can be dropped from the regression model given that X_1, and X_3 are retained. Use the F* test statistic and $\alpha = .05$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

e-) Does SSR(X_1)+SSR(X_2/X_1) equal SSR(X_2)+SSR(X_1/X_2) here? Must this always be the case? (5 points)