CSCI E-106:Assignment 7

Due Date: November 9, 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

Use the fortune data under the faraway r library, data(fortune,package="faraway"). The wealth in billions of dollars for 232 billionaires is given in fortune. (50 points, 10 points each) (Hint: refer to the interaction.pdf and rmd files for details)

a-)Plot the wealth as a function of age using a different plotting symbol for the different regions of the world. b-)Plot the wealth as a function of age with a separate panel for each region. c-)Determine a transformation on the response to facilitate linear modeling. d-)What is the relationship of age and region to wealth? e-)Check the assumptions of your model using appropriate diagnostics.

Problem 2

Refer to the CDI data set. A regression model relating serious crime rate (Y, total serious crimes divided by total population) to population density (X_1, total population divided by land area) and unemployment rate (X_3) is to be constructed. (30 points, 10 points each)

a-) Fit second-order regression model. Plot the residuals against the fitted values. How well does the second-order model appear to fit the data? What is R^2?

b-) Test whether or not all quadratic and interaction terms can be dropped from the regression model; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

c-) Instead of the predictor variable population density, total population (X_1 and land area (X_2) are to be employed as separate predictor variables, in addition to unemployment rate (X_3). The regression model should contain linear and quadratic terms for total population, and linear terms only for land area and unemployment rate. (No interaction terms are to be included in this model.) Fit this regression model and obtain R^2. Is this coefficient of multiple determination substantially different from the one for the regression model in part (a)?

Problem 3

Refer to the CDI data set. The number of active physicians (Y) is to be regressed against total population (X_1), total personal income (X_2), and geographic region (X_3, X_4 , X_5). (20 points, 10 points each)

(“(”Geographic region classification is that used by the U.S. Bureau of the Census, where: 1 = NE, 2 = NC, S = 3, W=4")

a-) Fit a first-order regression model. Let $X_3 = 1$ if NE and 0 otherwise, $X_4 = 1$ if NC and 0 otherwise, and $X_5 = 1$ if S and 0 otherwise. Examine whether the effect for the northeastern region on number of active physicians differs from the effect for the north central region by constructing an appropriate 90 percent confidence interval. Interpret your interval estimate.

b-) Test whether any geographic effects are present; use $\alpha = .10$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?