HOMEWORK #7

Read:

- Chapter 4 in ‘Regression Analysis by Example’.

R Assignment:

Solve the following question using R. Hand in your R code, figures and answers to all questions.

1. Census data was collected on the 50 states and Washington, D.C. We are interested in determining whether average lifespan (LIFE) is related to the ratio of males to females in percent (MALE), birth rate per 1,000 people (BIRTH), divorce rate per 1,000 people (DIVO), number of hospital beds per 100,000 people (BEDS), percentage of population 25 years or older having completed 16 years of school (EDUC) and per capita income (INCO).

   The data can be found at

   www.stat.columbia.edu/~martin/W2024/Data/Census.txt

   (a) Fit a multiple regression using LIFE as the response variable, and the other six variables as the explanatory variables. Write down the regression equation.

   (b) Make boxplots of each of the 6 explanatory variables. Are there any notable features?

   (c) Plot the residuals against the fitted values. Are there any notable points. In particular look for points with large residuals or that may be influential.

   (d) Compute and plot the leverage of each point. Identify any points that have a leverage larger than 0.5.

   (e) Study the observations identified in part (d). What notable features do these points have?

   (f) Compute the Cook’s distance for each point. Identify any points that have a leverage larger than 1. Are these the same observations as those seen in part (d)?

   (g) Plot the residuals against the variable BEDS. Specifically mark the point corresponding to Washington, D.C. What can you say about this observation?

   (h) Remove the observation corresponding to Washington, D.C. and refit the model. Are there any notable differences with the model fit in part (a)?

   (i) Plot the studentized residuals against each of the 6 explanatory variables. Specifically mark the observation corresponding to Utah. What is notable about this state?

   (j) Remove the observation corresponding to Utah and refit the model. Are there any notable differences with the model fit in part (a)? In particular, how does Utah’s exclusion impact the R^2 value?