Homework 3

PSTAT Summer 2023

Due date: September 4th, 2023 at 23:59 PT

1. This question uses the *cereal* data set.
   The data set *cereal* contains measurements for a set of 77 cereal brands. For this assignment only consider the following variables:

- Rating: Quality rating
- Protein: Amount of protein.
- Fat: Amount of fat.
- Fiber: Amount of fiber.
- Carbo: Amount of carbohydrates.
- Sugars: Amount of sugar.
- Potass: Amount of potassium.
- Vitamins: Amount of vitamins.
- Cups: Portion size in cups.

Our goal is to study how *rating* is related to all other 8 variables.

(a) **(2 pts)** Run a multiple linear regression model after removing observations 5,21 and 58. Calculate the fitted response values and the residuals from the linear model mentioned above. Use *head* function to show the first 5 entries of the fitted response values and the first 5 entries of the residuals.

(b) **(2 pts)** Use a graphical diagnostic approach to check if the random errors have constant variance. Briefly explain what diagnostics method you used and what is your conclusion.

(c) **(2 pts)** Use a graphical method to check if the random errors follow a normal distribution. What do you conclude?

(d) **(3 pts)** Run a *Shapiro-Wilk* test to check if the random errors follow a normal distribution. What is the null hypothesis in this test? What is the p-value associated with the test? What is your conclusion?

(e) **(3 pts)** Plot successive pairs of residuals. Do you find serial correlation among observations?

(f) **(3 pts)** Run a *Durvin-Watson* test to check if the random errors are uncorrelated. What is the null hypothesis in this test? What is the p-value associated with the test? What is your conclusion?

(g) **(2 pts)** Compute the hat matrix $\boldsymbol{H}$ in this data set (you don't need to show the entire matrix). Verify numerically that $\sum_{i=1}^{n} H_{ii} = p^* = p + 1$.

(h) **(2 pts)** Check graphically if there is any high-leverage point. What is the criterion you used?

(i) **(2 pts)** Compute the standardized residuals. Without drawing a plot, is there any outlier? What is the criterion you used?

(j) **(2 pts)** Calculate the Cook's distance. How many observations in this data set have a Cook's distance that is greater than $4/n$?

(k) **(2 pts)** Check whether the response needs a Box-Cox transformation. If a Box-Cox transformation is necessary, what would be the form of the transformation?