

# Homework 2

PSTAT 126 Winter 2023

Due date: February 17th, at 23:59 PT

1. This question uses the *cereal* data set available in the Homework Assignment 2 on Canvas. The following command can be used to read the data into R. Make sure the “cereal.txt” file is in the same folder as your R/Rmd file.

```
Cereal <- read.table("cereal.txt",header=T)
str(Cereal)
```

```
## 'data.frame':    77 obs. of  16 variables:
## $ name      : chr  "100%_Bran" "100%_Natural_Bran" "All-Bran" "All-Bran_with_Extra_Fiber" ...
## $ mfr       : chr  "N" "Q" "K" "K" ...
## $ type      : chr  "C" "C" "C" "C" ...
## $ calories: int   70 120 70 50 110 110 130 90 90 ...
## $ protein  : int   4 3 4 4 2 2 2 3 2 3 ...
## $ fat      : int   1 5 1 0 2 2 0 2 1 0 ...
## $ sodium   : int  130 15 260 140 200 180 125 210 200 210 ...
## $ fiber    : num   10 2 9 14 1 1.5 1 2 4 5 ...
## $ carbo    : num   5 8 7 8 14 10.5 11 18 15 13 ...
## $ sugars   : int   6 8 5 0 8 10 14 8 6 5 ...
## $ potass   : int  280 135 320 330 -1 70 30 100 125 190 ...
## $ vitamins: int   25 0 25 25 25 25 25 25 25 ...
## $ shelf    : int   3 3 3 3 3 1 2 3 1 3 ...
## $ weight   : num   1 1 1 1 1 1 1 1.33 1 1 ...
## $ cups     : num   0.33 1 0.33 0.5 0.75 0.75 1 0.75 0.67 0.67 ...
## $ rating   : num  68.4 34 59.4 93.7 34.4 ...
```

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) (4pts) Explore the data and perform a descriptive analysis of each variable, include any plot/statistics that you find relevant (histograms, scatter diagrams, correlation coefficients). Did you find any outlier? If yes, is it reasonable to remove this observation? why?

(b) (3pts) Use the `lm` function in R to fit the MLR model with *rating* as the response and the other 8 variables as predictors. Display the summary output.

- (c)(3pts) Which predictor variables are statistically significant under the significance threshold value of 0.01?
- (d)(2pts) What proportion of the total variation in the response is explained by the predictors?
- (e)(3pts) What is the null hypothesis of the global F-test? What is the p-value for the global F-test? Do the 7 predictor variables explain a significant proportion of the variation in the response?
- (f)(2pts) Consider testing the null hypothesis  $H_0 : \beta_{carbo} = 0$ , where  $\beta_{carbo}$  is the coefficient corresponding to *carbohydrates* in the MLR model. Use the t value available in the summary output to compute the p-value associated with this test, and verify that the p-value you get is identical to the p-value provided in the summary output.
- (g)(4pts) Suppose we are interested in knowing if either *vitamins* or *potass* had any relation to the response *rating*. What would be the corresponding null hypothesis of this statistical test? Construct a F-test, report the corresponding p-value, and your conclusion.
- (h)(3pts) Use the summary output to construct a 99% confidence interval for  $\beta_{protein}$ . What is the interpretation of this confidence interval?
- (i)(3pts) What is the predicted *rating* for a cereal brand with the following information:
- Protein=3
  - Fat=5
  - Fiber=2
  - Carbo=13
  - Sugars=6
  - Potass=60
  - Vitamins=25
  - Cups=0.8
- (j). (3pts) What is the 95% prediction interval for the observation in part (i)? What is the interpretation of this prediction interval?

Q2.(20pts) Consider the MLR model with  $p$  predictors:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

If we define  $\hat{\sigma}^2 = \frac{SSR}{n-p^*}$ , with  $p^* = p + 1$ . Use theoretical results from the lectures to show that  $\hat{\sigma}^2$  is an unbiased estimator of  $\sigma^2$ . Find  $V(\hat{\sigma}^2)$ .