Homework 2

PSTAT Winter 2023

Due date: February 17th, 2023 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.csv",header=T, sep = ",")
str(Cereal)
```

```
## 'data.frame':    77 obs. of  17 variables:
##  $ X       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ name    : chr  "100% Bran" "100% Natural Bran" "All-Bran" "All-Bran with Extra Fiber" ...
##  $ manuf   : chr  "N" "Q" "K" "K" ...
##  $ type    : chr  "cold" "cold" "cold" "cold" ...
##  $ calories: int  70 120 70 50 110 110 110 130 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 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?

I notice that the values carbo and cups are outliers in our data since their p-values are significantly higher than any of the other values.

(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.

```
model <- lm(rating~protein+fat+fiber+carbo+sugars+potass+vitamins+cups, data= Cereal)
summary(model)
```

```
##
## Call:
## lm(formula = rating ~ protein + fat + fiber + carbo + sugars +
##     potass + vitamins + cups, data = Cereal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5603  -3.2485  -0.4155   2.3679   9.2403
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51.57435    4.21658  12.231  < 2e-16 ***
## protein      1.96222    0.66433   2.954 0.004309 **
## fat         -4.00155    0.63099  -6.342 2.13e-08 ***
## fiber        3.24519    0.63885   5.080 3.16e-06 ***
## carbo       -0.01803    0.16384  -0.110 0.912708
## sugars      -1.68219    0.16337 -10.297 1.63e-15 ***
## potass      -0.02537    0.02140  -1.185 0.239948
## vitamins    -0.10262    0.02568  -3.997 0.000161 ***
## cups         0.49932    2.75464   0.181 0.856698
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.609 on 68 degrees of freedom
## Multiple R-squared:  0.9037, Adjusted R-squared:  0.8923
## F-statistic: 79.74 on 8 and 68 DF,  p-value: < 2.2e-16
```

(c)(3pts) Which predictor variables are statistically significant under the significance threshold value of 0.01?

The variables which are statistically significant under the threshold value of 0.01 are protein, fat, fiber, sugars, and vitamins. I know that because each variable has a p-value less than 0.01.

(d)(2pts) What proportion of the total variation in the response is explained by the predictors?

We can measure the proportion of the total variation in the response explained by the predictors by looking at our R^2 value. In this case, our R^2 value is equal to .904. So the proportion of the total variation in the

response explained by the predictors is 90.4%. We know that values closer to 1 indicate a good fit for the data, we can say that this value means that our model is a good fit for our data.

(e)(3pts) What is the null hypothesis of the global F-test? What is the p-value for the global F-test? Do the 8 predictor variables explain a significant proportion of the variation in the response?

The null hypothesis of the global F-test looks like $H_0 : \beta_{protein} = \beta_{fat} = \beta_{fiber} = \beta_{carbo} = \beta_{sugars} = \beta_{potass} = \beta_{vitamins} = \beta_{cups} = 0$
The interpretation here is that the null model and the full model are equivalent. The p-value here for the global F-test is equal to $< 2e^{-16}$ which is far less than our $\alpha = .01$ and $\alpha = .05$. This means that we reject our H_0 and we can conclude that our 8 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 β_{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.

```
n = 77
p = 8
T1 = -.11
p_value = pt(q = T1, df = n - p - 1) * 2
p_value
```

```
## [1] 0.9127334
```

This p-value corresponds to the p-value found in the summary.

(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.

The null hypothesis of this test would look like $H_0 : \beta_{vitamins} = \beta_{potass} = 0$.

```
mod_Cereal <- lm(rating~protein+fat+fiber+carbo+sugars+potass+vitamins+cups, data= Cereal)
mod_cereal <- lm(rating ~ protein+fat+fiber+carbo+sugars+cups, data = Cereal)
anova(mod_cereal,mod_Cereal)
```

```
## Analysis of Variance Table
##
## Model 1: rating ~ protein + fat + fiber + carbo + sugars + cups
## Model 2: rating ~ protein + fat + fiber + carbo + sugars + potass + vitamins +
##     cups
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     70 1817.4
## 2     68 1444.7  2    372.79 8.7737 0.0004076 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of our F-test here is equal to 0.0041. We will reject our H_0 and conclude that the full model with the predictors vitamins and potassium is better than the model without it.

(h)(3pts) Use the summary output to construct a 99% confidence interval for $\beta_{protein}$. What is the interpretation of this confidence interval?

```
CI.beta_protein <- c(model$coefficients[2] - qt(.995, df = model$df.residual)*.6643,
                     model$coefficients[2] + qt(.995, df = model$df.residual)*.6643)
CI.beta_protein
```

```
##   protein   protein
## 0.2017725 3.7226705
```

```
confint(model, level = .99)[2,]
```

```
##     0.5 %    99.5 %
## 0.2017021 3.7227409
```

We can see from the interval I created and the interval created from the confint function that a 99% confidence interval given the t-value from the summary output is equal to an interval from .2017 to 3.72274. The interpretation for this is that if we were to take many samples of protein for all 77 observations, we would find that the value would fall in this interval 99% of the time.

(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

```
y <- (model$coefficients[1]) + (model$coefficients[2]*3) + (model$coefficients[3]*5) + (model$coefficie
y
```

```
## (Intercept)
##    29.92808
```

Our predicted rating according to our given values is equal to 29.93

(j). (3pts) What is the 95% prediction interval for the observation in part (i)? What is the interpretation of this prediction interval?

```r
x <- model.matrix(model)
x0 <- c(1, 3,5,2,13,6,60,25,.8)
xtxi <- solve (t (x) %*% x)
RSE <- 4.61
SE.y <- RSE*sqrt(1 + (t(x0) %*% xtxi %*% x0))
CI.y <- c(y - qt(.975, df = 68)*SE.y, y + qt(.975, df = 68)*SE.y)
CI.y
```

```
## [1] 19.21355 40.64260
```

The interpretation here is that if given multiple samples of $\hat{y}$, the value would lie in this interval 95% of the time.

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

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

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 σ^2. Find $V(\hat{\sigma}^2)$.

Let us begin with calculating $\mathbb{E}[\hat{\epsilon}^T\hat{\epsilon}]$. Let us remember that $H = X(X^TX)^{-1}X^T$, $M = (I - H)$ and $\hat{\epsilon} = My$

$\mathbb{E}[\hat{\epsilon}^T\hat{\epsilon}] = \mathbb{E}[y^TM^TMy]$. Since M is idempotent and symmetric, we can say $M^2 = M$ and $M^T = M$. Thus, we have $\mathbb{E}[\hat{\epsilon}^T\hat{\epsilon}] = \mathbb{E}[y^TMy]$. Now we must make assumptions on X that it is fixed. Thus, we get $\mathbb{E}[\hat{\epsilon}^T\hat{\epsilon}] = M\mathbb{E}[y^Ty] + \sigma^2 tr(M)$. We know that $M\mathbb{E}[y^Ty] = 0$ so we get $\mathbb{E}[\hat{\epsilon}^T\hat{\epsilon}] = \sigma^2 tr(M)$ where $tr(M) = n - p^*$. Thus, we can finally say that $\mathbb{E}[\hat{\epsilon}^T\hat{\epsilon}] = \sigma^2(n - p^*)$. Since $\mathbb{E}[n - p^*] = n - p^*$, we find that $\mathbb{E}[\frac{\hat{\epsilon}^T\hat{\epsilon}}{n-p^*}] = \sigma^2$. Thus, proving that $\hat{\sigma}^2$ is an unbiased estimator of σ^2.

Now we will find $V(\hat{\sigma}^2)$.

$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-p^*}$

$V(\hat{\sigma}^2) = V(\frac{\sum_{i=1}^n e_i^2}{n-p^*}) = \frac{1}{(n-p^*)^2} V(\sum_{i=1}^n e_i^2) = \frac{2\sigma^4}{(n-p^*)^2}$ $V(\hat{\sigma}^2) = \frac{2MSE^2}{n-p^*}$