Homework 2

Alex Wako

2023-02-06

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 ...
```

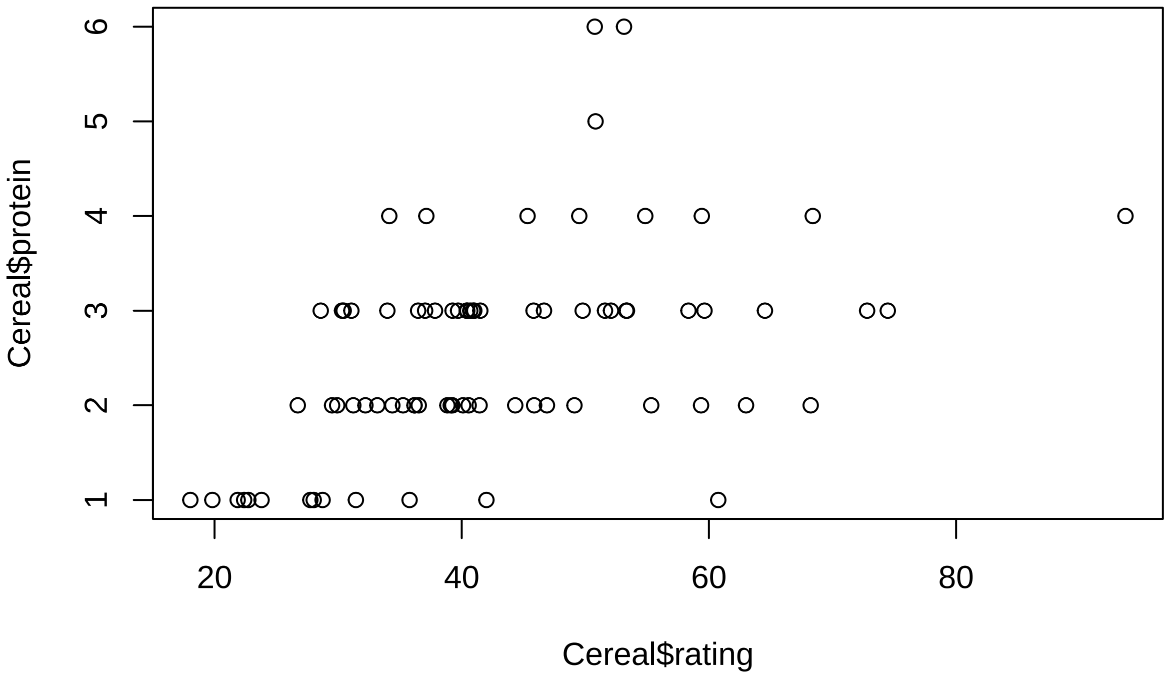
The data set *cereal* contains measurements for a set of 77 cereal brands. For this assignment only consider the following variables:

- 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?

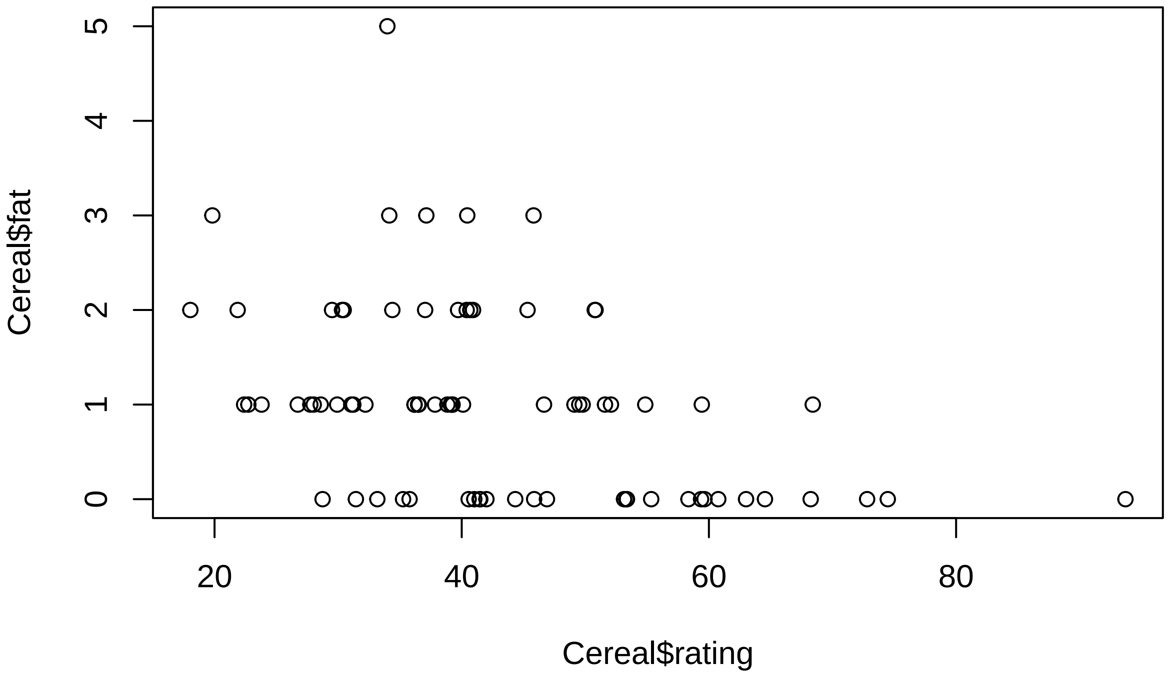
```
plot(Cereal$rating, Cereal$protein)
```



```
cor(Cereal$rating, Cereal$protein)
```

```
## [1] 0.4706185
```

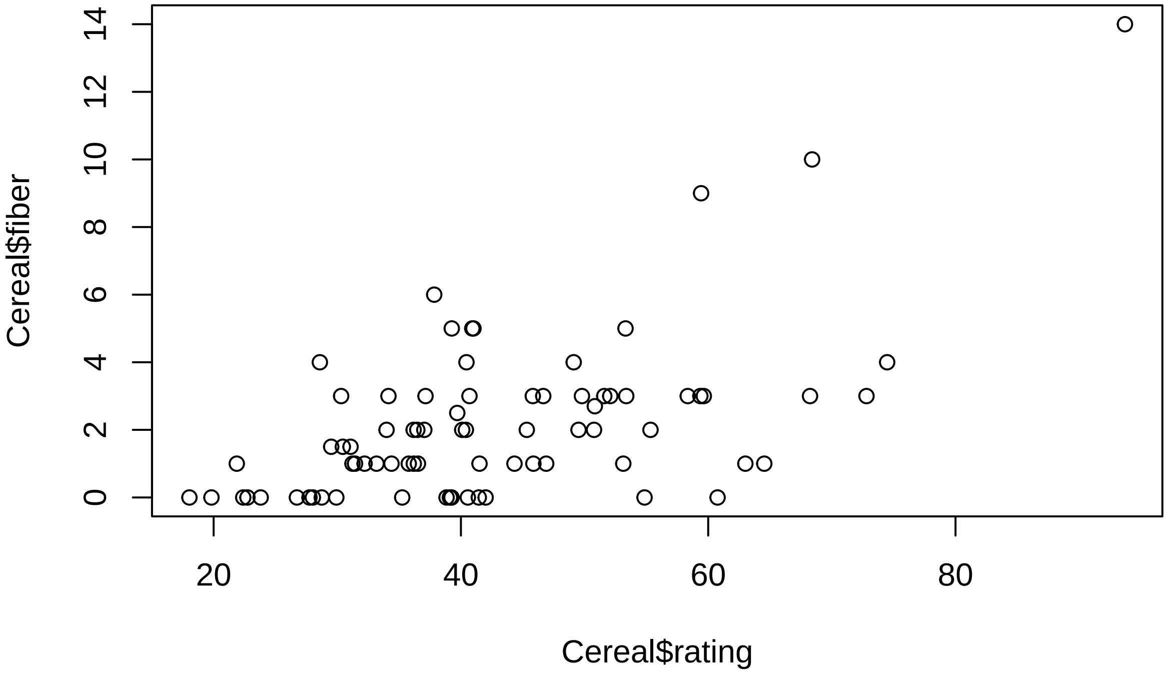
```
plot(Cereal$rating, Cereal$fat)
```



```
cor(Cereal$rating, Cereal$fat)
```

```
## [1] -0.4092837
```

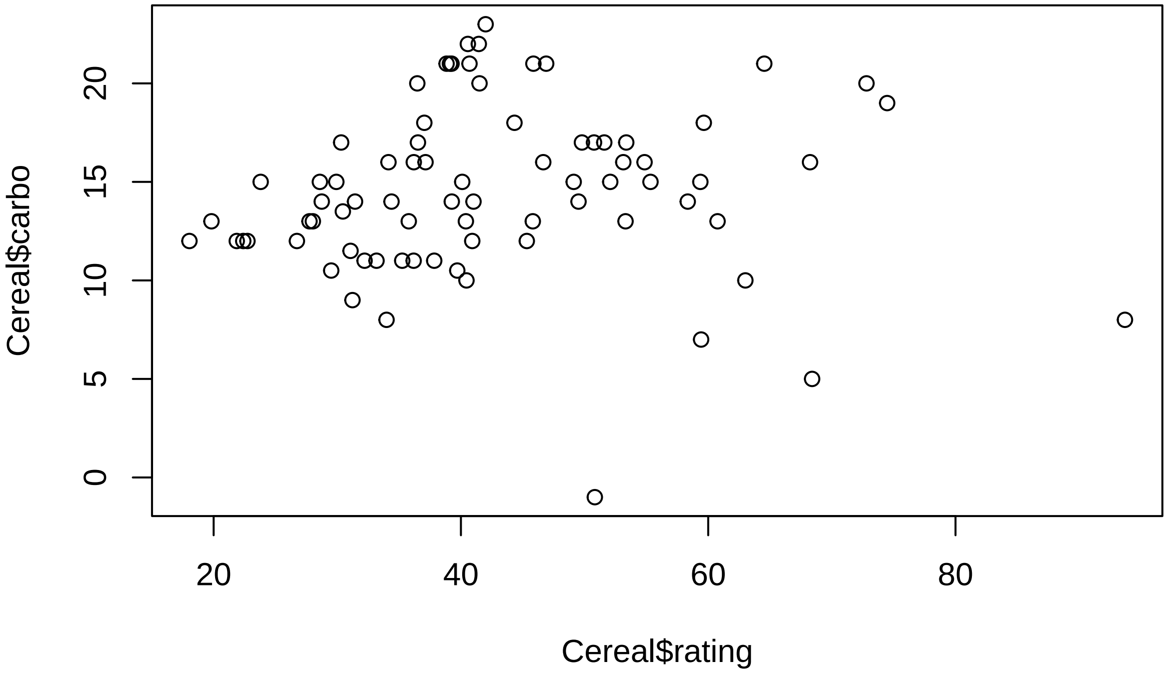
```
plot(Cereal$rating, Cereal$fiber)
```



```
cor(Cereal$rating, Cereal$fiber)
```

```
## [1] 0.5841604
```

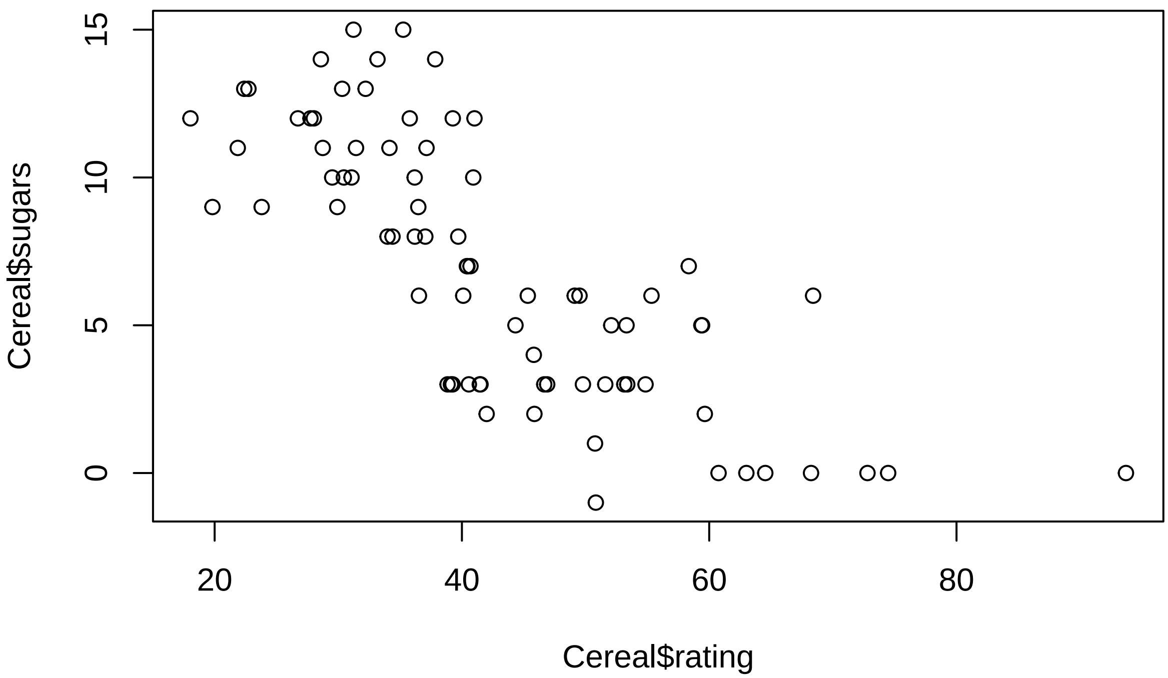
```
plot(Cereal$rating, Cereal$carbo)
```



```
cor(Cereal$rating, Cereal$carbo)
```

```
## [1] 0.05205466
```

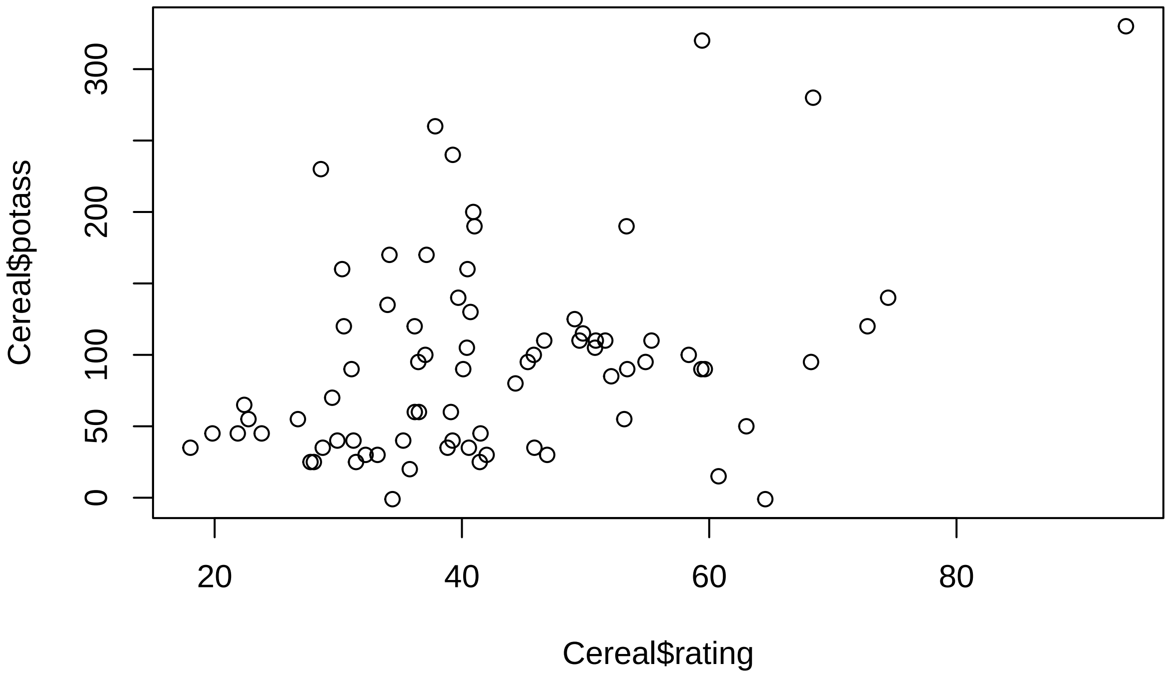
```
plot(Cereal$rating, Cereal$sugars)
```



```
cor(Cereal$rating, Cereal$sugars)
```

```
## [1] -0.7596747
```

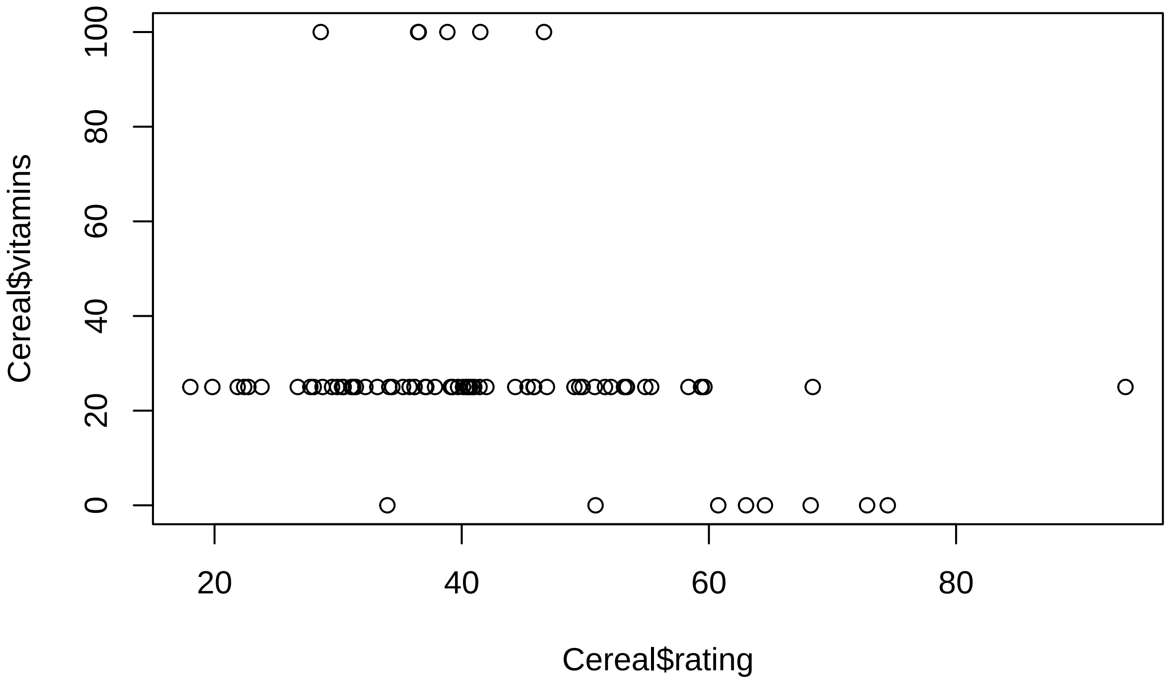
```
plot(Cereal$rating, Cereal$potass)
```



```
cor(Cereal$rating, Cereal$potass)
```

```
## [1] 0.3801654
```

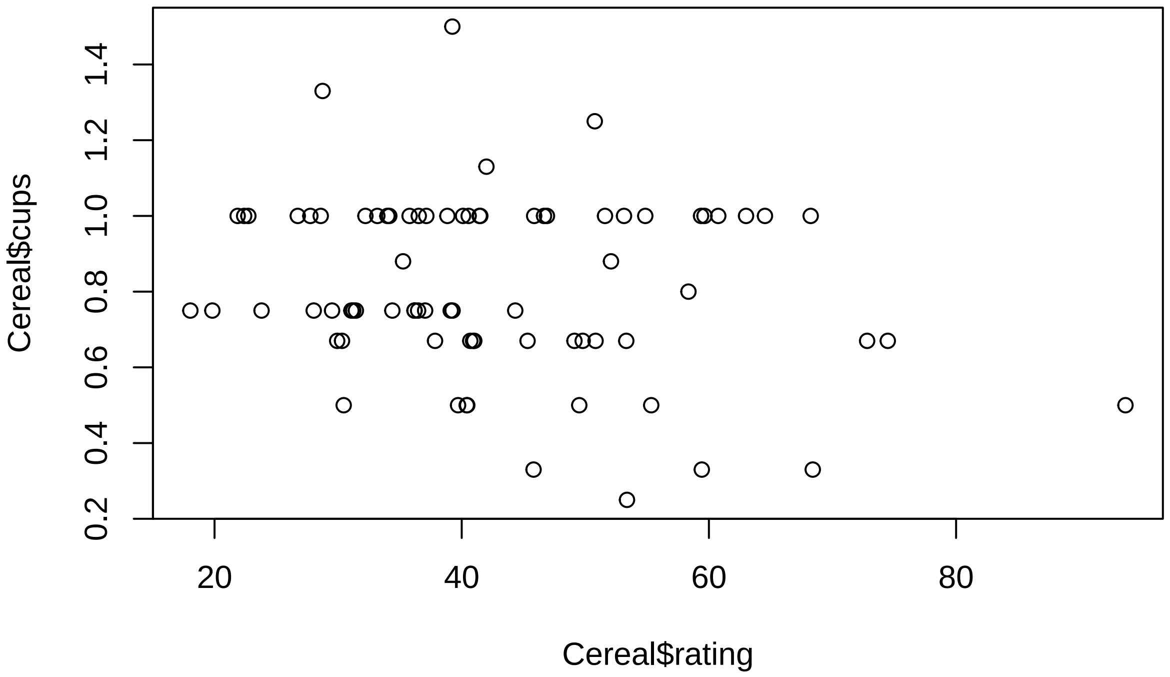
```
plot(Cereal$rating, Cereal$vitamins)
```



```
cor(Cereal$rating, Cereal$vitamins)
```

```
## [1] -0.2405436
```

```
plot(Cereal$rating, Cereal$cups)
```



```
cor(Cereal$rating, Cereal$cups)
```

```
## [1] -0.2031601
```

Looking at the plots, there seems to be no outlier but an outlier in the ratings. One of the cereals has a rating of 93.7 even though the range of most of the ratings sit between 20<y<70. The leverage should not be removed from the data because the leverage follows the linear trend of most of the predictors, but some predictors may be statistically insignificant to the best linear model.

(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.

```r
lm_Cereal1 <- lm(rating ~ protein + fat + fiber + carbo + sugars +
    potass + vitamins + cups, data = Cereal)
summary(lm_Cereal1)
```

```
##
## 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?

From the summary table, we can see that the variables of protein, fat, fiber, sugars, and vitamins 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?

The R-Squared value tells us that 90.37% of the total variation in the resposne 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?

The null hypothesis of the global F-test is when the model with no predictors and the model with predictors are the same. The global p-value of the F-test is shown in the summary as less than 2.2eˆ-16. The 7 predictor variables do explain a significant proportion of the variation in the response as more than 90% of the variaition is explained.

(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.

```
lm_Cereal2 <- lm(rating ~ protein + fat + fiber + sugars + potass +
    vitamins + cups, data = Cereal)
summary(lm_Cereal2)
```

```
##
## Call:
## lm(formula = rating ~ protein + fat + fiber + sugars + potass +
##     vitamins + cups, data = Cereal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2888  -3.2055  -0.4897   2.3898   9.2857
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51.27765    3.21836  15.933  < 2e-16 ***
## protein      1.96866    0.65699   2.996  0.00379 **
## fat         -3.98457    0.60743  -6.560 8.27e-09 ***
## fiber        3.26700    0.60295   5.418 8.29e-07 ***
## sugars      -1.67464    0.14720 -11.376  < 2e-16 ***
## potass      -0.02581    0.02086  -1.238  0.22009
## vitamins    -0.10352    0.02416  -4.286 5.79e-05 ***
## cups         0.46174    2.71375   0.170  0.86539
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.576 on 69 degrees of freedom
## Multiple R-squared:  0.9037, Adjusted R-squared:  0.8939
## F-statistic: 92.45 on 7 and 69 DF,  p-value: < 2.2e-16
```

As carbohydrates were not statistically significant to the original model, when the new model is created without the carbohydrate predictor, the coefficient of determination, 0.9037, and the p-value, less than 2.2eˆ-16, is still the same as the original model.

(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 will be when the model relating rating to vitamins and/or potassium does not differ to rating without relation to vitamins and/or potassium.

```
# Creating F-test
fullmodel <- lm(rating ~ protein + fat + fiber + sugars + potass +
    vitamins + cups, data = Cereal)
nullmodel <- lm(rating ~ protein + fat + fiber + sugars + cups,
    data = Cereal)
(anova <- anova(nullmodel, fullmodel))
```

```
## Analysis of Variance Table
##
## Model 1: rating ~ protein + fat + fiber + sugars + cups
## Model 2: rating ~ protein + fat + fiber + sugars + potass + vitamins +
##     cups
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     71 1883.3
## 2     69 1444.9  2    438.35 10.466 0.0001072 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Finding p-value
(pval <- 1 - pf(anova$F[2], 2, 69))
```

```
## [1] 0.0001071727
```

From the summary table and calculated p-value, we can conclude to reject the null hypothesis. The two predictors are statistically significant in our model and the p-value is too small to consider the null hypothesis.

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

```
lower_bound = lm_Cereal1$coefficients["protein"] - (0.66433 *
    qt(p = 0.005, df = 68, lower.tail = FALSE))
upper_bound = lm_Cereal1$coefficients["protein"] + (0.66433 *
    qt(p = 0.005, df = 68, lower.tail = FALSE))
```

Protein has a confidence interval of (0.201693, 3.72275).

(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

```
# Data frame for new predictor variables
new_Cereal = data.frame(protein = 3, fat = 5, fiber = 2, carbo = 13,
    sugars = 6, potass = 60, vitamins = 25, cups = 0.8)

# Prediction
prediction <- predict(lm_Cereal1, newdata = new_Cereal)
```

The predicted rating of the given predictor variables is 29.9280796.

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

```
predict(lm_Cereal1, newdata = new_Cereal, interval = "confidence",
    level = 0.95)
```

```
##        fit      lwr      upr
## 1 29.92808 24.43562 35.42054
```

The values printed from the function show the possible values of rating given the model with a 95% accuracy. This means that the model is not guaranteed to be correct and the true value of the rating ranges from 24.43562 to 35.42054.

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

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

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 $V(\hat{\sigma}^2)$.

$SSR = \hat{\epsilon}^T\hat{\epsilon} = (M\epsilon)^T(M\epsilon)$

$= \epsilon^T M\epsilon$

$\hat{\sigma}^2 = \frac{SSR}{n-p*}$

$E[\hat{\sigma}^2] = \frac{E(\epsilon^T M\epsilon)}{n-p*}$

$= \frac{E(\epsilon^T M\epsilon)}{n-p*}$

Since $\frac{\epsilon^T M\epsilon}{\sigma^2}$ is a Chi-squared distribution, the mean of $\epsilon^T M\epsilon$ is $(n-p*)\sigma^2$

$= \sigma^2$

$Bias[\sigma^2] = E[\hat{\sigma}^2] - \sigma^2 = \sigma^2 - \sigma^2 = 0$

$V(\hat{\sigma}^2) = \frac{1}{(n-p^*)^2} V(\epsilon^T M\epsilon)$

$V(\epsilon^T M\epsilon) = 2tr[(M\epsilon)^2] + 4\mu^T M\epsilon M\mu$

$V(\hat{\sigma}^2) = \frac{2tr[(M\epsilon)^2] + 4\mu^T M\epsilon M\mu}{(n-p^*)^2}$

Appendix

```
knitr::opts_chunk$set(echo = TRUE, tidy.opts = list(width.cutoff = 60),
    tidy = TRUE)
Cereal <- read.table("cereal.csv", header = T, sep = ",")
str(Cereal)
plot(Cereal$rating, Cereal$protein)
cor(Cereal$rating, Cereal$protein)
plot(Cereal$rating, Cereal$fat)
cor(Cereal$rating, Cereal$fat)
plot(Cereal$rating, Cereal$fiber)
cor(Cereal$rating, Cereal$fiber)
plot(Cereal$rating, Cereal$carbo)
cor(Cereal$rating, Cereal$carbo)
plot(Cereal$rating, Cereal$sugars)
cor(Cereal$rating, Cereal$sugars)
plot(Cereal$rating, Cereal$potass)
cor(Cereal$rating, Cereal$potass)
plot(Cereal$rating, Cereal$vitamins)
cor(Cereal$rating, Cereal$vitamins)
plot(Cereal$rating, Cereal$cups)
cor(Cereal$rating, Cereal$cups)
lm_Cereal1 <- lm(rating ~ protein + fat + fiber + carbo + sugars +
    potass + vitamins + cups, data = Cereal)
summary(lm_Cereal1)
lm_Cereal2 <- lm(rating ~ protein + fat + fiber + sugars + potass +
    vitamins + cups, data = Cereal)
```

```
summary(lm_Cereal2)
# Creating F-test
fullmodel <- lm(rating ~ protein + fat + fiber + sugars + potass +
    vitamins + cups, data = Cereal)
nullmodel <- lm(rating ~ protein + fat + fiber + sugars + cups,
    data = Cereal)
(anova <- anova(nullmodel, fullmodel))

# Finding p-value
(pval <- 1 - pf(anova$F[2], 2, 69))
lower_bound = lm_Cereal1$coefficients["protein"] - (0.66433 *
    qt(p = 0.005, df = 68, lower.tail = FALSE))
upper_bound = lm_Cereal1$coefficients["protein"] + (0.66433 *
    qt(p = 0.005, df = 68, lower.tail = FALSE))
# Data frame for new predictor variables
new_Cereal = data.frame(protein = 3, fat = 5, fiber = 2, carbo = 13,
    sugars = 6, potass = 60, vitamins = 25, cups = 0.8)

# Prediction
prediction <- predict(lm_Cereal1, newdata = new_Cereal)
predict(lm_Cereal1, newdata = new_Cereal, interval = "confidence",
    level = 0.95)
```