HW2_126

Zejie Gao

2023-02-07

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 50 110 110 110 130 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 ...
```

```
Cereal <- as.data.frame(Cereal)
```

(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?

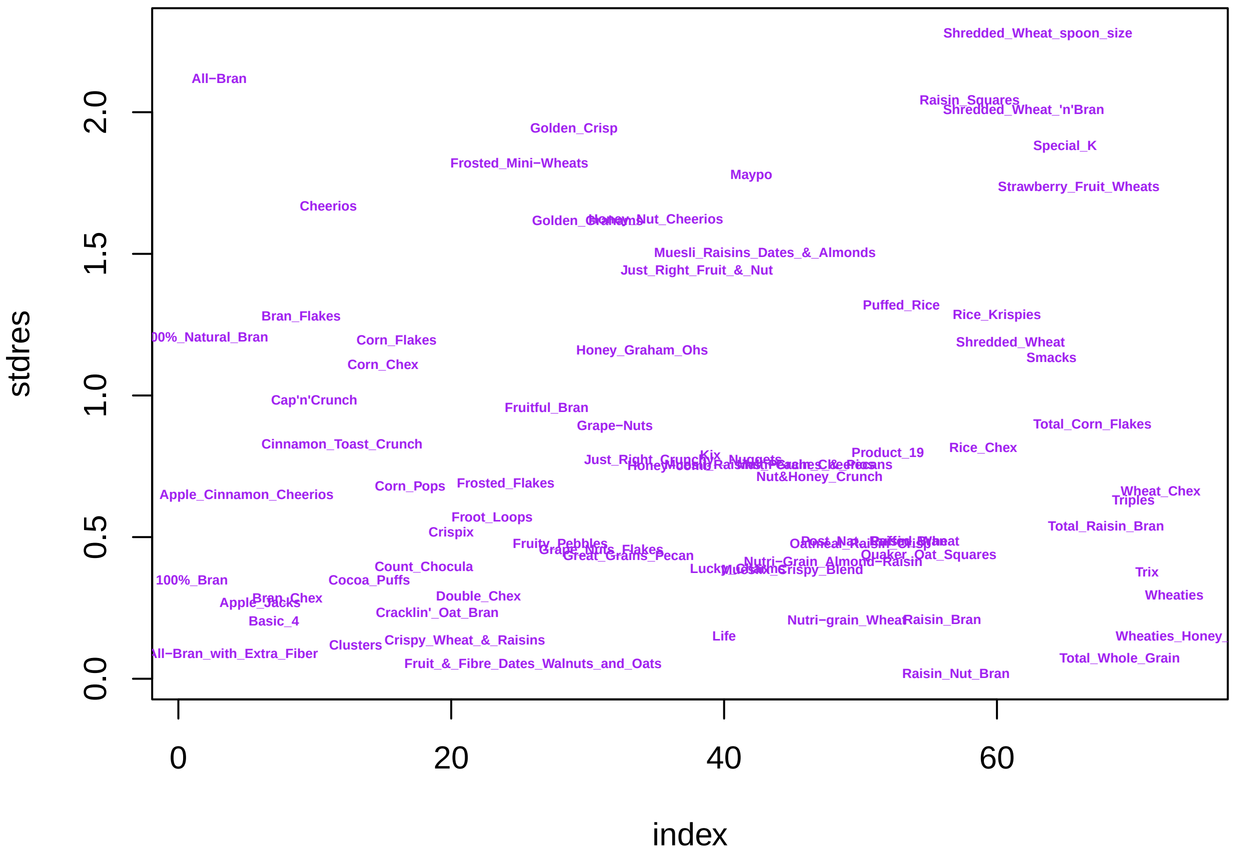
Based on the minimum of descriptive statistics, summary_statistics, numbers blow 0 in predictors carbo, sugars, and potass are impossible in the real life. Those value may comes from entry mistake, so we should remove them first.

```
Cereal <- Cereal|>
  select(name, rating, protein, fat, fiber, carbo, sugars, potass, vitamins
         , cups)
summary_statistics <- summary(Cereal)
summary_statistics
```

```
##      name               rating         protein           fat
##  Length:77          Min.   :18.04   Min.   :1.000   Min.   :0.000
##  Class :character   1st Qu.:33.17   1st Qu.:2.000   1st Qu.:0.000
##  Mode  :character   Median :40.40   Median :3.000   Median :1.000
##                     Mean   :42.67   Mean   :2.545   Mean   :1.013
##                     3rd Qu.:50.83   3rd Qu.:3.000   3rd Qu.:2.000
##                     Max.   :93.70   Max.   :6.000   Max.   :5.000
##      fiber            carbo          sugars           potass
##  Min.   : 0.000   Min.   :-1.0   Min.   :-1.000   Min.   : -1.00
##  1st Qu.: 1.000   1st Qu.:12.0   1st Qu.: 3.000   1st Qu.: 40.00
##  Median : 2.000   Median :14.0   Median : 7.000   Median : 90.00
##  Mean   : 2.152   Mean   :14.6   Mean   : 6.922   Mean   : 96.08
##  3rd Qu.: 3.000   3rd Qu.:17.0   3rd Qu.:11.000   3rd Qu.:120.00
##  Max.   :14.000   Max.   :23.0   Max.   :15.000   Max.   :330.00
##     vitamins           cups
##  Min.   :  0.00   Min.   :0.250
##  1st Qu.: 25.00   1st Qu.:0.670
##  Median : 25.00   Median :0.750
##  Mean   : 28.25   Mean   :0.821
##  3rd Qu.: 25.00   3rd Qu.:1.000
##  Max.   :100.00   Max.   :1.500
```

```r
Cereal_model <- Cereal|> # filter all the value greater than zero
  filter(carbo >= 0,
         sugars >= 0,
         potass >= 0
  )
```

```r
Cereal_fit <- lm(rating ~ protein + fat + fiber + carbo + sugars + potass
                 + vitamins + cups, data = Cereal_model)
r <- rstandard(Cereal_fit )
data.sres <- data.frame(index = seq(length(r)),
                        stdres = abs(r), names = Cereal_model$name)
par(mar = c(4, 4, 0.5, 0.5))
plot(stdres ~ index, data = data.sres, col = "White", pch = NULL)
text(stdres ~ index, labels = names, data = data.sres, cex = 0.4, font = 2
     , col = "purple")
abline(h = 3, col = "red", lty = 2)
```



Since we define the outliers as points that does not fit the model well, which means data point for which yi - yi_hat is large. By plotting standardized residuals vs index, we can find out that there is no outliers in our observation. This is because that none of the index satisfy the rule of thumb, which also equivalent to abs(r) >= 3.

(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.

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

```
##
## Call:
## lm(formula = rating ~ protein + fat + fiber + carbo + sugars +
##     potass + vitamins + cups, data = Cereal_model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1868 -3.0482 -0.4195  1.9820  9.3381
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 58.23622    4.45056  13.085  < 2e-16 ***
## protein      2.26615    0.64782   3.498 0.000851 ***
## fat         -4.12831    0.61233  -6.742 4.94e-09 ***
## fiber        2.63691    0.66389   3.972 0.000181 ***
## carbo       -0.39997    0.19262  -2.077 0.041805 *
## sugars      -1.90439    0.16915 -11.258  < 2e-16 ***
## potass      -0.01515    0.02233  -0.678 0.500020
## vitamins    -0.09156    0.02475  -3.699 0.000448 ***
## cups         0.50040    2.60766   0.192 0.848421
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.352 on 65 degrees of freedom
## Multiple R-squared:  0.9144, Adjusted R-squared:  0.9038
## F-statistic: 86.77 on 8 and 65 DF,  p-value: < 2.2e-16
```

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

protein, fat, fiber, sugars and vitamins are satistically significant under the significant threshold value of 0.01 because their p-valuea are less than 0.01. We successfully reject the H0: betai = 0

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

```r
r_squared <- summary(Cereal_fit)$r.squared
r_squared
```

```
## [1] 0.9143766
```

(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 that none of the predictor variables are statistically significant in explaining the response variable. The p-value is the probability of observing a test statistic at least as extreme as the observed results, assuming that the null hypothesis is true. It is often expressed as "Pr(>F)", which is the probability of observing a value larger than the F-statistic on the F-distribution. In the output provided, the p-value for the global F-test is less than 2.2e-16, which is far smaller than 0.05. Therefore, we reject the null hypothesis and conclude that the 8 predictor variables explain a significant proportion of the variation in the response.

```r
mod_M <- lm(rating ~ ., Cereal_model[,sapply(Cereal_model, is.numeric)])
# Larger model with all the predictors
mod_1 <- lm(rating ~ 1, Cereal_model[,sapply(Cereal_model, is.numeric)])
# Smaller model with only intercept
anova1 <- anova(mod_1, mod_M) ; anova1
```

```
## Analysis of Variance Table
##
## Model 1: rating ~ 1
## Model 2: rating ~ protein + fat + fiber + carbo + sugars + potass + vitamins +
##     cups
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     73 14377
## 2     65  1231  8     13146 86.767 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(f)(2pts) Consider testing the null hypothesis H0 : 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.

```
n <- dim(Cereal_model)[1]
p <- 8 # number of predictors
round(coefficients(summary(Cereal_fit)), 8)
```

```
##                Estimate Std. Error     t value   Pr(>|t|)
## (Intercept) 58.23622403 4.45055538  13.0851588 0.00000000
## protein      2.26614977 0.64782314   3.4980995 0.00085079
## fat         -4.12830850 0.61232731  -6.7419964 0.00000000
## fiber        2.63691160 0.66389237   3.9718962 0.00018110
## carbo       -0.39997157 0.19261759  -2.0765060 0.04180535
## sugars      -1.90439043 0.16915303 -11.2583876 0.00000000
## potass      -0.01514622 0.02233132  -0.6782504 0.50002046
## vitamins    -0.09155823 0.02475491  -3.6985887 0.00044786
## cups         0.50040163 2.60766381   0.1918965 0.84842149
```

```
pval_carbo = pt(q = -2.0765060 , df = n - p - 1) * 2
pval_carbo
```

```
## [1] 0.04180535
```

```
mod_M <- lm(rating ~ ., Cereal_model[,sapply(Cereal_model, is.numeric)])
mod_2 <- lm(rating ~ protein + fat + fiber +  sugars + potass + vitamins +cups
            ,Cereal_model[,sapply(Cereal_model, is.numeric)]) # Smaller model
anova2 <- anova(mod_2, mod_M); anova2
```

```
## Analysis of Variance Table
##
## Model 1: rating ~ protein + fat + fiber + sugars + potass + vitamins +
##     cups
## Model 2: rating ~ protein + fat + fiber + carbo + sugars + potass + vitamins +
##     cups
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     66 1312.7
## 2     65 1231.0  1    81.661 4.3119 0.04181 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pval_carbo1 <- 1 - pf(anova2$F[2], 1, 65); pval_carbo1
```

```
## [1] 0.04180535
```

(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 the F-test is that the coefficients for vitamins and potass are both equal to zero, and the alternative hypothesis is that at least one of them is non-zero. The output shows that the F-statistic is 6.9863 and the corresponding p-value is 0.001785. Since the p-value is less than 0.05, we reject the null hypothesis that both vitamins and potass coefficients are zero, and conclude that there is evidence of a significant relationship between at least one of the predictors and the rating response variable.

```
mod_M <- lm(rating ~ ., Cereal_model[,sapply(Cereal_model, is.numeric)])
mod_3 <- lm(rating ~ protein + fat + fiber + carbo + sugars + cups
            ,Cereal_model[,sapply(Cereal_model, is.numeric)])
anova3 <- anova(mod_3, mod_M); anova3
```

```
## 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     67 1495.6
## 2     65 1231.0  2    264.62 6.9863 0.001785 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(h)(3pts) Use the summary output to construct a 99% confidence interval for beta(protein). What is the interpretation of this confidence interval? Interpretation: If we were to repeat the sampling and regression process many times, 99% of the resulting intervals would contain the true value of beta(protein). It is also an indicator of the precision of our estimate for the protein variable.

```
confint(Cereal_fit, "protein", level = 0.99)
```

```
##              0.5 %   99.5 %
## protein 0.5470834 3.985216
```

(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

```
new_data <- data.frame(protein=3, fat=5, fiber=2, carbo=13, sugars=6, potass=60
                       , vitamins=25, cups=0.8)
predicted_rating <- predict(Cereal_fit, newdata=new_data); predicted_rating
```

```
##        1
## 30.24357
```

(j). (3pts) What is the 95% prediction interval for the observation in part (i)? What is the interpretation of this prediction interval? If we were to repeat the sampling and regression process many times, 95% of the resulting intervals would contain the true value of the predicted response variable (y), which means that there is a 95% chance that the true value of the response variable for a new observation falls between 19.96214 and 40.525.

```
predicted_CIs_rating <- predict(Cereal_fit, newdata=new_data
                                , interval = "prediction")
predicted_CIs_rating
```

```
##        fit      lwr    upr
## 1 30.24357 19.96214 40.525
```

2.(20pts) Consider the MLR model with p predictors:

$$
\begin{aligned}
& E\left(\hat{\sigma}\right) = E\left(\frac{SSR}{n-p^*}\right) \\
& = E\left(\frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-p^*}\right) \qquad \text{since } \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{\sigma^2} \sim X^2\left(n-p^*\right) \\
& = \frac{1}{n-p^*} E\left(\hat{\varepsilon}^T\hat{\varepsilon}\right) \qquad \text{Thus, } E\left(\frac{\hat{\varepsilon}^T\hat{\varepsilon}}{\sigma^2}\right) = n-p^* \\
& = \frac{\sigma^2}{n-p^*} E\left(\frac{\hat{\varepsilon}^T\hat{\varepsilon}}{\sigma^2}\right) \\
& = \frac{\sigma^2}{n-p^*}\left(n-p^*\right) \\
& = \sigma^2
\end{aligned}
$$

$$
\begin{aligned}
& \operatorname{Var}\left(\hat{\sigma}^2\right) = \operatorname{Vav}\left(\frac{SSR}{n-p^*}\right) \\
& = \operatorname{Var}\left(\frac{\hat{\varepsilon}^T\varepsilon}{n-p^*}\right) \\
& = \frac{1}{\left(n-p^*\right)^2} \operatorname{Var}\left(\hat{\varepsilon}^T\hat{\varepsilon}\right) \\
& = \frac{\sigma^4}{\left(n-p^*\right)^2} \operatorname{Var}\left(\frac{\hat{\varepsilon}^T\hat{\varepsilon}}{\sigma^2}\right) \\
& = \frac{\sigma^4}{\left(n-p^*\right)^2} 2 \times \left(n-p^*\right) \\
& = \frac{2\sigma^4}{\left(n-p^*\right)} \qquad \operatorname{Var}\left(\frac{\hat{\varepsilon}^T\hat{\varepsilon}}{\sigma^2}\right) = 2 \times \left(n-p^*\right)
\end{aligned}
$$