Homework 3

PSTAT Summer 2024

Due date: July 26th, 2024 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.

```
Cereal <- read.table("cereal.txt",header=T)
Cereal <- Cereal[-c(5, 21 ,58), ]
cereal_model <- lm(rating~protein+fat+fiber+carbo+sugars+potass+vitamins+cups, Cereal)
head(cereal_model$fitted.values, 5)
```

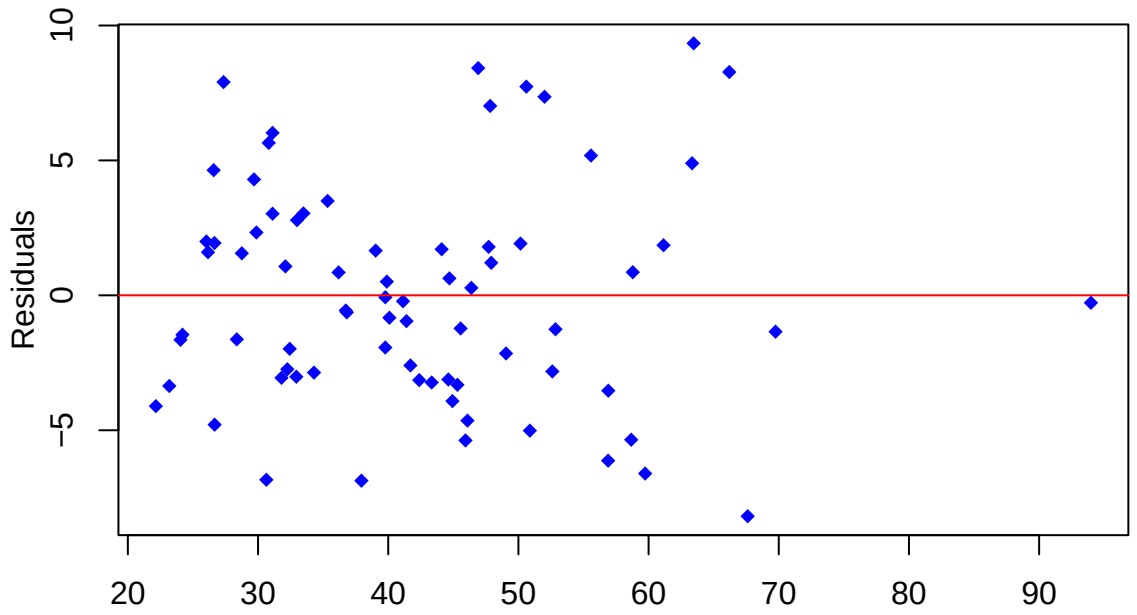
```
##        1        2        3        4        6
## 69.75066 29.68772 67.61235 93.98080 32.24978
```

```
head(cereal_model$residuals, 5)
```

```
##          1          2          3          4          6
## -1.3476910  4.2959597 -8.1868456 -0.2758917 -2.7402368
```

(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.

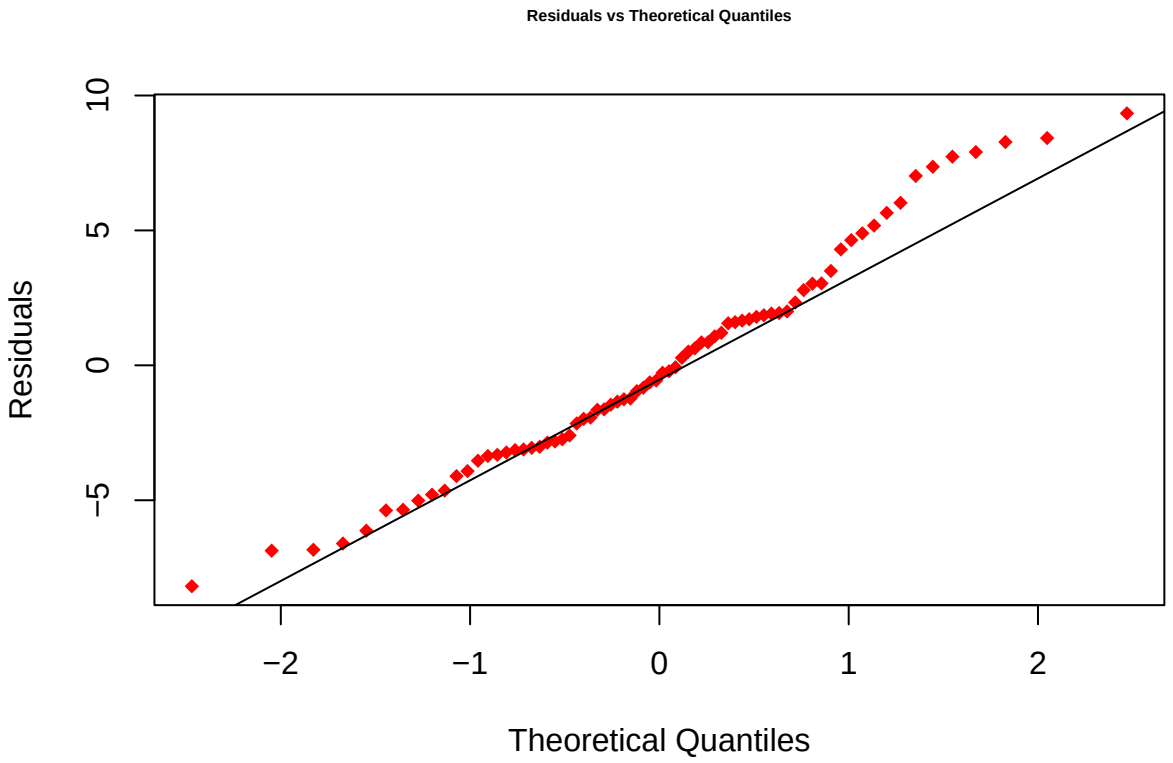
```
plot(fitted(cereal_model), residuals(cereal_model), xlab="",ylab="", col="blue", pch=18)
mtext(side=2, text="Residuals", line=2)
abline(h=0, col="red")
```



I used a plot of the residuals (vertical axis) vs. the fitted values (horizontal axis) to test if the data shows Homoscedasticity, or constant symmetrical variance. It seems that the random errors for the data does indeed have constant variance.

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

```
qqnorm(residuals(cereal_model), ylab="Residuals",main="", pch=18, col="red")
qqline(residuals(cereal_model))
title("Residuals vs Theoretical Quantiles", cex.main=0.5)
```



Because the points on the qq-plot lie mostly along the straight diagonal line with some minor deviations towards the ends, we can safely assume that the data is normally distributed.

(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?

The null hypothesis for this test is H_0 :The residuals are normally distributed.

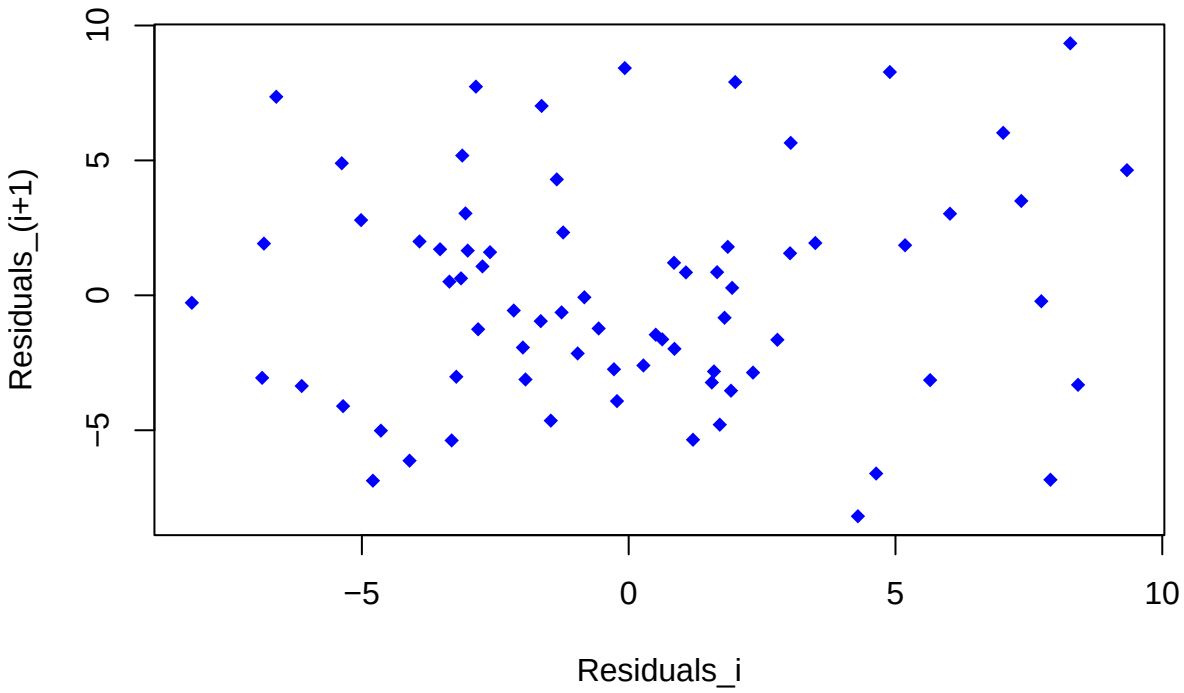
```
shapiro.test(residuals(cereal_model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(cereal_model)
## W = 0.97607, p-value = 0.1728
```

Assuming our $\alpha = .05$, we will not reject the null hypothesis because our p-value is 0.1728.

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

```
plot(residuals(cereal_model)[-c(length(residuals(cereal_model)))],residuals(cereal_model)[-c(1)], xlab=
```



There does not seem to be any correlation between the pairs of residuals.

(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?

The null hypothesis for the Durvin-Watson test is H_0 :Uncorrelated errors.

```
dwtest(cereal_model)
```

```
##
##  Durbin-Watson test
##
## data:  cereal_model
## DW = 1.8414, p-value = 0.2041
## alternative hypothesis: true autocorrelation is greater than 0
```

Assuming our $\alpha = .05$, we will not reject the null hypothesis because our p-value is 0.2041.

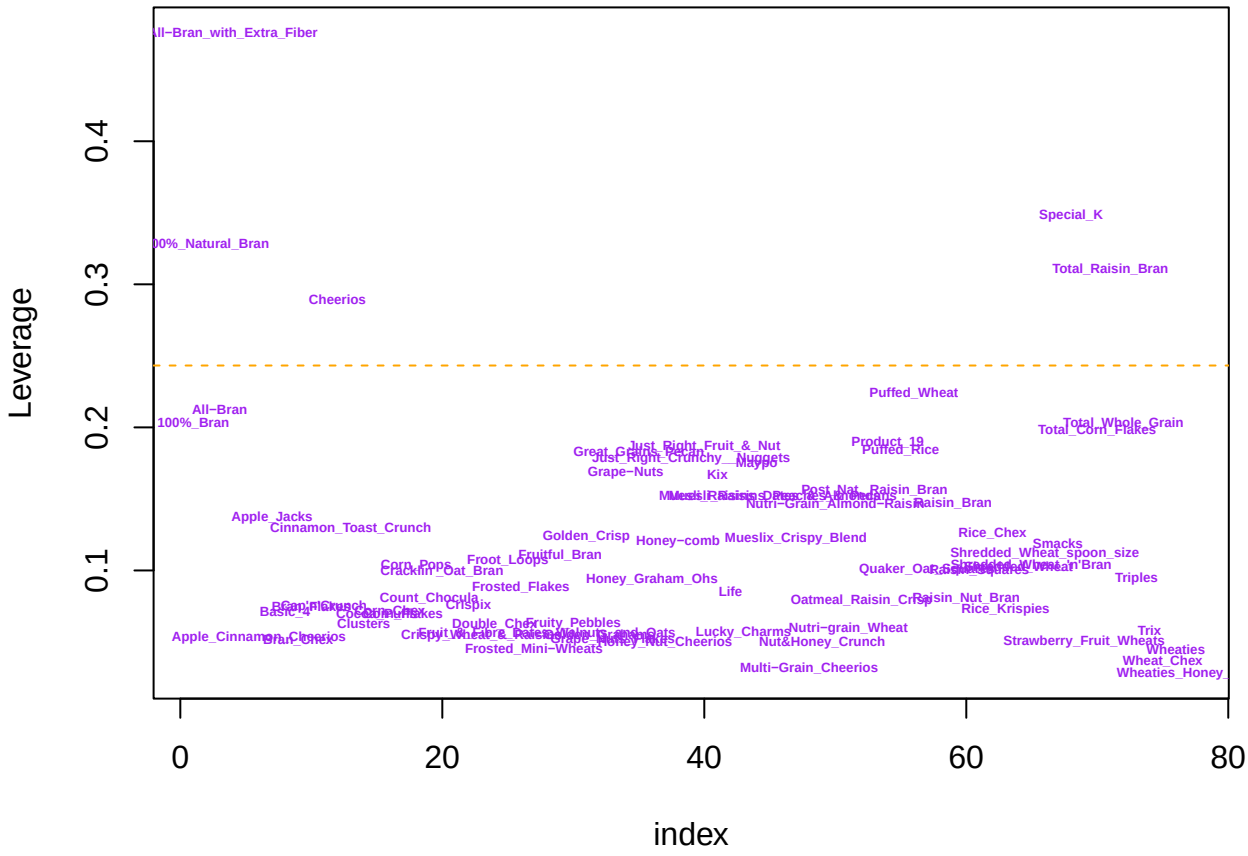
(g) **(2 pts)** Compute the hat matrix $\boldsymbol{H}$ in this data set (you don't need to show the entire matrix). Verify numerically that $\sum_{i=1}^{n} H_{ii} = p^* = p + 1$.

```
X <- model.matrix(cereal_model)
cereal_hat <- X%*%solve(t(X)%*%X)%*%t(X)
all.equal(sum(diag(cereal_hat)), length(coefficients(cereal_model)))
```

```
## [1] TRUE
```

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

```
cereal_leverage <- data.frame(index=c(as.integer(rownames(Cereal))), Leverage=diag(cereal_hat), namesC=(
par(mar = c(4, 4, 0.5, 0.5))
plot(Leverage ~ index, data=cereal_leverage, col="white", pch=NULL)
text(Leverage ~ index, labels=namesC, data=cereal_leverage, cex=0.4, font=2, col="purple")
abline(h = 2*sum(diag(cereal_hat))/dim(cereal_leverage)[1], col="orange", lty=2)
```



We can see that rows 100%_Natural_Bran, All-Bran_with_Extra_Fiber, Cheerios, Special_K, and Total_Raisin_Bran are all high-leverage points. If any leverage values were larger than $2*\bar{h}$, they are considered large.

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

```
observation <- 1
s_residuals <- rstandard(cereal_model)

for(x in s_residuals) {
  if(abs(x) > 3 || abs(x) == 3){
      print(glue("Observation number {observation} is an outlier with a standardized residual value {x}"
  }
  observation <- observation + 1
}
```

If the absolute value of any standard residuals were greater than or equal to 3, they are considered outliers. There are no data points that are outliers.

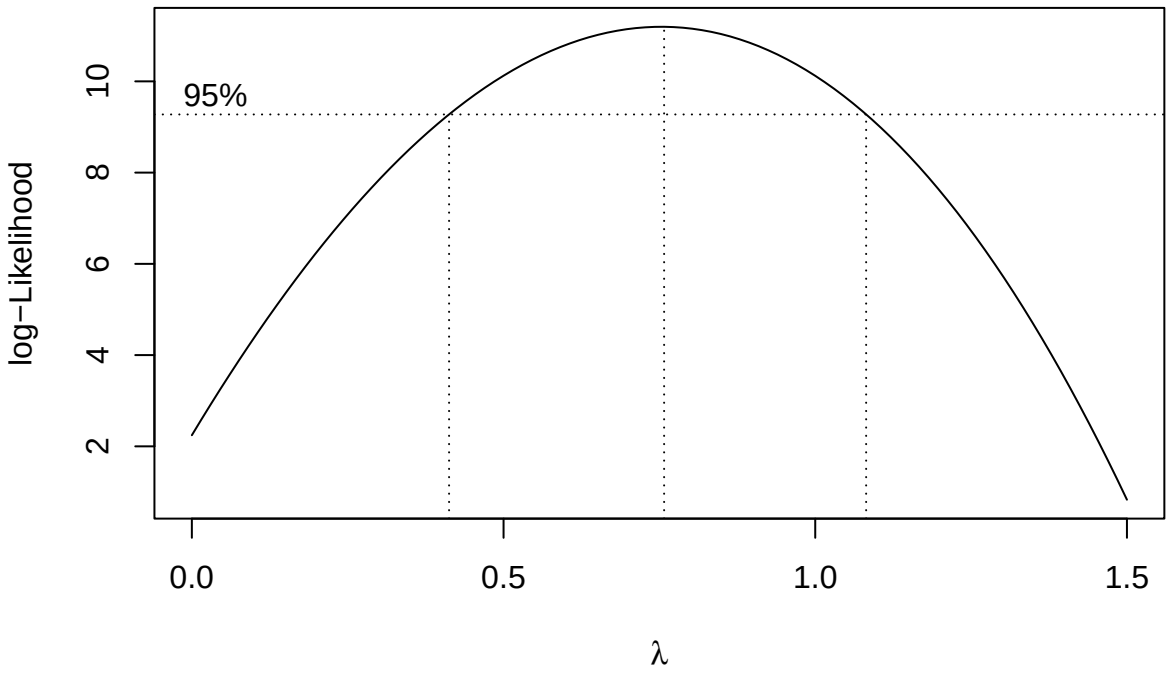
(j) **(2 pts)** Calculate the Cook's distance. How many observations in this data set have a Cook's distance that is greater than $4/n$?

```
observations <- 0
cook_distance <- cooks.distance(cereal_model)
for(x in cook_distance){
  if(x > (4/length(cook_distance))){
    observations <- observations + 1
  }
}
print(glue("There are {observations} total observations with cook's distance greater than 4/n"))
```

```
## There are 7 total observations with cook's distance 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?

```
boxcox(cereal_model, lambda=seq(0, 1.5, by=0.5))
```



Because 1 is contained in the 95% CI for $\hat{\lambda}$, no transformation is necessary.