CMTH 642 Data Analytics: Advanced Methods

Assignment 2 (10%)

Rui Zhang

500736315

```r
USDA_Clean <- read.csv("USDA_Clean.csv")
```

1. Read the csv file (USDA__Clean.csv) in the folder and assign it to a data frame. (3 points)

```r
str(USDA_Clean)
```

2. Check the datatypes of the attributes. (3 points)

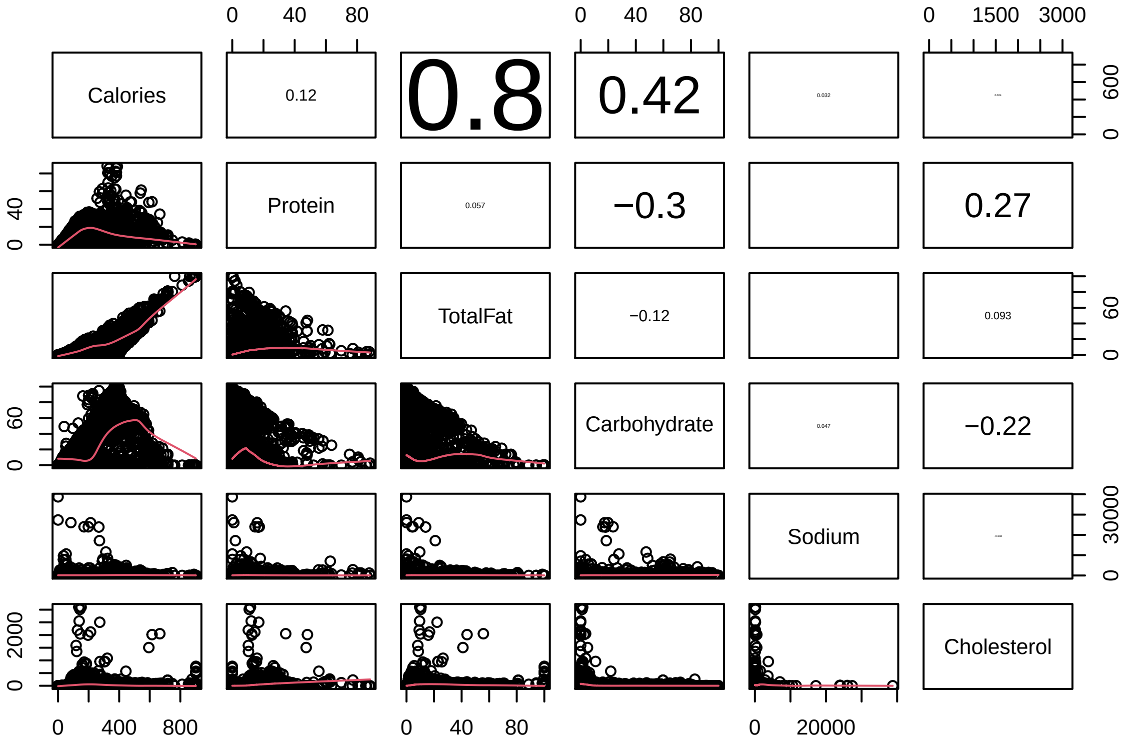
```
## 'data.frame':    6310 obs. of  21 variables:
##  $ X           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ ID          : int  1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 ...
##  $ Description : chr  "BUTTER,WITH SALT" "BUTTER,WHIPPED,WITH SALT" "BUTTER OIL,ANHYDROUS" "CHEESE,B
##  $ Calories    : int  717 717 876 353 371 334 300 376 403 387 ...
##  $ Protein     : num  0.85 0.85 0.28 21.4 23.24 ...
##  $ TotalFat    : num  81.1 81.1 99.5 28.7 29.7 ...
##  $ Carbohydrate: num  0.06 0.06 0 2.34 2.79 0.45 0.46 3.06 1.28 4.78 ...
##  $ Sodium      : int  714 827 2 1395 560 629 842 690 621 700 ...
##  $ Cholesterol : int  215 219 256 75 94 100 72 93 105 103 ...
##  $ Sugar       : num  0.06 0.06 0 0.5 0.51 ...
##  $ Calcium     : int  24 24 4 528 674 184 388 673 721 643 ...
##  $ Iron        : num  0.02 0.16 0 0.31 0.43 0.5 0.33 0.64 0.68 0.21 ...
##  $ Potassium   : int  24 26 5 256 136 152 187 93 98 95 ...
##  $ VitaminC    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VitaminE    : num  2.32 2.32 2.8 0.25 0.26 ...
##  $ VitaminD    : num  1.5 1.5 1.8 0.5 0.5 ...
##  $ HighSodium  : int  1 1 0 1 1 1 1 1 1 1 ...
##  $ HighCals    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ HighSugar   : int  0 0 0 0 0 0 0 1 0 1 ...
##  $ HighProtein : int  0 0 0 1 1 1 1 1 1 1 ...
##  $ HighFat     : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
USDA_Clean_Select <- USDA_Clean[,c("Calories", "Protein", "TotalFat", "Carbohydrate", "Sodium", "Choles
cor(USDA_Clean_Select)
```

3. Visualize the correlation among Calories, Protein, Total Fat, Carbohydrate, Sodium and Cholesterol. (7 points)

```
##                Calories      Protein     TotalFat Carbohydrate       Sodium
## Calories     1.00000000  0.122122537  0.804495022   0.42460618  0.032321026
## Protein      0.12212254  1.000000000  0.057035611  -0.30471117 -0.003489485
## TotalFat     0.80449502  0.057035611  1.000000000  -0.12434291  0.002916089
## Carbohydrate 0.42460618 -0.304711167 -0.124342914   1.00000000  0.046838692
## Sodium       0.03232103 -0.003489485  0.002916089   0.04683869  1.000000000
## Cholesterol  0.02391933  0.269854840  0.093289601  -0.21937986 -0.017774863
##              Cholesterol
## Calories      0.02391933
## Protein       0.26985484
## TotalFat      0.09328960
## Carbohydrate -0.21937986
## Sodium       -0.01777486
## Cholesterol   1.00000000
```

```
panel.cor <- function(x, y, ...)
{
  par(usr = c(0, 1, 0, 1))
  txt <- as.character(format(cor(x, y), digits=2))
  text(0.5, 0.5, txt, cex = 6* abs(cor(x, y)))
}
pairs(USDA_Clean_Select,lower.panel=panel.smooth, upper.panel=panel.cor)
```



```r
cor(USDA_Clean[,c("Calories", "TotalFat")])
```

4. Is the correlation between Calories and Total Fat statistically significant? Why? (7 points)

```
##          Calories TotalFat
## Calories 1.000000 0.804495
## TotalFat 0.804495 1.000000
```

```r
# The correlation between Calories and Total Fat statistically is significant, the correlation coeffici
```

```r
lm(Calories~Protein+TotalFat+Carbohydrate+Sodium+Cholesterol, data=USDA_Clean)
```

5. Create a Linear Regression Model, using Calories as the dependent variable Protein, Total Fat, Carbohydrate, Sodium and Cholesterol as the independent variables. (7 points)

```
##
## Call:
## lm(formula = Calories ~ Protein + TotalFat + Carbohydrate + Sodium +
##     Cholesterol, data = USDA_Clean)
```

```
##
## Coefficients:
##  (Intercept)       Protein      TotalFat  Carbohydrate        Sodium
##    3.9882753     3.9891994     8.7716980     3.7432001     0.0003383
##  Cholesterol
##    0.0110138
```

```
# Calories = 3.9882753 + 3.9891994xProtein + 8.7716980xTotalFat + 3.7432001xCarbohydrate + 0.0003383xSo
```

6. Write the Linear Regression Equation, using Calories as the dependent variable whereas Protein, TotalFat, Carbohydrate, Sodium and Cholesterol as the independent variables. (7 points)

```
calory.lm <- lm(Calories~Protein+TotalFat+Carbohydrate+Sodium+Cholesterol,data=USDA_Clean)
anova(calory.lm)
```

7. Which independent variable is the least significant? Why? (7 points)

```
## Analysis of Variance Table
##
## Response: Calories
##                Df    Sum Sq   Mean Sq    F value    Pr(>F)
## Protein         1   2728899   2728899 7.6197e+03 < 2.2e-16 ***
## TotalFat        1 116762840 116762840 3.2603e+05 < 2.2e-16 ***
## Carbohydrate    1  61215495  61215495 1.7093e+05 < 2.2e-16 ***
## Sodium          1       789       789 2.2031e+00    0.1378
## Cholesterol     1     11014     11014 3.0753e+01  3.05e-08 ***
## Residuals    6304   2257685       358
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
summary(aov(Calories~Protein+TotalFat+Carbohydrate+Sodium+Cholesterol, data=USDA_Clean))
```

```
##                Df    Sum Sq   Mean Sq   F value   Pr(>F)
## Protein         1   2728899   2728899 7.620e+03  < 2e-16 ***
## TotalFat        1 116762840 116762840 3.260e+05  < 2e-16 ***
## Carbohydrate    1  61215495  61215495 1.709e+05  < 2e-16 ***
## Sodium          1       789       789 2.203e+00    0.138
## Cholesterol     1     11014     11014 3.075e+01 3.05e-08 ***
## Residuals    6304   2257685       358
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
# Sodium is the least significant because its p value is 0.1378 bigger than 0.1
```

```r
pred.calory = predict(calory.lm, data.frame(Protein=0.1,TotalFat=35,Carbohydrate=405,Sodium=440,Cholest
pred.calory
```

8. A new product is just produced with the following data: Protein=0.1, TotalFat=35, Carbohydrate=405, Sodium=440, Cholesterol=70, Sugar=NA, Calcium=35, Iron=NA, Potassium=35, VitaminC=10, VitaminE=NA, VitaminD=NA. Based on the model you created, what is the predicted value for Calories? (7 points)

```
##        1
## 1828.312
```

```r
Calories <- 3.9882753 + 3.9891994*0.1 + 8.7716980*35 + 3.7432001*405 + 0.0003383*440 + 0.0110138*70
Calories
```

```
## [1] 1828.312
```

```r
(44440-440)/440*100
```

9. If the Sodium amount increases from 440 to 44440 (10000% increase), how much change will occur on Calories in percent? Explain why? (7 points)

```
## [1] 10000
```

```r
pred.calory10000 = predict(calory.lm, data.frame(Protein=0.1,TotalFat=35,Carbohydrate=405,Sodium=44440,
(pred.calory10000 - pred.calory)/pred.calory
```

```
##           1
## 0.008141547
```

```r
# calories changes by 0.814%, sodium has little smallest coefficient 0.0003383, it has little effect on
```

10. A study of primary education asked elementaty school students to retell two book articles that they read earlier in the week. The first (Article 1) had no pictures, and the second (Article 2) was illustrated with pictures. An expert listened to recordings of the students retelling each article and assigned a score for certain uses of language. Higher scores are better. Here are the data for five readers in this study:

Article 1 0.40 0.72 0.00 0.36 0.55

Article 2 0.77 0.49 0.66 0.28 0.38

```r
#$H_0$: the median score from two book articles are identical
#$H_a$:  the median score from two book articles are different
```

A) What are $H_0$ and $H_a$ ? (5 points)

```
# this is paired experiement
```

B) Is this a paired or unpaired experiment? (5 points)

```
# Wilcoxon signed-rank test for a paired experiment
```

C) Based on your previous answer, which nonparametric test statistic would you use to compare the medians of Article 1 and Article 2. (5 points)

```
article1 <- c(0.40, 0.72, 0.00, 0.36, 0.55)
article2 <- c(0.77, 0.49, 0.66, 0.28, 0.38)
test <- wilcox.test(article1, article2, alternative="two.sided", paired=TRUE)
test
```

D) Use a nonparametric test statistic to check if there is a statistically significant difference between the medians of Article 1 and Article 2. (5 points)

```
##
##  Wilcoxon signed rank exact test
##
## data:  article1 and article2
## V = 6, p-value = 0.8125
## alternative hypothesis: true location shift is not equal to 0
```

```
# Don’t reject hypotheses Since the p-value 0.8123 is greater than 0.05, we can say that the medians of
```

E) Will you accept or reject your Null Hypothesis? $(\alpha = 0.05)$ Do illustrations improve how the students retell an article or not? Why? (5 points)

11. Two companies selling toothpastes with the lable of 100 grams per tube on the package. We randomly bought eight toothpastes from each company A and B from random stores. Afterwards, we scaled them using high precision scale. Our measurements are recorded as follows:

Company A: 97.1 101.3 107.8 101.9 97.4 104.5 99.5 95.1

Company B: 103.5 105.3 106.5 107.9 102.1 105.6 109.8 97.2

```
# this is unpaired experiment
```

A) Is this a paired or unpaired experiment? (5 points)

```
# Wilcoxon rank sum exact test
```

B) Based on your previous answer, which nonparametric test statistic would you use to compare the medians of Company A and Company B. (5 points)

```
companyA <- c(97.1, 101.3, 107.8, 101.9, 97.4, 104.5, 99.5, 95.1)
companyB <- c(103.5, 105.3, 106.5, 107.9, 102.1, 105.6, 109.8, 97.2)
companies <- c(companyA, companyB)
wilcox.test(companyA, companyB, alternative="two.sided")
```

C) Use a nonparametric test statistic to check if there is a statistically significant difference between the medians of Company A and Company B. (5 points)

```
##
##  Wilcoxon rank sum exact test
##
## data:  companyA and companyB
## W = 13, p-value = 0.04988
## alternative hypothesis: true location shift is not equal to 0
```

```
# We reject the Null hypothesis
# Since the p-value 0.04988 is less than 0.05, we conclude that there is a statistically significant di
```

D) Will you accept or reject your Null Hypothesis? $(\alpha = 0.05)$ **Are packaging process similar or different based on weight measurements? Why? (5 points)** This is the end of Assignment 2

Ceni Babaoglu, PhD