CMTH 642 Data Analytics: Advanced Methods

Assignment 2 (10%)

[Ilhak Park]

[DJ0 & 501072432]

```
USDAClean <-read.csv(file="USDA_Clean.csv", sep =",")
```

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

```
sapply (USDAClean, class)
```

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

```
##            X           ID  Description     Calories      Protein     TotalFat
##    "integer"    "integer"  "character"    "integer"    "numeric"    "numeric"
## Carbohydrate       Sodium  Cholesterol        Sugar      Calcium         Iron
##    "numeric"    "integer"    "integer"    "numeric"    "integer"    "numeric"
##    Potassium     VitaminC     VitaminE     VitaminD   HighSodium     HighCals
##    "integer"    "numeric"    "numeric"    "numeric"    "integer"    "integer"
##    HighSugar  HighProtein      HighFat
##    "integer"    "integer"    "integer"
```

```
subset <- USDAClean[,c("Calories","Protein","TotalFat","Carbohydrate", "Sodium","Cholesterol")]
cor(subset)
```

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

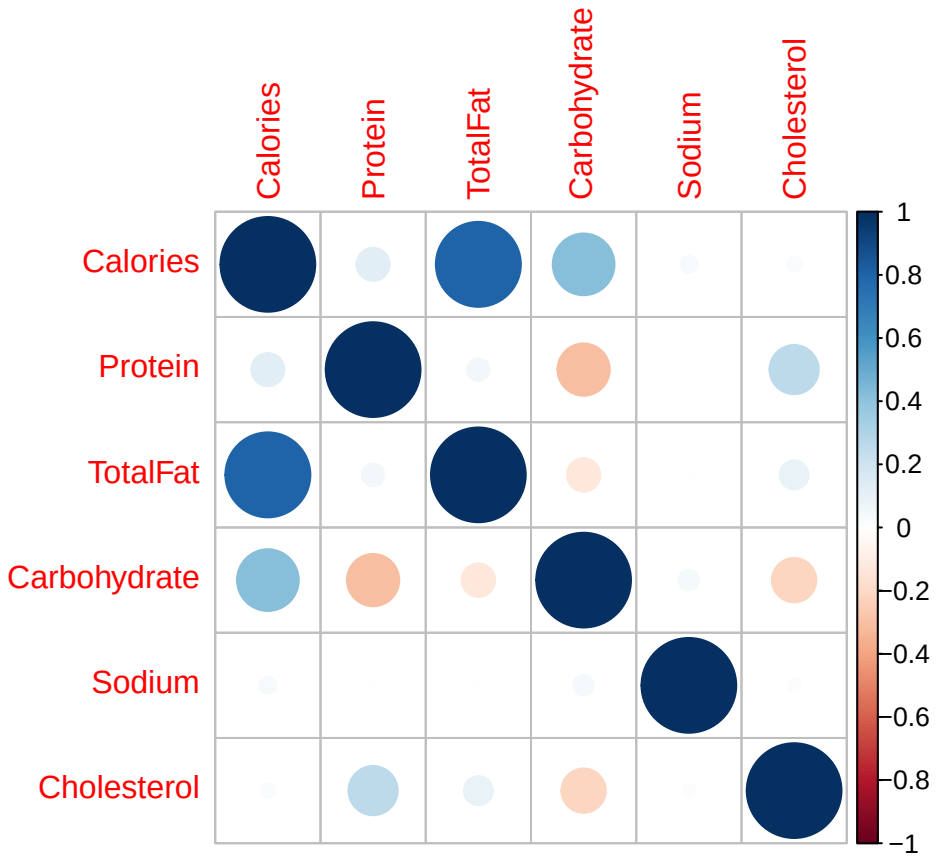
```
##                Calories      Protein     TotalFat Carbohydrate       Sodium
## Calories     1.00000000  0.122122537  0.804495022   0.42460618  0.032321026
## Protein      0.12212254  1.000000000  0.057035611  -0.30471117 -0.003489485
## TotalFat     0.80449502  0.057035611  1.000000000  -0.12434291  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
```

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
corrplot(cor(subset), method="circle")
```




```
cor (USDAClean$Calories, USDAClean$TotalFat)
```

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

```
## [1] 0.804495
```

```
#The correlation between Calories and Total Fat statistically significant because
#their correlation coefficient is about 0.804495 which is close to a perfect positive correlation, 1.0.
```

```
Calories_lm <- lm(Calories~ Protein+TotalFat+Carbohydrate+Sodium+Cholesterol, data= USDAClean)
Calories_lm
```

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

```
Calories_lm$coefficients
```

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)

```
##  (Intercept)      Protein     TotalFat Carbohydrate       Sodium  Cholesterol
## 3.9882752613 3.9891994394 8.7716980068 3.7432000604 0.0003383021 0.0110138110
```

```
#Calories=3.9882752613+3.9891994394*Protein+8.7716980068*TotalFat
#+3.7432000604*Carbohydrate+0.0003383021*Sodium+0.0110138110*Cholesterol
```

```
anova(Calories_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
```

```
#The sodium is the least significant because it has a p-value of 0.1378,
#while independent variables have much smaller p-values.
#When an independent variable has a p-value of 0.1378,
#it is difficult to predict changes to the data.
```

```
NewData = data.frame(Protein=0.1, TotalFat=35, Carbohydrate=405, Sodium=440, Cholesterol=70, Calcium=35
predict(Calories_lm, NewData)
```

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
```

```
#The predicted value for Calories is 1828.312.
```

```
NewData2 = data.frame(Protein=0.1, TotalFat=35, Carbohydrate=405, Sodium=44440, Cholesterol=70, Calcium=
predict(Calories_lm, NewData2)
```

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
## 1843.198
```

```
(1843.198-1828.312)/1828.12 *100
```

```
## [1] 0.8142792
```

```
#The Calories will increase by 0.8142792%. Since Sodium is the least significant independent variable,
#the Calories has only increased by a small amount even though the Sodium amount has increased by 10000
```

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

```
#H_0: The population means of the scores from the two groups are equal, mu_1 = mu_2.
#The null hypothesis states that illustrations do not affect how students retell the article.
#H_a: The population means of the scores from the two groups are not equal, mu_1 != mu_2.
#The alternative hypothesis states that illustrations affect how students retell the article.
```

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

```
#This is a paired experiment because both article 1 and 2 are tested for each student.
```

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

```
#Since comparing two groups with nonparametric statistics and paired experiment,
#Wilcoxon signed rank test should be used.
```

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)
wilcox.test(article1, article2, paired=T)
```

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
```

```
#Since p-value is greater than 0.05, we do not reject the null hypothesis;
#therefore, illustrations do not improve how the students retell an article.
#We accept the null hypothesis that the population means of the scores from the two groups are equal
#because given the confidence level is 0.05, the p-value, 0.8125, is much greater.
```

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 an unpaired experiment because the toothpastes A and B
#are not bought from the same store.
```

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

```
#Since comparing two groups with nonparametric statistics and unpaired experiment,
#Wilcoxon rank sum test should be used.
```

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)
wilcox.test(CompanyA, CompanyB, paired = FALSE)
```

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
```

```
#Since the p-value is less than 0.05, we reject the Null Hypothesis.
#The packaging process between the two companies are different based on weight measurement
#because we reject the Null Hypothesis that the two population means are the same.
```

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