CMTH 642 Data Analytics: Advanced Methods

Assignment 2 (10%)

Ajay Herod

D10 500764104

#### 1. Read the csv file (USDA\_Clean.csv) in the folder and assign it to a data frame. (3 points)

USDA\_Clean <- read.csv(file = 'USDA\_Clean.csv')

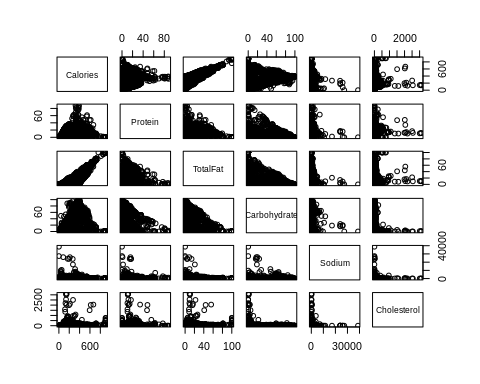
#### 2. Check the datatypes of the attributes. (3 points)

head(USDA\_Clean)

## X ID Description Calories Protein TotalFat Carbohydrate Sodium  
## 1 1 1001 BUTTER,WITH SALT 717 0.85 81.11 0.06 714  
## 2 2 1002 BUTTER,WHIPPED,WITH SALT 717 0.85 81.11 0.06 827  
## 3 3 1003 BUTTER OIL,ANHYDROUS 876 0.28 99.48 0.00 2  
## 4 4 1004 CHEESE,BLUE 353 21.40 28.74 2.34 1395  
## 5 5 1005 CHEESE,BRICK 371 23.24 29.68 2.79 560  
## 6 6 1006 CHEESE,BRIE 334 20.75 27.68 0.45 629  
## Cholesterol Sugar Calcium Iron Potassium VitaminC VitaminE VitaminD  
## 1 215 0.06 24 0.02 24 0 2.32 1.5  
## 2 219 0.06 24 0.16 26 0 2.32 1.5  
## 3 256 0.00 4 0.00 5 0 2.80 1.8  
## 4 75 0.50 528 0.31 256 0 0.25 0.5  
## 5 94 0.51 674 0.43 136 0 0.26 0.5  
## 6 100 0.45 184 0.50 152 0 0.24 0.5  
## HighSodium HighCals HighSugar HighProtein HighFat  
## 1 1 1 0 0 1  
## 2 1 1 0 0 1  
## 3 0 1 0 0 1  
## 4 1 1 0 1 1  
## 5 1 1 0 1 1  
## 6 1 1 0 1 1

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

plot(USDA\_Clean[4:9])



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

# Yes, there is a correlation between Calories and Total Fat. The correlation appears to be statistically significant and shows a strong positive linear trend. This means that as either attribute increases so does the other.

#### 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)

summary(lm (Calories ~ Protein + TotalFat + Carbohydrate + Sodium + Cholesterol, data = USDA\_Clean))

##   
## Call:  
## lm(formula = Calories ~ Protein + TotalFat + Carbohydrate + Sodium +   
## Cholesterol, data = USDA\_Clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -191.087 -3.832 0.426 5.147 291.011   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.9882753 0.4832629 8.253 < 2e-16 \*\*\*  
## Protein 3.9891994 0.0233550 170.807 < 2e-16 \*\*\*  
## TotalFat 8.7716980 0.0143291 612.158 < 2e-16 \*\*\*  
## Carbohydrate 3.7432001 0.0091404 409.522 < 2e-16 \*\*\*  
## Sodium 0.0003383 0.0002189 1.545 0.122   
## Cholesterol 0.0110138 0.0019861 5.545 3.05e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.92 on 6304 degrees of freedom  
## Multiple R-squared: 0.9877, Adjusted R-squared: 0.9877   
## F-statistic: 1.009e+05 on 5 and 6304 DF, p-value: < 2.2e-16

#### 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)

formula <- lm(Calories ~ Protein + TotalFat + Carbohydrate + Sodium + Cholesterol, data = USDA\_Clean)  
formula

##   
## 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 = 3.99Protein + 8.77TotalFat + 3.74Carbohydrate + 0.0003Sodium + 0.011Cholestrol + 3.99

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

#Sodium is the least significant independent variable because the coefficient in the linear regression equation is the lowest. This means it is the least significant variable for Calories.

#### 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)

newProduct <- data.frame("Protein" = 0.1, "TotalFat" = 35, "Carbohydrate" = 405, "Sodium" = 440, "Cholesterol" = 70)  
predict(formula, newProduct)

## 1   
## 1828.312

#### 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)

newProduct2 <- data.frame("Protein" = 0.1, "TotalFat" = 35, "Carbohydrate" = 405, "Sodium" = 44440, "Cholesterol" = 70)  
((predict(formula, newProduct2)/predict(formula, newProduct))-1)\*100

## 1   
## 0.8141547

#There is only a 0.81% increase in calories for a 10000% increase in sodium. This is because sodium has a very low coefficient making it insignificant as a dependent variable to calories even with a 10000% increase.

#### 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

#### A) What are and ? (5 points)

#Illustrations do not improve how students retell an article  
#H0: p >= 0.05  
#illustrations improve how students retell an article   
#Ha: p < 0.05

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

#This is a unpaired 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)

#Wilcoxon rank sum 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)

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)

##   
## Wilcoxon rank sum test  
##   
## data: Article1 and Article2  
## W = 10, p-value = 0.6905  
## alternative hypothesis: true location shift is not equal to 0

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

#The p-value is 0.69 indicating to accept the null hypothesis. A p-value higher than 0.05 is not statistically significant and indicates strong evidence for the null hypothesis. This means that the illustrations do not improve how the students retell an article.

#### 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

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

#This is a unpaired experiment.

#### 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)

#Wilcoxon rank sum test

#### 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)

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)

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

#### D) Will you accept or reject your Null Hypothesis? () Are packaging process similar or different based on weight measurements? Why? (5 points)

#The p-value is 0.049 indicating to reject the null hypothesis. A p-value less than 0.05 is statistically significant and indicates strong evidence against the null hypothesis. This means the packaging process is different based on weight measurements.

This is the end of Assignment 2

Ceni Babaoglu, PhD