CMTH 642 Data Analytics: Advanced Methods

Assignment 2 (10%)

[Daria Yip]

[DHA - Student#500721106]

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

USDAclean <- read.csv("C:/Users/samda/Documents/Ryerson/CIND 642 - R/USDA\_Clean.csv", header = T, sep = ",")

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

sapply(USDAclean, class)

## 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"

#Can also use str(USDAclean)

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

subset\_USDAclean <- data.frame(USDAclean[, c('Calories', 'Carbohydrate', 'Protein', 'TotalFat', 'Sodium', 'Cholesterol')])  
cor(subset\_USDAclean) #check correlation values

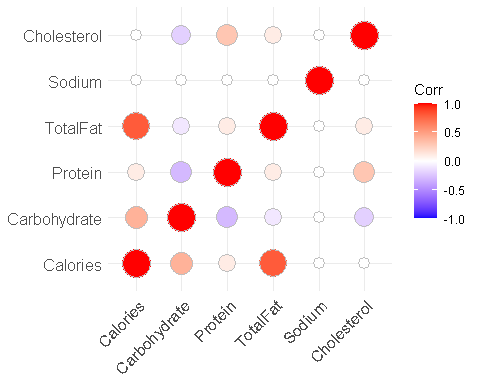
## Calories Carbohydrate Protein TotalFat Sodium  
## Calories 1.00000000 0.42460618 0.122122537 0.804495022 0.032321026  
## Carbohydrate 0.42460618 1.00000000 -0.304711167 -0.124342914 0.046838692  
## Protein 0.12212254 -0.30471117 1.000000000 0.057035611 -0.003489485  
## TotalFat 0.80449502 -0.12434291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

#Correlogram with Circles  
#install.packages("ggcorrplot")  
library(ggcorrplot)

## Warning: package 'ggcorrplot' was built under R version 4.0.2

## Loading required package: ggplot2

corr <- round(cor(subset\_USDAclean), 1)  
ggcorrplot(corr, method = "circle")



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

cor.test(USDAclean$Calories,USDAclean$TotalFat, method = "pearson")

##   
## Pearson's product-moment correlation  
##   
## data: USDAclean$Calories and USDAclean$TotalFat  
## t = 107.58, df = 6308, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.7956139 0.8130305  
## sample estimates:  
## cor   
## 0.804495

#The correlation between Calories and Total Fat is #statistically significant because the p-value is < 2.2e-16 as  
#shown below, which is much below the significance threshold of  
#0.05. Therefore, as the p value is below 0.05, the correlation  
#is statistically significant.

#### 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)

model <- lm(Calories ~ Protein + TotalFat + Carbohydrate + Sodium + Cholesterol , data = USDAclean)

#### 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)

#Based on the formula y = a + bx, the equation could be   
#rewritten as Calories = 3.9882752613 + 3.9891994394\*Protein + #8.7716980068\*TotalFat + 3.7432000604\*Carbohydrate + #0.0003383021\*Sodium + #0.0110138110\*Cholesterol  
  
coef(model) #find the coefficients and intercept for the model

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

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

#Using ANOVA, you can tell which independent variable is least  
#significant.  
anovamodel <- anova(model)  
anovamodel

## 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

#This table shows us the P value for each variable. While each  
#of the other variabes have a P-value less than 2.2e-16 or a #p-value of 3.05e-08, we have sodium with a p-value of 0.1378.  
#This is not even under 0.05, which signifies an insignificant  
#independent variable and the least significant in this model.

#### 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)

stats <- data.frame(Protein = 0.1, TotalFat = 35, Carbohydrate = 405, Sodium = 440, Cholesterol = 70, Calcium = 35, Potassium = 35, VitaminC = 10)  
original <- predict(model, stats)  
original

## 1   
## 1828.312

#Therefore, the predicted value for Calories is 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)

stats2 <- data.frame(Protein = 0.1, TotalFat = 35, Carbohydrate = 405, Sodium = 44440, Cholesterol = 70, Calcium = 35, Potassium = 35, VitaminC = 10)  
((predict(model, stats2) - original)/original)\*100

## 1   
## 0.8141547

#Therefore, there will be a change in calories of 0.81% as an  
#increase. The "slope" or coefficient of Sodium tells you how  
#much Calories will change for each extra unit of Sodium. This #can also be manually calculated using the linear equation #setup earlier with the coefficient of Sodium,  
#0.0003383021, and then multiply it by the difference in 44440  
#and 440 to find the change in its contribution to the #Calories.

#### 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)

#H0: The median of the scores from both groups are the #same/equal.  
#HA: The median of the scores from both groups are not equal/are #different.

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

#This is an paired experiment because it studies subjects  
#at two different times and tests the students' recollection  
#for Story 1 and Story 2.

#### 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)

#Based on it being an paired test and a non-parametric test  
#statistic is required, we should use the Wilcoxon Signed Rank  
#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, paired = TRUE)

##   
## 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

#based off the result, we can see that the p-value is 0.8125,  
#which is statistically insignificant is it is above the   
#threshold of 0.05.

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

#Based off the p-value of the last question, we will accept   
#our Null Hypothesis. Therefore, illustrations do not improve  
#how the students retell an article as that is the Null  
#Hypothesis we have accepted.

#### 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 an unpaired experiment due to us buying random  
#toothpastes and comparing two different subjects.

#### 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)

#Based on it being an unpaired experiment, I would use the  
#non-parametric test of Wilcoxon Rank Sum Test or Mann-Whitney  
#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, paired = FALSE)

##   
## 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

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

#Based on the alpha of 0.05 and our p-value of 0.04988, we can  
#reject our Null Hypothesis. Therefore, all packaging processes  
#are different based on weight measurements as we can deduce  
#from our p-value.

This is the end of Assignment 2

Ceni Babaoglu, PhD