Stats 110 Homework 1

Owen Lin

1.
   a. There is a positive association between height of athlete and length of jump
   b. It looks like there is a linear relationship and a line would fit well though the data
   c. 0.7 could be a possible value of the correlation coefficient between distance and height, it demonstrate a positive, relative strong (but not perfect) relationship
   d. Both response and explanatory variable are quantitative.
   e. The predicted length of the jump for an 72 inches tall athlete is 82.27
   f. We would expected an increase of 1.0534 inches jump length on average when the height of athlete is increase by 1.
   g. No, the intercept gave an estimate of jump lenth when the height of athlete is 0 inches - not useful at all.
   h. No, it is only an observational study, thus the coach can not conclude that there exists a causal relationship
   i. If response variable Y is recorded in feet instead, the intercept estimate will decrease to 0.5357. Because it means Y = 6.4285 inches when X = 0 => Y = 6.4285 inches = 0.5357 feet when X = 0
   j. If explanatory variable X is recorded in feet instead, the intercept estimate will stay the same. Because it means Y = 6.4285 inches when X = 0 inch => Y = 6.4285 feet when X = 0 feet = 0 inch
   k. The correlation coefficient will not change regardless of the units. That the whole reason why it was invented… Covariance is what will be affected by the unit, but in correlation coefficient, we take units in to account by dividing covariance with the standard deviation of X and Y.

2.
   a. A randomized experiement can’t really be conducted, because it is not plausible to tell the participant that he/she can only spend certain number of hours outdoor. Even if it is achieved by some mean, it will affect many other things such as the emotion of participants, which can be a confounding variable.
   b. Researchers can’t conclude a causal relationship from an observational study because of the confounding variables.
   c. Those who spend more time outdoor might spend more time on jogging. And as jogging time increases, the number of miles increases. It could well be that the number of miles someone jogs each day is the true explanatory variable that cause the blood pressure to decrease, but we only observed the number of outdoor hours.

3.
   - AcceptStatus: A=accepted to med school and D=denied **categorical**
   - Acceptance: 1=accepted and 0=denied **categorical**
   - Sex: F=female and M=male **categorical**
   - BCPM: Bio/Chem/Physics/Math grade point average **quantitative**
   - GPA: College grade point average **quantitative**
   - VR: Verbal reasoning subscore on the MCAT **quantitative**
   - PS: Physical science subscore **quantitative**
   - WS: Writing sample subscore **quantitative**
   - BS: Biological science subscore **quantitative**
   - MCAT: Score on the MCAT (sum of VR, PS, WS, and BS) **quantitative**
   - Apps: Number of medical schools applied to **quantitative**

4. a. response variable: **Verbal reasoning subscore**, explanatory variable: **Sex**
   b. response variable: **Acceptance**, explanatory variable: **Sex**
   c. response variable: **MCAT score**, explanatory variable: **GPA**
   d. response variable: **Acceptance**, explanatory variable: **Bio/Chem/Physics/Math GPA**

5. ○ The mean and the five number summary for GPA for those who were admitted

```
mcat <- read.table("D:\\Coding\\BigData\\Stats110\\MedGPA.txt", fill = TRUE, header = TRUE) # no lint
summary(mcat[mcat$Acceptance == 1, ]$GPA)
mean(mcat[mcat$Acceptance == 1, ]$GPA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   3.140   3.545   3.715   3.693   3.888   3.970
## [1] 3.693333
```

- The mean and the five number summary for GPA for those who were denied admission to med school

```
summary(mcat[mcat$Acceptance == 0, ]$GPA)
mean(mcat[mcat$Acceptance == 0, ]$GPA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.720   3.290   3.380   3.385   3.610   3.770
## [1] 3.3852
```

b. The average GPA for those who were admitted is higher then those who were denied admission.

6. a.

$$y = 3.923 + 9.104x$$

where Y is MCAT score and X is GPA

```
model <- lm(MCAT ~ GPA, data = mcat)
model
```

```
## 
## Call:
## lm(formula = MCAT ~ GPA, data = mcat)
## 
## Coefficients:
## (Intercept)          GPA  
##       3.923        9.104
```

b.

$$\sum_{i=1}^{55}[Y_i - (3.923 + 9.104X_i)]^2$$

c. We would expected an increase of 9.104 MCAT score on average when the GPA is increase by 1

d. We expected an average of 3.923 MCAT score when the GPA of the student is 0. It is not a useful interpretation because student with a GPA of 0 is unlikely to graduate from the college.

e. Predicting price:

```
x <- 3.923 + 9.104 * 3
print(paste("Predicted MCAT score for some one that has a GPA of 3.0 is", x)) # nolint
y <- 3.923 + 9.104 * 4
print(paste("Predicted MCAT score for some one that has a GPA of 4.0 is", y)) # nolint
```

```
## [1] "Predicted MCAT score for some one that has a GPA of 3.0 is 31.235"
## [1] "Predicted MCAT score for some one that has a GPA of 4.0 is 40.339"
```

f. The predicted difference in MCAT scores for two people who differ in GPA by 2.0 is 9.104*2 = 18.208

g. No, we can not make a causal conclusion because it is an observational study instead of a randomized experiement (which is infeasible).

7. In an observation study, many confounding variables are involved. But in a randomized experiement, potential effect from confounding variables are mitigated. The idea is that we randomly assign the sample to different assigned value of explanatory variable. Since the assignment is random, we can reasonably say that at each level of explanatory variable, there are about the same degree of confounding factors.