STAT 645 Assignment 1

Kat Gliszczynski

2022-08-26

Question 1- Edited for mof problem 3.4 of the book

Pearson correlation is zero because because it is a nonlinear function since it is a quadratic equation. Spearman's correlation uses the ranks instead of the values of the variables to calculate correlation however this still would result in a correlation close to 0 since the ranks will be based off of X^2 and -5 and 5 will still have the same rank.

```
x <- runif(1000,-5,5)
z = x^2
y = z = rnorm(1000,0,1)
corr <- cor(x,y)
cat("Sample Correlation of x and y =",corr)
```

```
## Sample Correlation of x and y = -0.04239677
```

In this example the correlation will always be close to 0 regardless of the values of x and y since we are using a quadratic function.

Question 2

```
a <- 3/5
b <- 4/26
ER <- a-b
RR <- a/b
OR <- (a/(1-a))/(b/(1-b))
cat("Excess Risk =", ER)
```

```
## Excess Risk = 0.4461538
```

```
cat("Relative Risk =",RR)
```

```
## Relative Risk = 3.9
```

```
cat("Odds Ratio =",OR)
```

```
## Odds Ratio = 8.25
```

Question 3- Problem 3.6

```
p1 <- 191/711
p0 <- 9/264
OR1 <- (p1/(1-p1))/(p0/(1-p0))
cat("Odds ratio comparing the risk of cancer in individuals who report consuming more than ten grams of
```

```
## Odds ratio comparing the risk of cancer in individuals who report consuming more than ten grams of to
```

```
p1 <- 191/200
p0 <- 520/775
OR2 <- (p1/(1-p1))/(p0/(1-p0))
cat("Odds ratio comparing the proportion of individuals reporting higher levels of consumption among cas
```

```
## Odds ratio comparing the proportion of individuals reporting higher levels of consumption among cases
```

Individuals who consume more than ten grams of tobacco per day are 10.407 times higher risk of cancer than those in the group with less consumption.

Individuals among cases are 10.407 times more likely to report higher levels of consumption compared to those among the controls.

Question 4

```
library(readr)
hersdata_LDL_1_ <- read.csv("hersdata_LDL(1).csv")
View(hersdata_LDL_1_)
```

First I created the subset for data of just white women and printed out some summary statistics before running a multi way ANOVA for LDL and exercise, diabetes and smoking. I am testing the Hypotheses: 1. Ho: There is no effect of exercise on LDL Ha: There is some effect of exercise on LDL 2. Ho: There is no effect of diabetes on LDL Ha: There is some effect of diabetes on LDL 3. Ho: There is no effect of smoking on LDL Ha: There is some effect of smoking on LDL

```
WhiteWomen <-subset(hersdata_LDL_1_, nonwhite == "no",select=c(nonwhite,LDL,exercise,diabetes,smoking))
dim(WhiteWomen)
```

```
## [1] 2451    5
```

```
names(WhiteWomen)
```

```
## [1] "nonwhite" "LDL"      "exercise" "diabetes" "smoking"
```

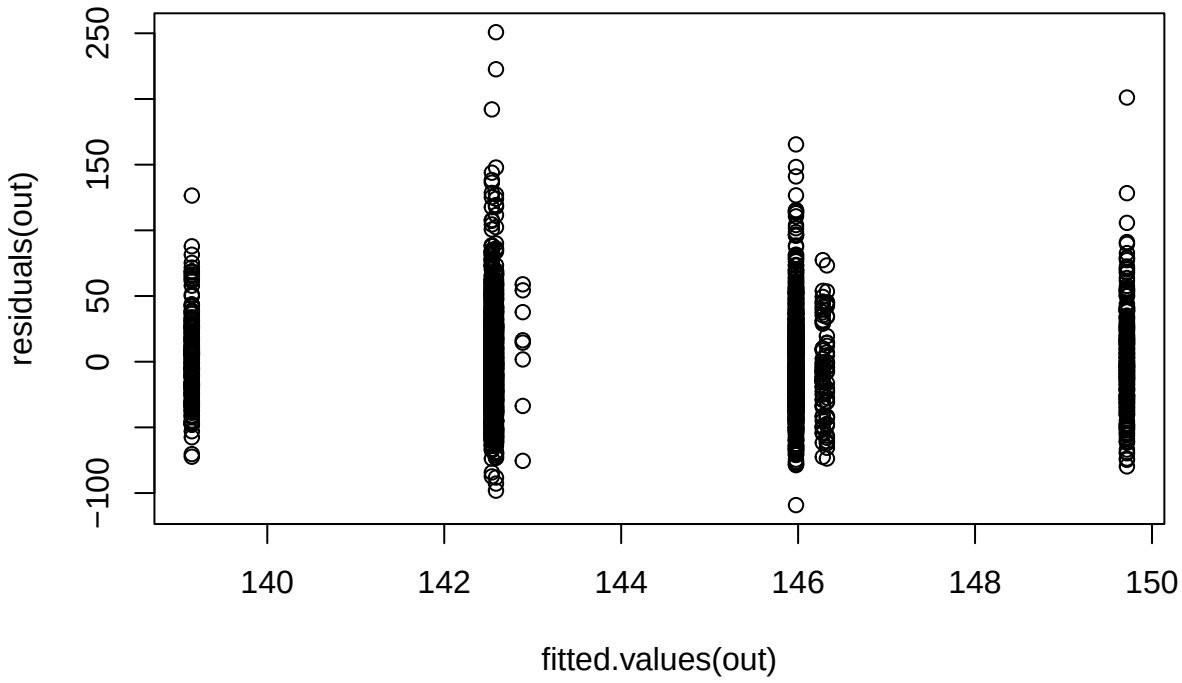
```
summary(WhiteWomen)
```

```
##    nonwhite              LDL          exercise           diabetes
##  Length:2451        Min.   : 36.8   Length:2451        Length:2451
##  Class :character   1st Qu.:119.4   Class :character   Class :character
##  Mode  :character   Median :140.8   Mode  :character   Mode  :character
##                     Mean   :144.3
##                     3rd Qu.:165.2
##                     Max.   :393.4
##                     NA's   :9
##    smoking
##  Length:2451
##  Class :character
##  Mode  :character
##
##
##
##
```

```
out = aov(LDL~exercise+diabetes+smoking, data=WhiteWomen)
```

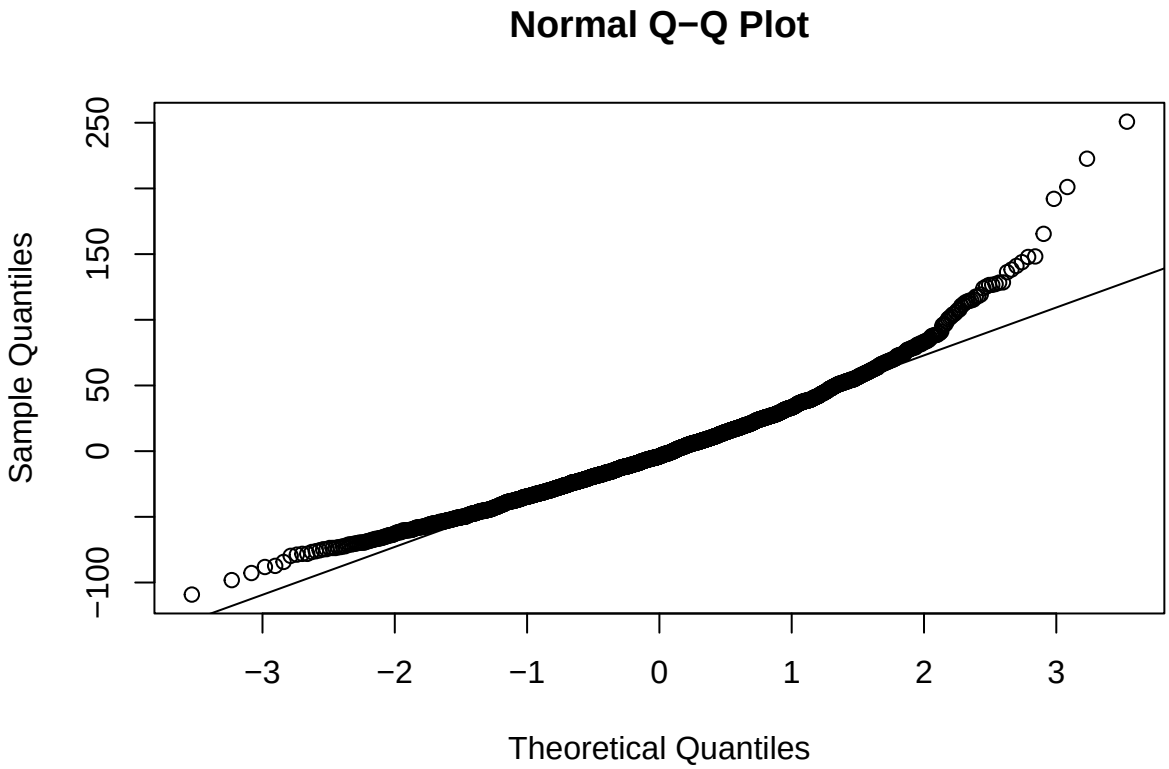
After running the multiway anova I now checked if the assumption of homogeneous variance is satisfied was met.

```
plot(fitted.values(out),residuals(out))
```



Now I checked normality of of the residuals with a Q-Q plot.

```
out2=qqnorm(residuals(out))
out3= lm(out2$y~out2$x)
abline(a=out3$coef[1],b=out3$coef[2])
```



After checking the homogeneity and normality of the residuals I was able to form a conclusion. In conclusion I rejected the null hypothesis at the 0.05 level for exercise and smoking since the p-value for those variables was lower than 0.05 so there is evidence of some effect of exercise on LDL in white women and some effect of smoking on LDL in white women.I failed to reject the null hypothesis at the 0.05 level so there is no evidence of an effect of diabetes on LDL in white women.

Question 5

I will be testing the following hypothesis during this question: Ho: mean difference between white and nonwhite SBP = 0 when controlling the effect of other variables, exercise, smoking and age. vs. Ho: mean difference between white and nonwhite SBP != 0 when controlling the effect of other variables, exercise, smoking and age.

First I need to group the data into groups by smoking, by exercise and by age: [44-55), [55,65), 65 and above.

```
U55NoExcNoSmoke <- subset(hersdata_LDL_1_, smoking == "no" & age <55 & exercise == "no")
U55NoExcYesSmoke <- subset(hersdata_LDL_1_, smoking == "yes" & age <55 & exercise == "no")
U55YesExcNoSmoke <- subset(hersdata_LDL_1_, smoking == "no" & age <55 & exercise == "yes")
U55YesExcYesSmoke <-subset(hersdata_LDL_1_, smoking == "yes" & age <55 & exercise == "yes")
```

```
NoExcNoSmoke55to65 <- subset(hersdata_LDL_1_, smoking == "no" &age >= 55& age < 65 & exercise == "no")
NoExcYesSmoke55to65 <- subset(hersdata_LDL_1_, smoking == "yes" &age >= 55& age < 65 & exercise == "no")
YesExcNoSmoke55to65 <- subset(hersdata_LDL_1_, smoking == "no" & age >= 55& age < 65 & exercise == "yes")
YesExcYesSmoke55to65 <-subset(hersdata_LDL_1_, smoking == "yes" & age >= 55& age < 65 & exercise == "yes")
O65NoExcNoSmoke <- subset(hersdata_LDL_1_, smoking == "no" & age >= 65 & exercise == "no")
O65NoExcYesSmoke <- subset(hersdata_LDL_1_, smoking == "yes" & age >= 65 & exercise == "no")
O65YesExcNoSmoke <- subset(hersdata_LDL_1_, smoking == "no" & age >= 65 & exercise == "yes")
O65YesExcYesSmoke <-subset(hersdata_LDL_1_, smoking == "yes" & age >= 65 & exercise == "yes")

group1a <-subset(U55NoExcNoSmoke, nonwhite == "yes")
group1b <-subset(U55NoExcNoSmoke, nonwhite == "no")
group2a <-subset(U55NoExcYesSmoke, nonwhite == "yes")
group2b <-subset(U55NoExcYesSmoke, nonwhite == "no")
group3a <-subset(U55YesExcNoSmoke, nonwhite == "yes")
group3b <-subset(U55YesExcNoSmoke, nonwhite == "no")
group4a <-subset(U55YesExcYesSmoke, nonwhite == "yes")
group4b <-subset(U55YesExcYesSmoke, nonwhite == "no")

group5a <-subset(NoExcNoSmoke55to65, nonwhite == "yes")
group5b <-subset(NoExcNoSmoke55to65, nonwhite == "no")
group6a <-subset(NoExcYesSmoke55to65, nonwhite == "yes")
group6b <-subset(NoExcYesSmoke55to65, nonwhite == "no")
group7a <-subset(YesExcNoSmoke55to65, nonwhite == "yes")
group7b <-subset(YesExcNoSmoke55to65, nonwhite == "no")
group8a <-subset(YesExcYesSmoke55to65, nonwhite == "yes")
group8b <-subset(YesExcYesSmoke55to65, nonwhite == "no")

group9a <-subset(O65NoExcNoSmoke, nonwhite == "yes")
group9b <-subset(O65NoExcNoSmoke, nonwhite == "no")
group10a <-subset(O65NoExcYesSmoke, nonwhite == "yes")
group10b <-subset(O65NoExcYesSmoke, nonwhite == "no")
group11a <-subset(O65YesExcNoSmoke, nonwhite == "yes")
group11b <-subset(O65YesExcNoSmoke, nonwhite == "no")
group12a <-subset(O65YesExcYesSmoke, nonwhite == "yes")
group12b <-subset(O65YesExcYesSmoke, nonwhite == "no")
```

Now I will randomly choose 2 white and 2 nonwhite from each group

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
newgroup1a <- sample_n(group1a,2,replace=FALSE)
newgroup1b <- sample_n(group1b,2,replace=FALSE)
newgroup2a <- sample_n(group2a,2,replace=FALSE)
newgroup2b <- sample_n(group2b,2,replace=FALSE)
newgroup3a <- sample_n(group3a,2,replace=FALSE)
newgroup3b <- sample_n(group3b,2,replace=FALSE)
newgroup4a <- group4a
newgroup4b <- sample_n(group4b,2,replace=FALSE)
newgroup5a <- sample_n(group5a,2,replace=FALSE)
newgroup5b <- sample_n(group5b,2,replace=FALSE)
newgroup6a <- sample_n(group6a,2,replace=FALSE)
newgroup6b <- sample_n(group6b,2,replace=FALSE)
newgroup7a <- sample_n(group7a,2,replace=FALSE)
newgroup7b <- sample_n(group7b,2,replace=FALSE)
newgroup8a <- sample_n(group8a,2,replace=FALSE)
newgroup8b <- sample_n(group8b,2,replace=FALSE)
newgroup9a <- sample_n(group9a,2,replace=FALSE)
newgroup9b <- sample_n(group9b,2,replace=FALSE)
newgroup10a <- sample_n(group10a,2,replace=FALSE)
newgroup10b <- sample_n(group10b,2,replace=FALSE)
newgroup11a <- sample_n(group11a,2,replace=FALSE)
newgroup11b <- sample_n(group11b,2,replace=FALSE)
newgroup12a <- sample_n(group12a,2,replace=FALSE)
newgroup12b <- sample_n(group12b,2,replace=FALSE)
```

Now average the SBP for white and nonwhite groups in each group and then create the groups

```
group1a <- mean(newgroup1a$SBP)
group1b <- mean(newgroup1b$SBP)
group2a <- mean(newgroup2a$SBP)
group2b <- mean(newgroup2b$SBP)
group3a <- mean(newgroup3a$SBP)
group3b <- mean(newgroup3b$SBP)
group4a <- mean(newgroup4a$SBP)
group4b <- mean(newgroup4b$SBP)
group5a <- mean(newgroup5a$SBP)
group5b <- mean(newgroup5b$SBP)
group6a <- mean(newgroup6a$SBP)
group6b <- mean(newgroup6b$SBP)
group7a <- mean(newgroup7a$SBP)
group7b <- mean(newgroup7b$SBP)
group8a <- mean(newgroup8a$SBP)
group8b <- mean(newgroup8b$SBP)
group9a <- mean(newgroup9a$SBP)
group9b <- mean(newgroup9b$SBP)
group10a <- mean(newgroup10a$SBP)
group10b <- mean(newgroup10b$SBP)
group11a <- mean(newgroup11a$SBP)
group11b <- mean(newgroup11b$SBP)
group12a <- mean(newgroup12a$SBP)
group12b <- mean(newgroup12b$SBP)

a <- c(group1a,group2a,group3a,group4a,group5a,group6a,group7a,group8a,group9a,group10a,group11a,group1
```

```
a
```

```
##  [1] 112.0 119.0 128.5 127.0 129.5 135.5 133.5 132.0 146.5 158.0 135.0 165.5
```

```
b <- c(group1b,group2b,group3b,group4b,group5b,group6b,group7b,group8b,group9b,group10b,group11b,group1
b
```

```
##  [1] 116.0 141.0 129.5 112.0 108.0 121.0 143.0 142.0 154.5 128.5 137.5 153.0
```

Run the Paired t-test

```
t.test(a,b,paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  a and b
## t = 0.6789, df = 11, p-value = 0.5112
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.725893 12.725893
## sample estimates:
## mean of the differences
##                       3
```

The p-value is greater than 0.05 so we therefore fail to reject the null hypothesis. There is not significant evidence that the mean difference of the expression is different from zero when controlling the effect of other variables, exercise, smoking and age.