Stats 402 ~ Homework 2

Group 4 ~ Britney Brown, Harrison DiStefano, Jaehui(Jaehee) Jeong, Lisa Kaunitz, Tianyang Liu, Yuandong (David) Sun, Jeremy Weidner

11/02/2020

Problem One

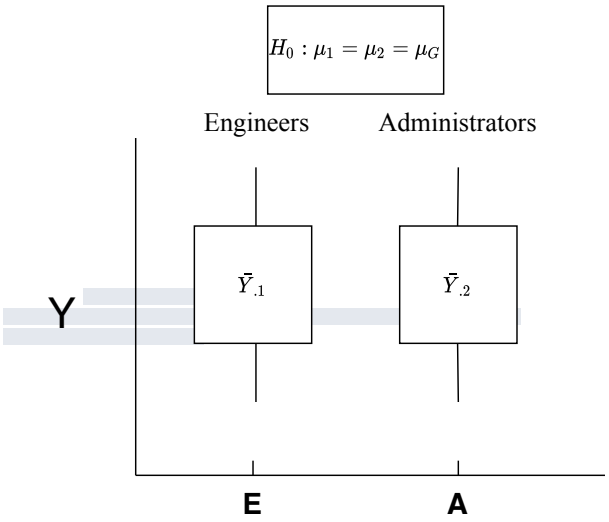
Question: If you were looking for a TA would you hire this person, Yes or no, and why? Back up your reasoning with drawing the right schematic, showing the relevant formulas, connecting your conceptual explanation to the formulas you used.

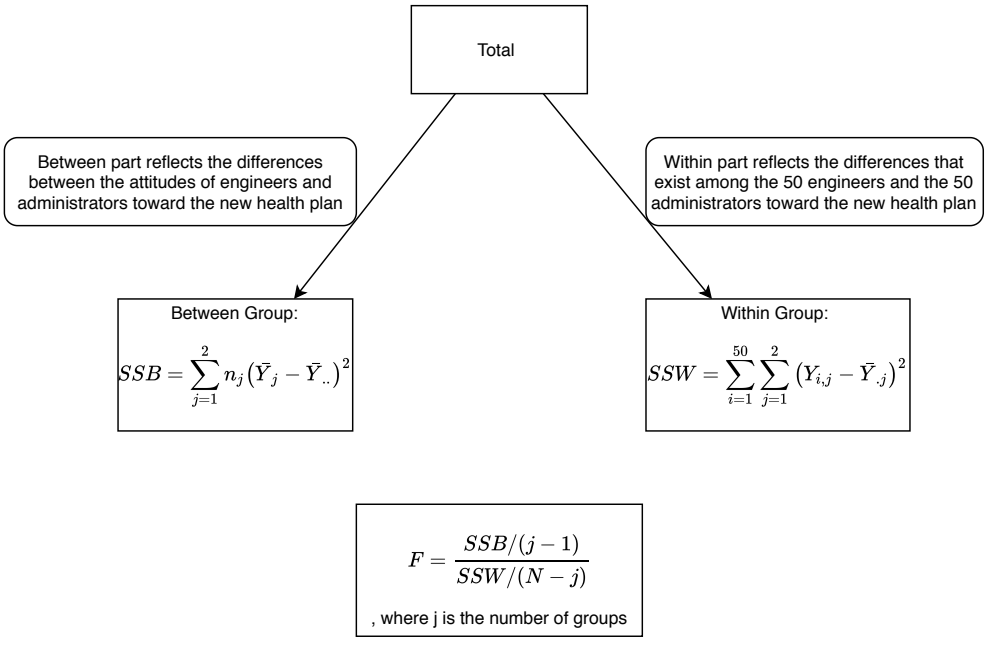
We would not hire this person. Their conceptual explanation to clarify the difference between the two groups (within and between) is incorrect and should be the other way around (as explained below). In addition, the mathematical approach to finding an F-value is incorrect and unclear. When referring to groups, the TA is not precise in the difference between engineers and administrators versus the difference in the between and within groups. A better conceptualization of mean square between and mean square within is as follows:

From a conceptual point of view, the within part (SSW) reflects the difference within the individuals of each group, comparing them to their group mean. In this example, we would use each individual attitude response, subtract the mean attitude response of its group, and then take the sum of squares for these results. We do this for both the engineer and administrator groups. Lastly, we take the sum of these two groups results. This is the part of the variance that we **cannot** explain.

The between part (SSB) reflects the differences between the attitudes of engineers and administrators toward the new health plan. In this example, we use the mean of each group to subtract the grand mean of all groups and square the result, then multiply by the number of samples of each group (n=50), and lastly take the sum of each group's result. This is the part of the variance that we **can** explain.

See the following page for a visual representation.





$$F = \frac{SSB/(j-1)}{SSW/(N-j)}$$

, where j is the number of groups

Problem Two

Using Houston real estate data posted in the homework folder on week three, answer the following questions.

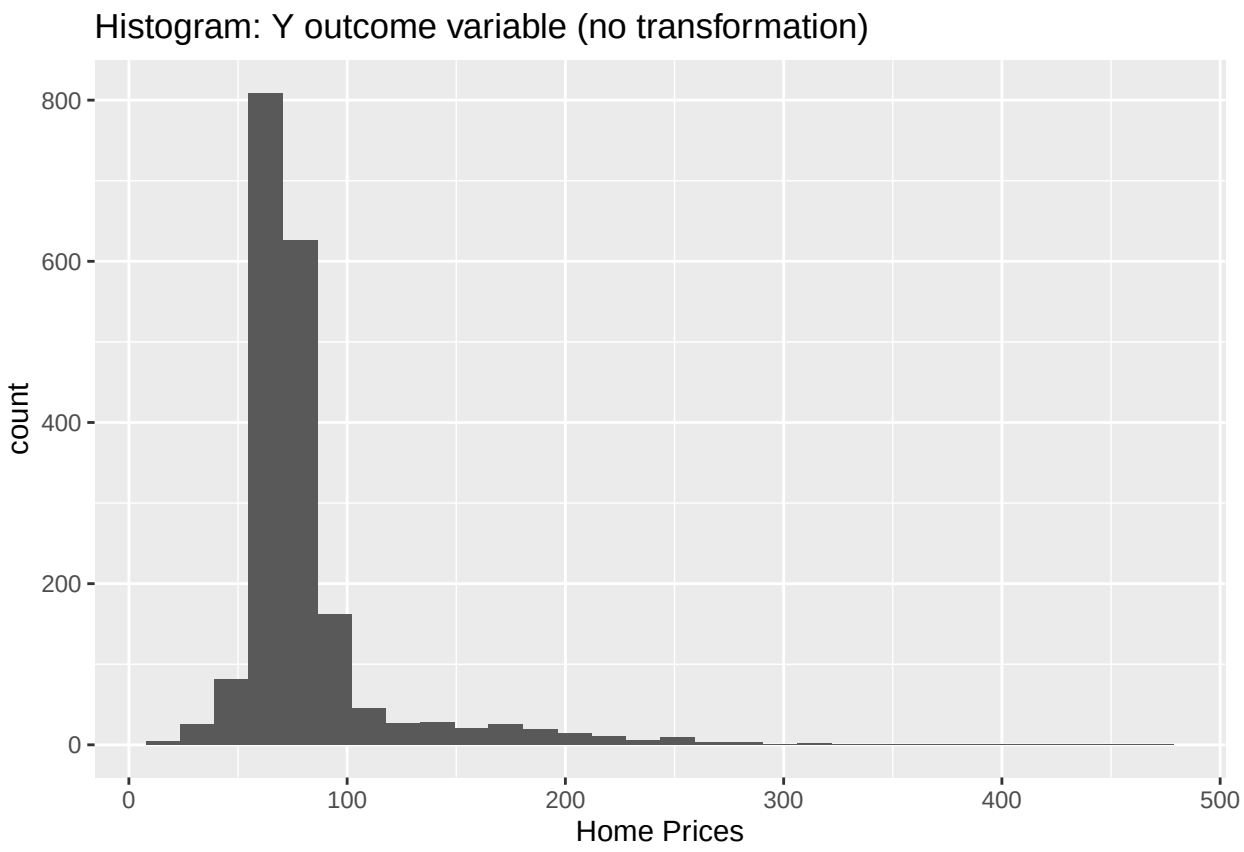
```
houstonrealesate <- read.csv("houstonrealesate.csv")
head(houstonrealesate) #view data
```

```
##       Yi ni   x1i   x2i
## 1 169.20  7 0.857 0.000
## 2  56.82  6 0.167 0.667
## 3  25.52  6 0.000 1.000
## 4  90.67  5 0.800 0.200
## 5  92.65  8 0.500 0.000
## 6  87.76  9 0.667 0.000
```

a) Draw the histogram of home prices (Y)

```
library(ggplot2)
ggplot(mapping = aes(x=houstonrealesate$Yi)) + geom_histogram() +
  xlab("Home Prices") + ggtitle("Histogram: Y outcome variable (no transformation)")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From the histogram above, we see that the outcome variable, home prices, is right skewed.

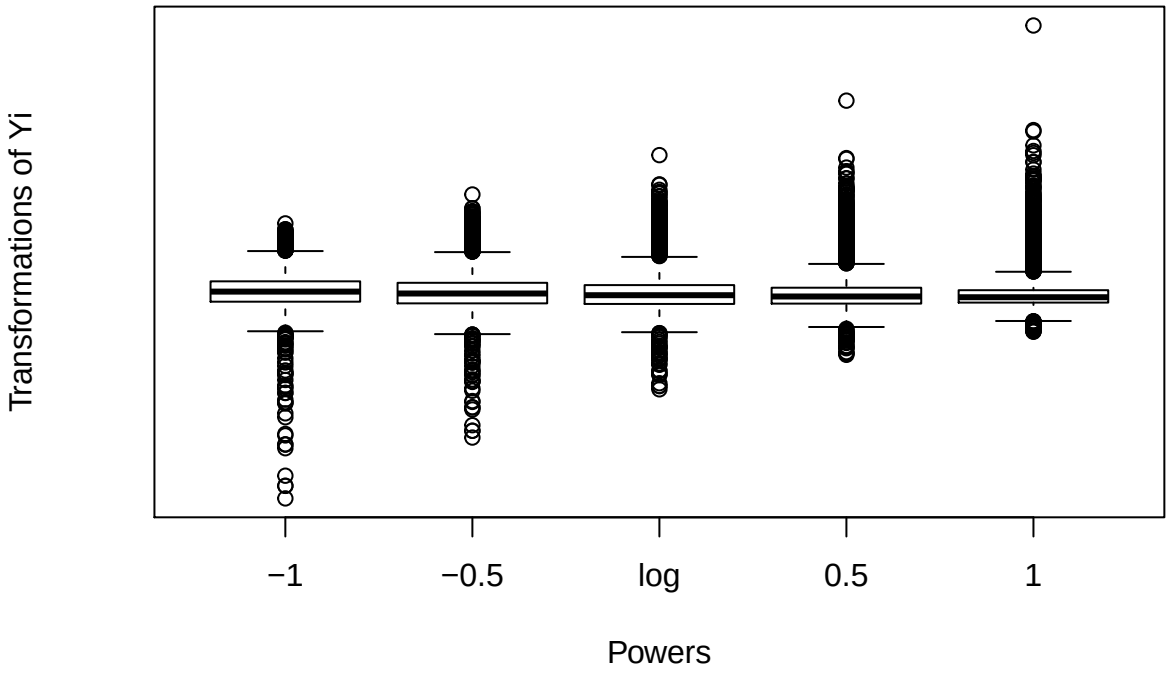
b) Conduct transformation using the package in library "car" – see lecture two – and decide which transformation if any help to solve the problem.

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.6.2
```

```
## Loading required package: carData
```

```
symbox(~Yi,data=houstonrealesate)
```



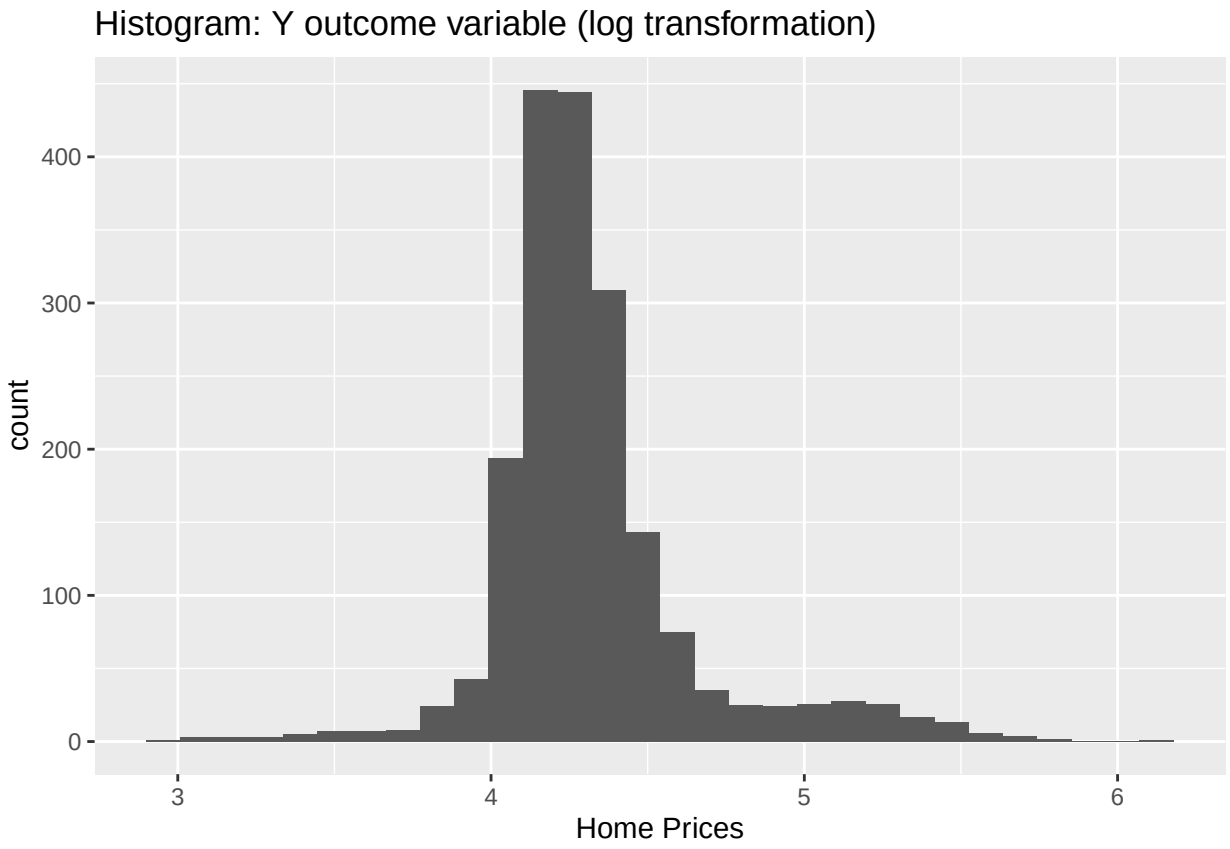
The above plot suggests that the log transformation of the Yi variable does the best job at creating a symmetric distribution.

c) Make the transformation that you think is best and draw the resulting histogram.

```
logYi <- log(houstonrealesate$Yi)

ggplot(mapping = aes(x=logYi)) + geom_histogram() +
  xlab("Home Prices") + ggtitle("Histogram: Y outcome variable (log transformation)")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



d) Watch the following videos to learn about "boxcox" transformation and use the following commands to calculate the exact value of lambda needed for making the boxcox transformation. After you made the transformation draw the histogram.

```
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 3.6.2
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

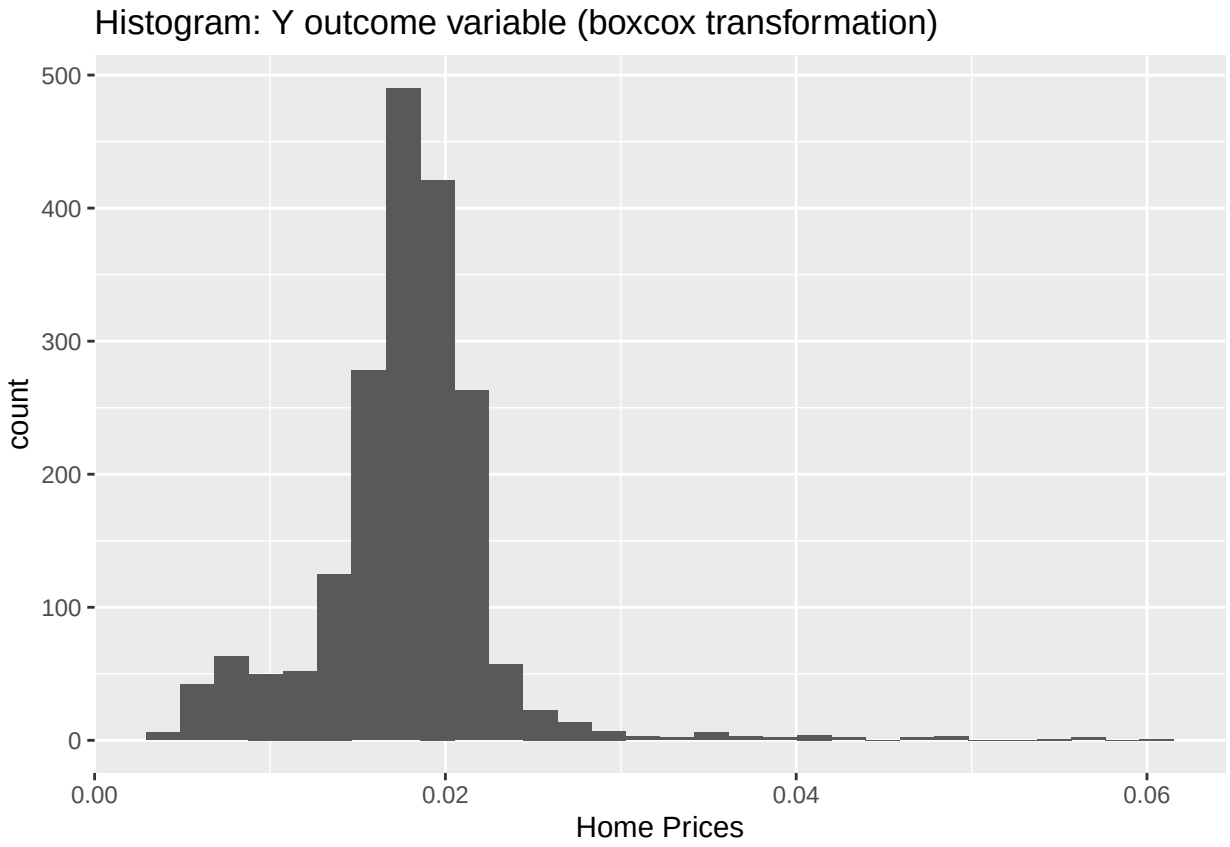
```
lambda <- BoxCox.lambda(houstonrealesate$Yi)
lambda #recommended lambda for transformation
```

```
## [1] -0.9423097
```

```
boxcoxYi <- houstonrealesate$Yi^lambda

ggplot(mapping = aes(x=boxcoxYi)) + geom_histogram() +
  xlab("Home Prices") + ggtitle("Histogram: Y outcome variable (boxcox transformation)")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



e) Which histogram looks best?

The log transformation does a much better job at creating a more symmetric, normally spread distribution.

f) Install the library (moments) to calculate the skewness of a histogram. Using this library, calculate the skewness of Yi and the histograms resulting from log and boxcox transformation. Decide which transformation did the best job of improving the skewness.

```
library(moments)
```

```
#no transformation
skewness(houstonrealesate$Yi)
```

```
## [1] 3.377336
```

```
#log transformation
skewness(logYi)
```

```
## [1] 1.274064
```

```
#boxcox transformation
skewness(boxcoxYi)
```

```
## [1] 1.608676
```

The skewness of the log transformation is approximately 1.27, which is close to the 1 for a normally distributed histogram. Compared to the higher skewness values for the original data (3.377) and the boxcox transformation (1.608676), this indicates that the log transformation does the best job at improving the skewness.

g) Use the Shapiro test to test for the normality of Yi , logYi, and Yi resulting from boxcox transformation. The null hypothesis is that the histogram of interest is normal vs. the alternative hypothesis that it is not. Use the following command.

```
#no transformation
shapiro.test(houstonrealesate$Yi) #reject the null, histogram is NOT normal
```

```
##
##  Shapiro-Wilk normality test
##
## data:  houstonrealesate$Yi
## W = 0.64268, p-value < 2.2e-16
```

```
#log transformation
shapiro.test(logYi) #reject the null, histogram is NOT normal
```

```
##
##  Shapiro-Wilk normality test
##
## data:  logYi
## W = 0.85284, p-value < 2.2e-16
```

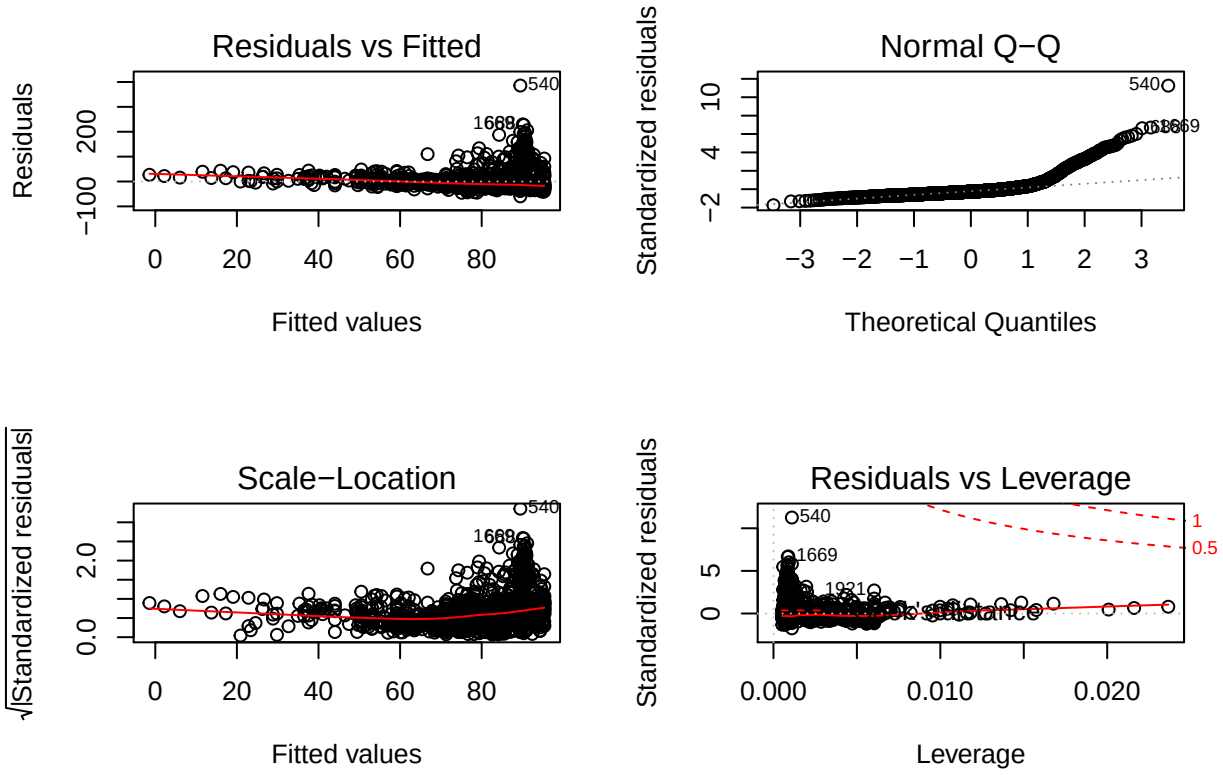
```
#boxcox transformation
shapiro.test(boxcoxYi) #reject the null, histogram is NOT normal
```

```
##
##  Shapiro-Wilk normality test
##
## data:  boxcoxYi
## W = 0.84442, p-value < 2.2e-16
```

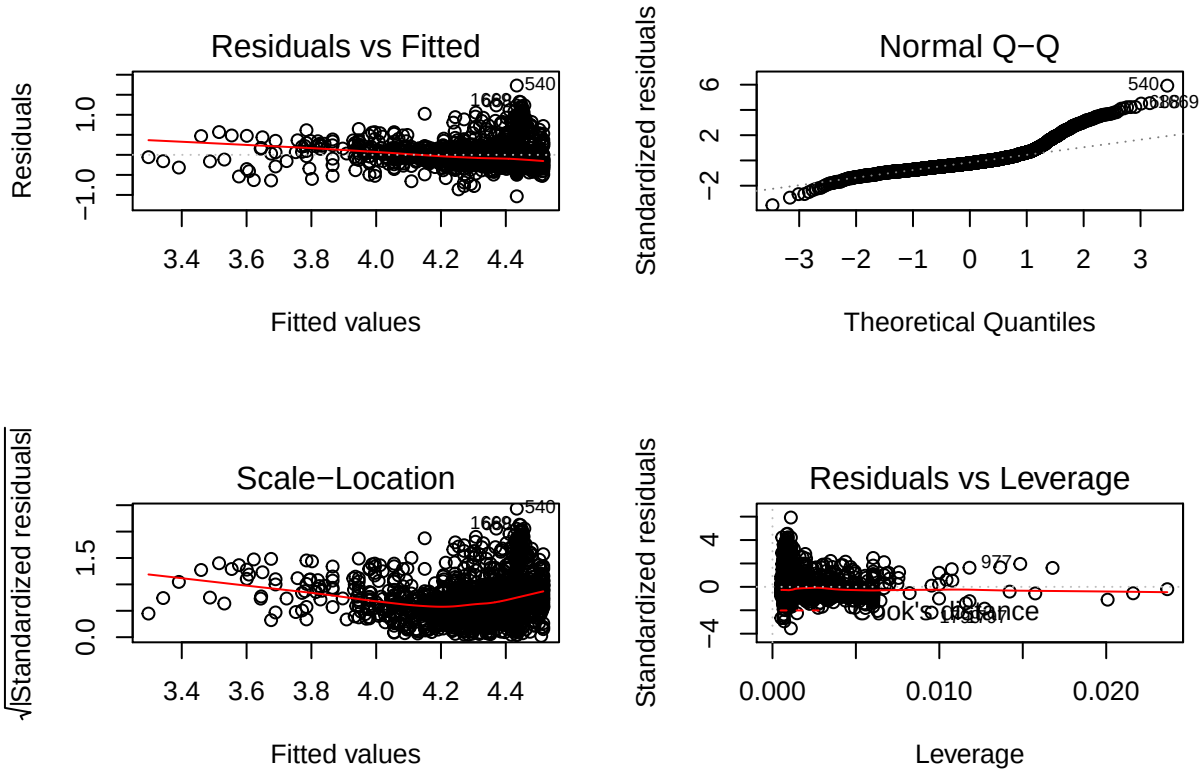
According to the Shapiro Test, the original data is not normally distributed, which is visually obvious in the first, right-skewed histogram. The Shapiro Test also proves that the log and boxcox transformations still do not follow a normal distribution, despite their improvement over the non-transformed data.

h) Draw the plot of residuals and qq plot for Y, logYi, and the boxcox transformation of Yi. Compare the plots and decide which transformation did the best job of taking care of the problem and whether any of them solved the problem

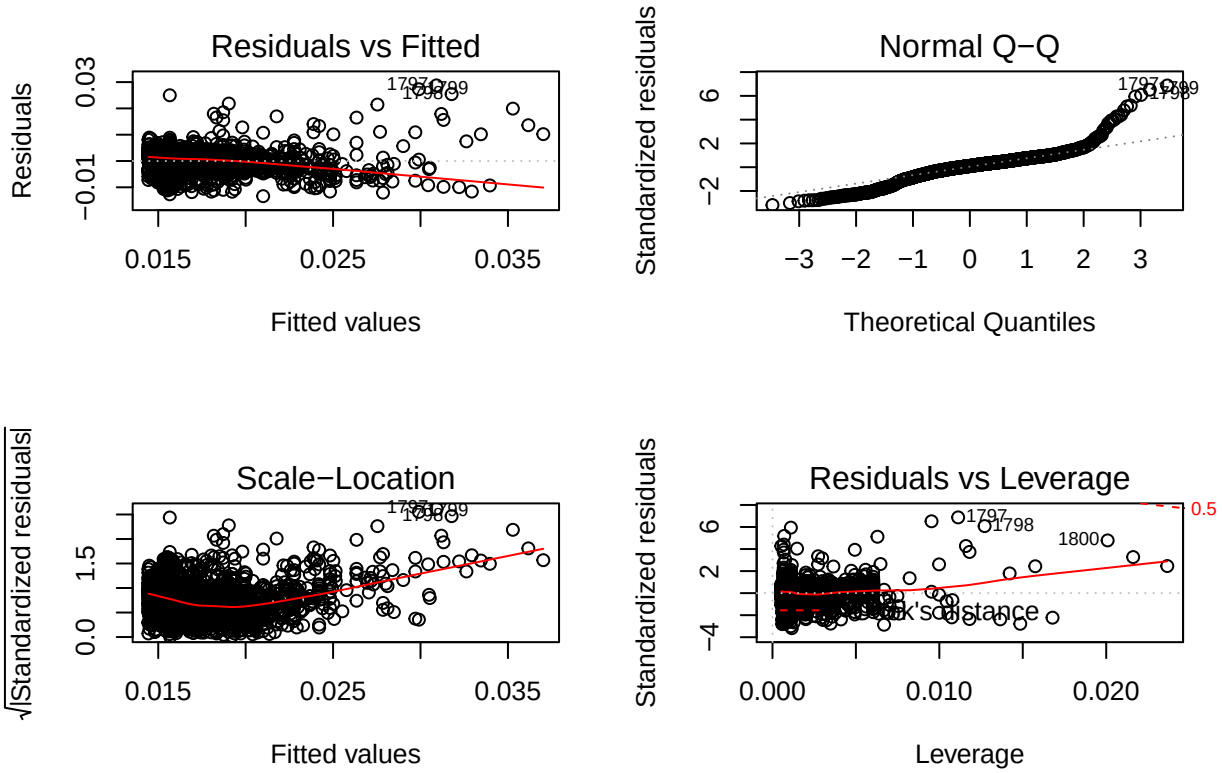
```
# y no transformation
m.original <- lm(Yi ~x1i + x2i, data=houstonrealesate)
par(mfrow=c(2,2))
plot(m.original)
```



```r
# y log transform
m.log <- lm(log(Yi) ~x1i + x2i, data=houstonrealesate)
par(mfrow=c(2,2))
plot(m.log)
```



```
# y boxcox transform
m.boxcox <- lm(Yi^lambda ~x1i + x2i, data=houstonrealesate)
par(mfrow=c(2,2))
plot(m.boxcox)
```



From the plots above, it looks like the log transformation does a slightly better job at attempting to fix the normality violation. However, while the log and boxcox transformations do improve the skewness of the house price variable, they still do not solve the normality assumption violation.

Problem Three

Using women powers data and codebook posted in the homework folder on week two:

a) Complete the following table

```
womenpowers <- read.csv("womenpowers.csv")
attach(womenpowers)
summary(womenpowers) #quick look at the variables
```

```
##       age         nonint     mwork          meduc           income
##  Min.   :14.08   no :2777   no :1840   Min.   : 0.00   Min.   :    0
##  1st Qu.:16.42   yes: 928   yes:1865   1st Qu.:11.00   1st Qu.: 4800
##  Median :18.50                         Median :12.00   Median : 7354
##  Mean   :18.36                         Mean   :11.61   Mean   : 8646
##  3rd Qu.:20.33                         3rd Qu.:12.00   3rd Qu.:10667
##  Max.   :22.00                         Max.   :20.00   Max.   :49497
##      nsibs         fundprot   tradrole
##  Min.   : 0.000   no :2667   A : 417
##  1st Qu.: 2.000   yes:1038   D :1675
##  Median : 3.000              SA: 131
##  Mean   : 3.328              SD:1482
##  3rd Qu.: 4.000
##  Max.   :17.000
```

```
#nonint: non-intact family structure at age 14 (yes or no)
intact <- womenpowers[nonint == "no",]
  mean1 <- mean(intact$income); var1 <- var(intact$income); n1 <- length(intact$income)
non_intact <- womenpowers[nonint == "yes",]
  mean2 <- mean(non_intact$income); var2 <- var(non_intact$income); n2 <- length(non_intact$income)

table <- rbind(c(mean1, var1, n1), c(mean2, var2, n2))
colnames(table) <- c("Mean of Income", "Variance of Income", "Sample Size")
rownames(table) <- c("Intact Family", "Non-Intact Family")
table
```

```
##                   Mean of Income Variance of Income Sample Size
## Intact Family           9182.084           37256692        2777
## Non-Intact Family       7042.962           24563880         928
```

```
(income.mean <- mean(income)) # Overall mean for income
```

```
## [1] 8646.293
```

```
(income.variance <- var(income)) # Overall variance for income
```

```
## [1] 34929285
```

b) Once you have complete the above table, calculate...

```
# 1) t-value for testing the null hypothesis:
# average income is similar for intact and non-intact families.
t_val <- (mean1 - mean2)/sqrt(var1/n1 + var2/n2)
t_val
```

```
## [1] 10.7109
```

```
#2) Confidence interval for the two-sample test of the mean. Interpret this interval within context.
c((mean1 - mean2) - qnorm(0.975)*sqrt(var1/n1 + var2/n2), #lower bound, uses two-tailed 95% CI
  (mean1 - mean2) + qnorm(0.975)*sqrt(var1/n1 + var2/n2)) #upper bound
```

```
## [1] 1747.689 2530.555
```

We are 95% confident that the true difference in mean family incomes in the population is between 1748 and 2531 units. In other words, non-intact families make 1748 to 2531 less than intact families.

```
#3) Sum of square between.
SSbetween <- n1*(mean1 - income.mean)^2 + n2*(mean2 - income.mean)^2
SSbetween
```

```
## [1] 3182780904
```

```
#4) Sum of square within.
SSwithin <- sum((intact$income - mean1)^2) + sum((non_intact$income - mean2)^2)
SSwithin
```

```
## [1] 126195292388
```

```
#5) F-value for analysis of variance.
j <- 2
N <- n1 + n2
F_val <- (SSbetween/(j-1))/(SSwithin/N-j)
F_val
```

```
## [1] 93.44409
```

```
#6) Compute and interpret R-squared within context.
R2 <- SSbetween / (SSbetween + SSwithin)
R2
```

```
## [1] 0.02460062
```

```
# all of these values can also be obtained using the following
# t.test(income~ nonint)
# summary(aov(income ~ nonint, data=womenpowers))
# summary(lm(income ~ nonint, data=womenpowers))
```

The high F-value indicates that the difference in the means for the non-intact family variable is statistically significant. However, the 2.5% R^2 is so low that it is not practically significant in prediction.

Problem Four

Using campus climate data, examine the relationship between students' perception of academic success and class comfort.

```
campusclimate <- read.csv("campusclimate.csv")
attach(campusclimate)
```

a) Recode Q10-A-5 based on the directions given below.

```
# Q10-A-5: I have performed academically as well as I expected I would.
# We will call this "students' perception of academic success"
table(campusclimate$Q10_A_5) #original
```

```
##
##    1    2    3    4    5    6
##  597 1690 1172 1424  475    9
```

```
#remove level 3 (neither) and 6 (unknown)
is.na(Q10_A_5) <- Q10_A_5 == 3 | Q10_A_5 == 6
Q10_A_5 <- factor(Q10_A_5)

library(car)
#recode all levels of agree to 'agree' and all levels of disagree to 'disagree'
academic.performance <- recode(Q10_A_5, "c('1','2') = 'agree'; c('4','5') = 'disagree'")
table(academic.performance) #recoded
```

```
## academic.performance
##    agree disagree
##     2287     1899
```

b) Recode the classcomfort based on the directions given below

```
table(campusclimate$classcomfort) #original
```

```
##
##    1    2    3    4    5    6
##  910 2913 1172  328   52    3
```

```
#remove NAs
is.na(classcomfort) <- classcomfort == 6 ; classcomfort <- factor(classcomfort)

#recode and reorder from least comfortable to most
classcomfort <- recode(classcomfort,"'1'='very comfortable'; '2'='comfortable'; '3'='meh';
                      '4'='uncomfortable' ; '5'='uncomfortable'",
                      levels = c("uncomfortable", "meh", "comfortable","very comfortable"))
table(classcomfort)
```

```
## classcomfort
##    uncomfortable              meh      comfortable very comfortable
##              380             1172             2913              910
```

c) After recoding, calculate chi-square to examine the relationship between students' perception of their academic success and their perception of class comfort. Make class comfort row and academic performance column

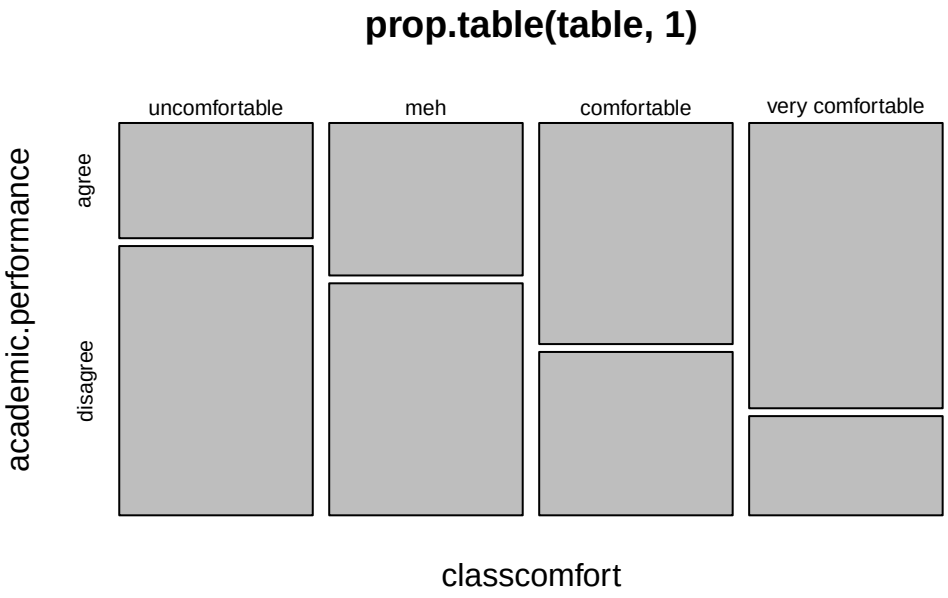
```
chisq.test(x = classcomfort, y= academic.performance) #reject null, so variables are dependent
```

```
##
##  Pearson's Chi-squared test
##
## data:  classcomfort and academic.performance
## X-squared = 281.07, df = 3, p-value < 2.2e-16
```

d) Calculate and interpret row percentage.

```
table <- table(classcomfort, academic.performance)
prop.table(table,1); plot(prop.table(table,1))
```

```
##                   academic.performance
## classcomfort           agree  disagree
##   uncomfortable    0.2996845 0.7003155
##   meh              0.3965909 0.6034091
##   comfortable      0.5750784 0.4249216
##   very comfortable 0.7420213 0.2579787
```



As class comfort increases, perceived satisfaction with academic performance increases as well.

e) Calculate the odds ratio of students' perception of academic success in a comfortable class climate compared to an uncomfortable class climate. Interpret the odds ratio you find within context

```
addmargins(table)
```

```
##                   academic.performance
## classcomfort       agree disagree  Sum
##   uncomfortable       95      222  317
##   meh                349      531  880
##   comfortable       1283      948 2231
##   very comfortable   558      194  752
##   Sum               2285     1895 4180
```

```
odds_comfort <- 1283/948 #comfortable agree/disagree
odds_uncomfort <- 95/222 #uncomfortable agree/disagree

#the odds ratio of students' perception of academic success in a comfortable class climate
total_odds <- odds_comfort/ odds_uncomfort
total_odds
```

```
## [1] 3.162625
```

From the odds ratio above we can see that the perception of academic success is 3.16 times higher for students who say they are comfortable, rather than students who feel uncomfortable in the class climate. In other words, students who feel comfortable will have a perception of academic success that is about three times higher than students who do not feel comfortable.

Problem Five

Suppose you were the TA, Given the following information, how would you explain questions "a", "b", and "c" to your students? Hint: See page 92 on the chapter on chi-square F-test in ANOVA.

a) What is the difference between α_j and $\hat{\alpha}_j$?

α_j is the difference between the mean of population of group j and the mean of the overall population.

$\hat{\alpha}_j$ is the difference between the mean of a sample from the population of group j and the mean of a sample from the overall population.

Note that the overall population includes the population of group j.

b) Given the above, explain why we say that if we fail to reject the null, the expected value of F is one.

If we fail to reject the null this means that the p-value for the t-test is greater than 0.05 and that $\hat{\alpha}_j$ is close to 0. This means that there is no statistical difference between $MS_{between}$ and MS_{within}. In other words, the amount of variation we are able to explain by comparing the difference between groups is not significantly more than the amount of variation we cannot explain, which is the difference within the groups. If the amount of variation explained is similar for both between and within, the ratio between them will be approximately 1.

c) Given the following output, do the findings represent $E(MS_{between})$ and $E(MS_{within})$? Yes or No? Explain your reasoning.

Yes, because the results show that the null hypothesis is false, we know that a treatment effect between groups does exist. Therefore, the expected $MS_{between}$ is an estimate of the population error variance in addition to a function of the squared treatment effect. On the other hand, if the null were true, then the expected $MS_{between}$ would only be an estimate of the population error variance. For either case, the expected MS_{within} will be the population error variance.

Problem Six

Suppose you were given the following output and plot, how would you explain the findings within context to a **non-statistical audience**.

a) The question that we are trying to answer.

Given the following output, it appears that we are using a dataset to determine whether there is a relationship between the birthweight of a child and the smoking habits of the mother. Specifically we are dividing the mothers into two groups (smoking and non-smoking) to observe and compare the difference in the birthweight across groups.

b) The significance of findings.

The test used to compare the difference in birthweight outputs a p-value close to 0. This p-value indicates the chance that the birthweights come from the same population, therefore we think it is extremely unlikely that the average birthweight of the two groups is the same. Specifically we have enough evidence to say that there most likely is a difference in birthweight of children born to mothers with a history of smoking vs. those without.

c) Explanation of error bars.

The graph has a point for the mean of the two groups in the sample we have collected. The error bar is a statistical attempt at showing the possible range for the mean if we observed the entire population instead of just taking this sample of around 1,000 observations. Based on the bars shown on the graph, we see that there is almost no chance that the actual population means are the same since the ends of the error bars they do not overlap.