STAT 217 Project 2

Ben Miller

Due: October26, 2020 by 11:00pm in Gradescope

A Comparison between Multiple Regression Models and CUN-BAE Equation to Predict Body Fat in Adults

For this project you are encouraged to work in groups of up to four people. Save all files associated with the project in your STAT 217 folder. Questions should be answered in this Markdown file in bold text. All code and output must be included in final submission. Knit your completed Markdown file to Word, save this as a PDF, and submit to Gradescope by the deadline. When submitting your project, please indicate which page each question is on. Failure to do so may result in a loss of points.

Read the paper by Fuster-Parra et al. (2015) available at https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0122291 and posted on D2L with some edits that someone should have identified before this was published. Their data set is posted in the author's original format, a .txt file called "Fuster_data1.txt". Code is provided below to import data from this format as it is slightly different from our .xlsx or .csv formats we have worked with before.

Citation:

- Fuster-Parra P, Bennasar-Veny M, Tauler P, Yañez A, López-González AA, et al. (2015) A Comparison between Multiple Regression Models and CUN-BAE Equation to Predict Body Fat in Adults. PLOS ONE 10(3): e0122291. https://doi.org/10.1371/journal.pone.0122291

1) Use the following code to read in the data set. Are there any missing values in the data set? What did they say about missing values in the paper?

```
library(readr)
Fuster_data1 <- read_table2("Fuster_data1.txt")

Fuster_data1$Gender <- factor(Fuster_data1$Gender)
levels(Fuster_data1$Gender) <- c("Male", "Female")

summary(Fuster_data1)
```

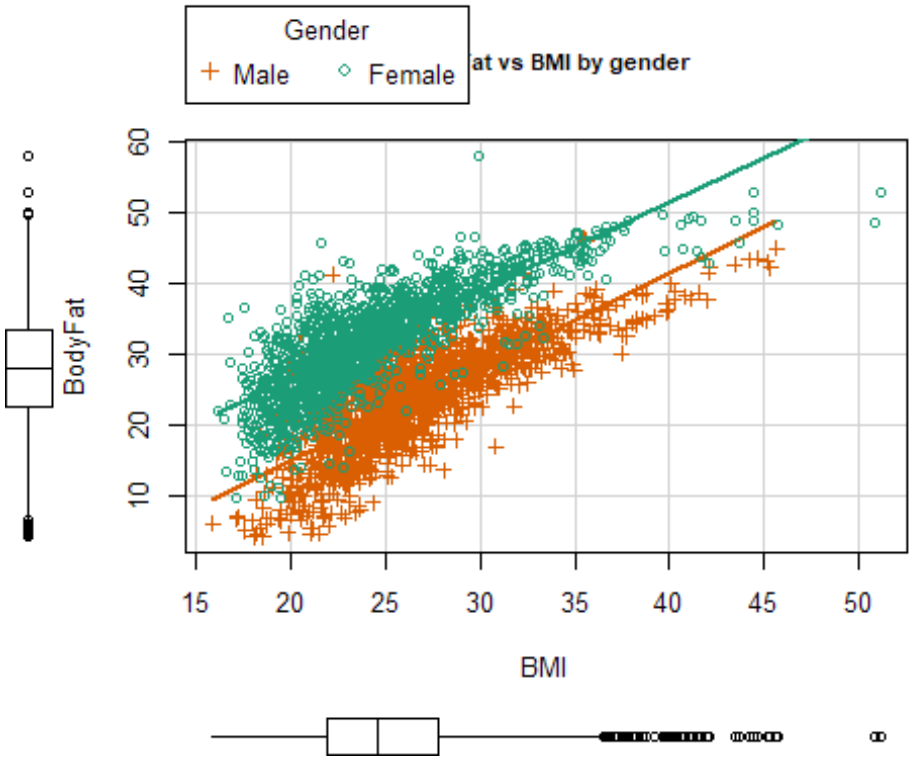
```
##       ID                  Age              BAI             BMI        
##  Length:3200        Min.   :16.00   Min.   :12.28   Min.   :15.84  
##  Class :character   1st Qu.:31.00   1st Qu.:25.24   1st Qu.:21.97  
##  Mode  :character   Median :39.00   Median :27.89   Median :24.61  
##                     Mean   :39.19   Mean   :28.67   Mean   :25.30  
##                     3rd Qu.:47.00   3rd Qu.:31.38   3rd Qu.:27.77
```

```
##                     Max.   :69.00   Max.   :65.88   Max.   :51.13
##     BodyFat         Gender
##  Min.   : 4.00   Male  :1474
##  1st Qu.:22.60   Female:1726
##  Median :28.00
##  Mean   :27.95
##  3rd Qu.:33.40
##  Max.   :58.20
```

The paper says that 3223 people were selected, but that 23 refused, so their total data set was 3200 people. We have the same length as the paper, and we have the same number of men and women as the paper did, so we are all good.

2) Make a plot of `BodyFat` as a function of `BMI` by the `Gender` of the subjects. Hint: Use the `scatterplot` function in the `car` package (see page 22 of the Week 7 (Day 20) notes). This will resemble their Figure 1 except that BMI is not log-transformed. Compare the relationships between body fat and BMI for the two genders. Do they have similar directions? Strengths? Are they linear? What is different between the two groups?

```
library(car)
scatterplot(BodyFat~BMI|Gender,data=Fuster_data1,pch=c(3,21),smooth=F,
boxplots="xy", col=c("#D95F02","#1B9E77"),main=list("Body Fat vs BMI by
gender",cex=0.85))
```

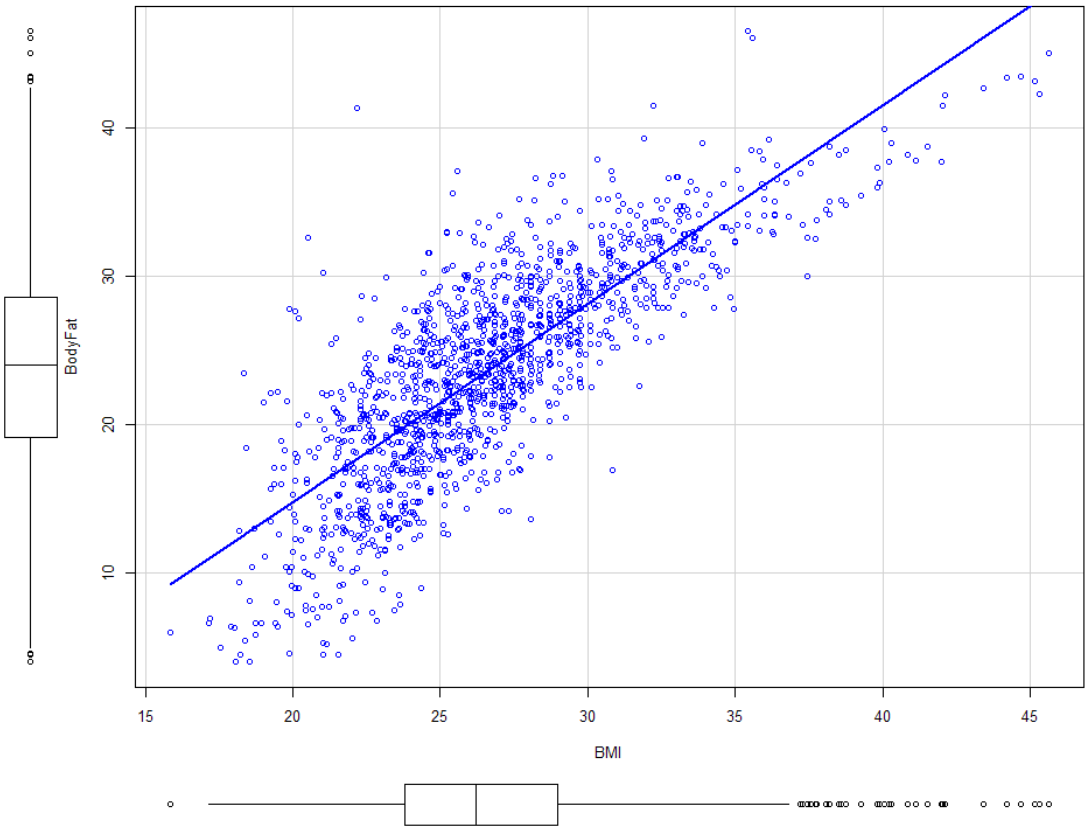


In Terms of the spread of the variables, it is fairly similar between the two groups. The Female group is generally consistently above the Male group. They both have almost the same slope, which is positive. They both look like pretty strong

relationships, too. They look quite linear, though the groups look like they have a slight curve to them. The differences between them are that the Male group appears to be more right tailed, and it has a wider range. The right tail also looks a little more spread out than the Female group. Other than that, they look very similar.

3) Add a comment after # to each line of code in the following code chunk that explains what each one is doing. Interpret the result from the last line of code (not inside the code chunk - write a sentence outside the chunk).

```
library(mosaic) # importing the mosaic package
d_m <- subset(Fuster_data1, Gender=="Male") # creating a subset of
Fuster_data1 that only has Males and storing it in d_m
scatterplot(BodyFat ~ BMI, data=d_m, smooth=F) # creating a scatterplot of
d_m
```




```
cor(BodyFat ~ BMI, data=d_m) # printing the correlation coefficient

## [1] 0.8012112
```

The correlation coefficient between Body Fat and BMI for Males is 0.8012, so there is a strong positive relationship between the variables.

4) With the data set `d_m`, fit a linear model for `BodyFat` (response) with `BMI` as a predictor. Write out the estimated regression model in the context of the problem.

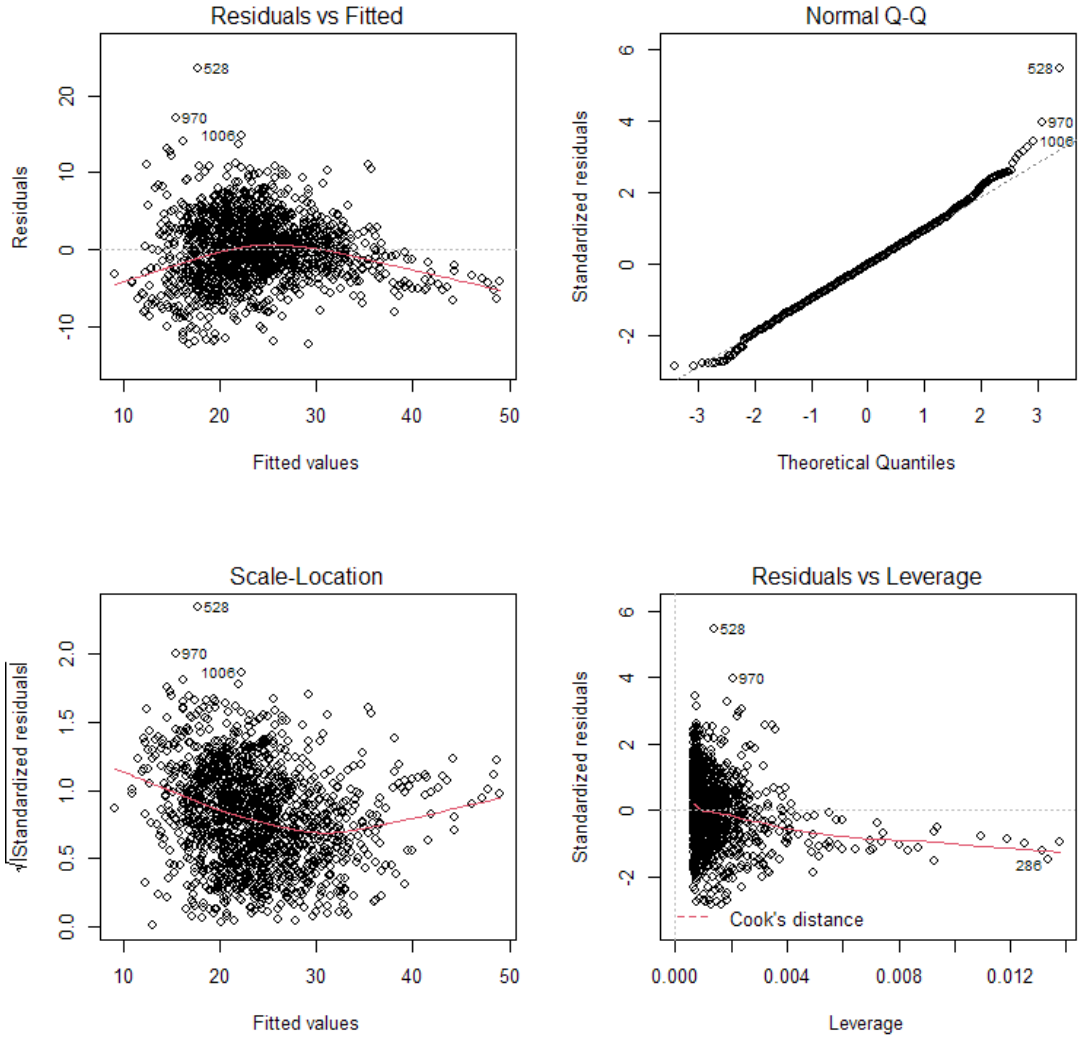
```
lm.male<- lm(BodyFat~BMI, data=d_m)
summary(lm.male)

##
## Call:
## lm(formula = BodyFat ~ BMI, data = d_m)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.3508  -2.8365  -0.0419   2.7103  23.6083
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.00976    0.70416  -17.05   <2e-16
## BMI           1.33855    0.02606   51.37   <2e-16
##
## Residual standard error: 4.312 on 1472 degrees of freedom
## Multiple R-squared:  0.6419, Adjusted R-squared:  0.6417
## F-statistic:  2639 on 1 and 1472 DF,  p-value: < 2.2e-16
```

$\widehat{BodyFat}$ = -12.0098 + 1.3386(BMI)

5) Generate our regular suite of diagnostic plots for the model you fit in the previous question (you'll need to replace the ? in the `plot` function below with the name of your model from the previous question). Assess the linearity and no influential observations conditions based on these plots. Be specific about which plot/information you are using for each assessment.

```
par(mfrow=c(2,2))
plot(lm.male)
```



Linearity: There is moderate to strong evidence against the assumption of linearity. In the scatterplot of the data, there was a noticeable curve in the data. Looking at the Residuals vs Fitted plot, there is a distinct curve in the residuals.

No Influential Observations: There are no Influential observations. Looking at the Residuals vs Leverage plot, there is no observation with a Cook's distance close to 0.5 or greater.

6) Identify an observation in the `d_m` data set (not the first one) and report the observed values of the `BMI` and the `BodyFat` for that subject. Use the previously fit model to generate a predicted body fat percentage for that subject. Then calculate the residual for that subject. Did the model under- or over-estimate the body fat for that subject? Show your work.

Observation number 1345 in the data text file with gender 0, which is a male so it's in d_m, has an observed BMI of 32.2831 and observed body fat percentage of 32.6.

$\widehat{BodyFat}$ = -12.0098 + 1.3386(32.2831) = 31.2044 $e_i = y_i - \hat{y}_i = 32.6 - 31.2044 =$ 1.3956%

The model slightly under-estimated this observation.

7) Interpret the slope coefficient from your model in context. Note that the units on `BMI` are kg/m^2 and the units on body fat percent are percentages.

For a one kg/m^2 increase in BMI, we estimate the body fat, on average, to increase by 1.3386%.

8) Provide a conclusion, in context, for the slope coefficient associated with `BMI` from your fitted model.

We have strong evidence against the null hypothesis that there is no linear relationship between BMI score and body fat percentage (t = 51.37, t_{1472}, p-value < 0.001). Thus, we conclude that there is a linear relationship between these variables.

9) Write a scope of inference for the results of the model fit with the `d_m` data, in context (this should **not** be a scope of inference for the full data set).

The authors randomly sampled from the population of government workers in Spain, these results are for the males in the data, and there was not random assignment, so we can say that BMI had a correlative relationship with body fat percentage in the population of male government workers in Spain.

10) On pages 2-3, the authors discuss a second data set "dataset B." What is the difference between this and the "full" data set provided in the S1 Dataset? How would that change the SOI you just wrote if you had been working with that data set?

Dataset B is a subset of the full data set only containing people with a BMI over 25. The scope of inference would change to a correlative relationship with BMI and body fat percentage in the population of government workers in Spain with a BMI over 25.

11) **NA. No other assistance**