STAT 217 Project 3

Ben Miller

Due: November 16, 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

These are the same data as used in Project 2. The code below will import and prepare the data for you.

```
library(effects)
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
```

1) What do the acronyms of `BIA`, `BMI`, and `BAI` used in the paper stand for and how are they calculated?

BIA is Bioelectrical Impedence Analysis and is calculated using impedence data obtained by a 50 KHz 0.8mA constant sine wave sent through participants standing barefoot on metal contacts. BMI is Body Mass Index and it is calculated by weight in kilograms divided by height in meters squared. BAI is Body Adiposity Index and it is calculated by the equation $BAI = (HipCircumferenceincm)/((heightinm)^{1.5} - 18)$

2) It is always good practice to verify that the data set that you think matches the one used in someone else's research actually does match. Write and report the output from code that generates the summary statistics that they report in their Table 1 for the `BAI` row (hint: use the `favstats` function). State the statistics that you are comparing and whether you get a match. Remember to discuss sample sizes too. If you don't get a match, discuss the differences.

```
library(mosaic)
favstats(BAI~Gender, data=Fuster_data1)
```

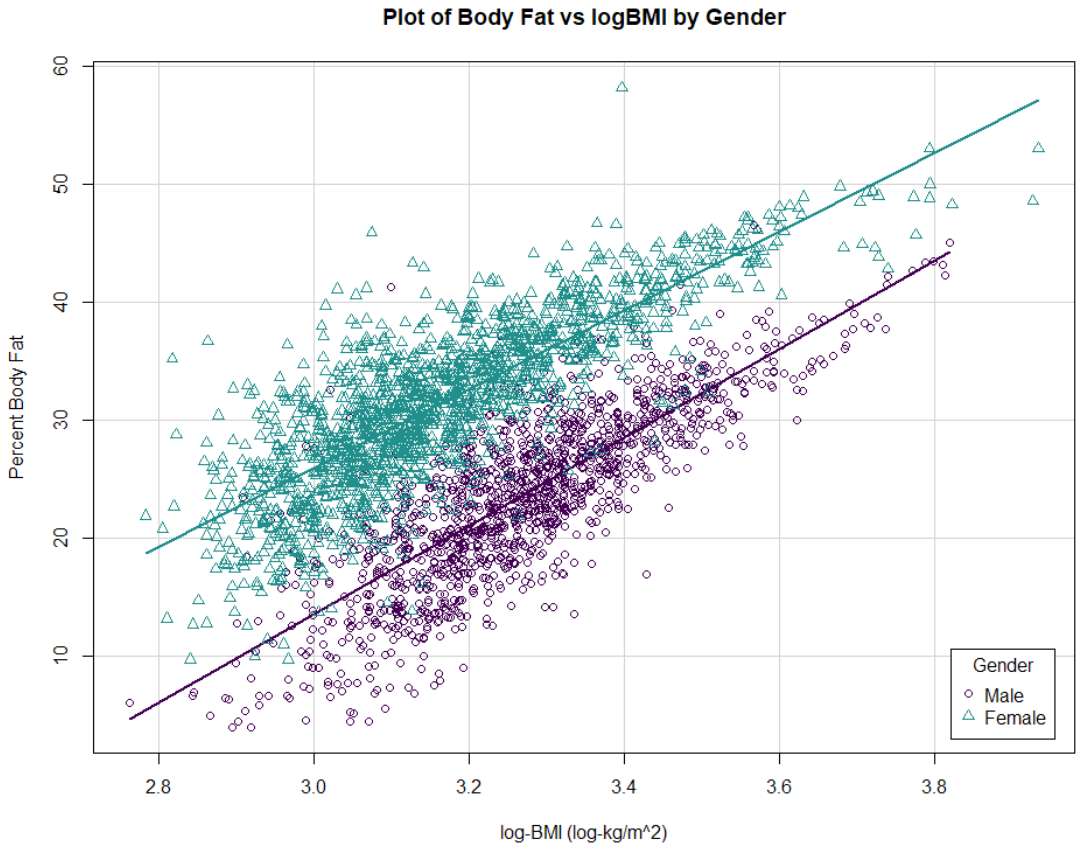
```
##   Gender     min      Q1   median      Q3     max     mean       sd    n
## 1   Male 14.4236 24.0412 26.23665 28.9202 46.9664 26.64943 3.944228 1474
## 2 Female 12.2770 26.6891 29.43420 33.2796 65.8754 30.38711 5.349370 1726
##   missing
## 1       0
## 2       0
```

The mean BAI for men in the paper is reported as 27 and the standard deviation is 4. We have 26.65 in the above results, and we have a standard deviation of 3.94. For women, the mean BAI in the paper is reported as 30 with a standard deviation of 5. We have a mean of 30.39 and a standard deviation of 5.35. The paper also states that there were 1474 men and 1726 women in their study which matches what we have. The only differences are that our results are a little more specific.

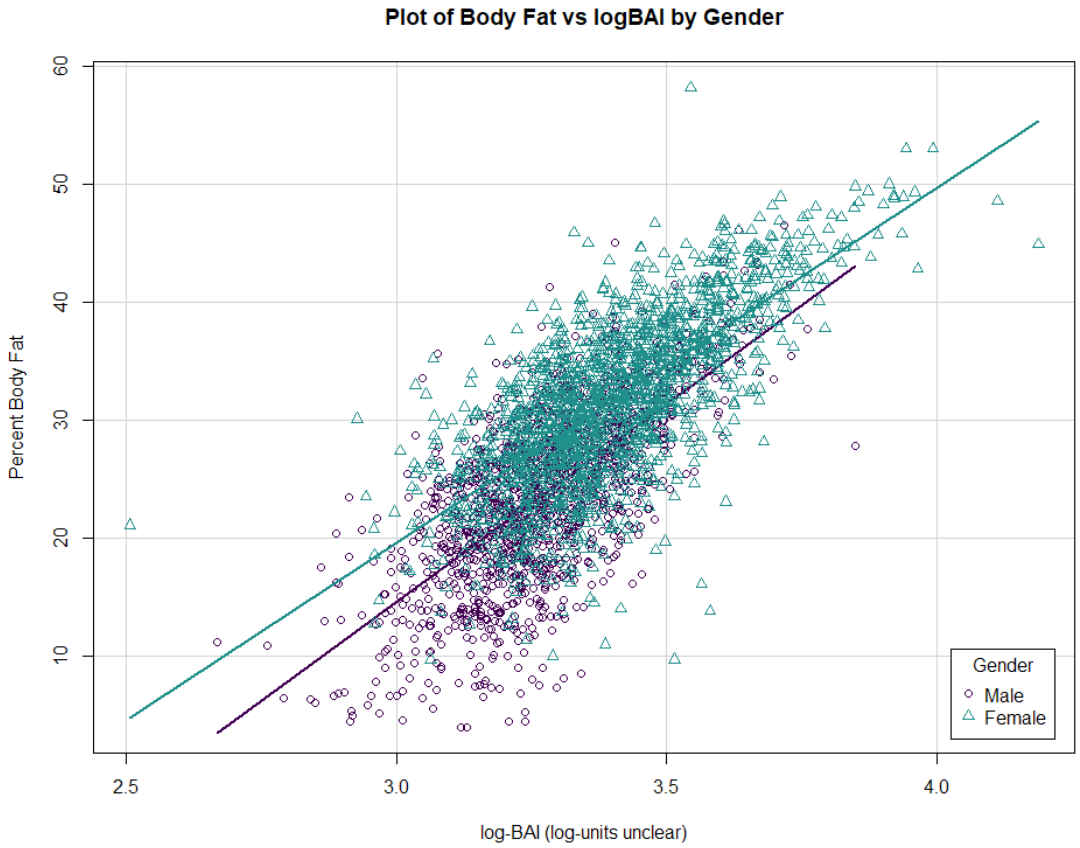
3) The following code creates new versions of the `BMI`, `BAI`, and `Age` variables that are `log` transformed. This doesn't quite match what the researchers did (they used log10), so don't expect results that match the researchers', but by using natural logs you can use the rules discussed in Chapter 7 for interpretation of log-transformed predictors (there are similar rules if you needed to use log10 but we have not discussed them).

The code also makes plots similar to their Figure 1. Run the code and provide the plots. No discussion needed.

```
Fuster_data1$logBMI <- log(Fuster_data1$BMI)
Fuster_data1$logBAI <- log(Fuster_data1$BAI)
Fuster_data1$logAge <- log(Fuster_data1$Age)

library(car)
library(viridis)
scatterplot(BodyFat ~ logBMI|Gender, data= Fuster_data1, col=viridis(3)[-3],
            xlab="log-BMI (log-kg/m^2)", ylab="Percent Body Fat",
            main="Plot of Body Fat vs logBMI by Gender",
            legend=c(title="Gender", coords="bottomright"), smooth=F)
```




```
scatterplot(BodyFat ~ logBAI|Gender, data= Fuster_data1, col=viridis(3)[-3],
            xlab="log-BAI (log-units unclear)", ylab="Percent Body Fat",
            main="Plot of Body Fat vs logBAI by Gender",
            legend=c(title="Gender", coords="bottomright"), smooth=F)
```



4) For the relationship between `logBAI` and `BodyFat`, which group has a steeper (larger) slope? Does it appear that an interaction term might be necessary between `logBAI` and `Gender`?

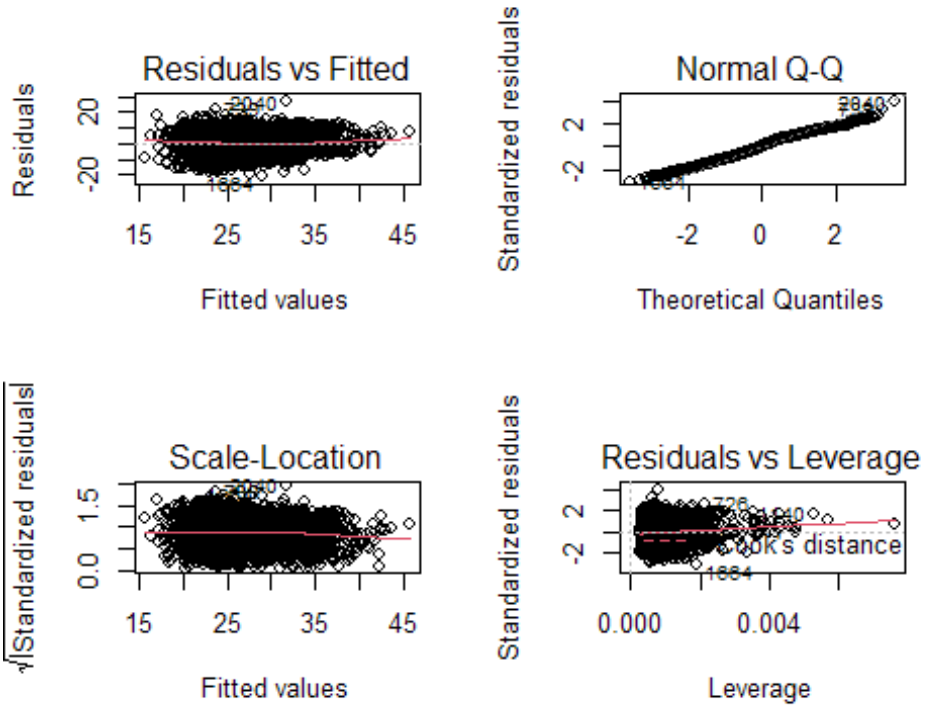
The Male group has a steeper slope than the Female group. It appears than an interaction term between 'logBAI' and 'Gender' might be necessary because the difference in slopes is noticeable and the sample set is quite large.

5) The researchers consider a wide array of competing models for the body fat responses. One model that they consider includes `logAge` and either `logBAI` or `logBMI` (their Models 2a and 2b). We are not convinced that they needed to log-transform `Age` in these models. The following code fits our versions of these models **without** transforming `Age` and produces a suite of information for each model. Run this code and provide the `summary()` output for each model. No discussion needed.

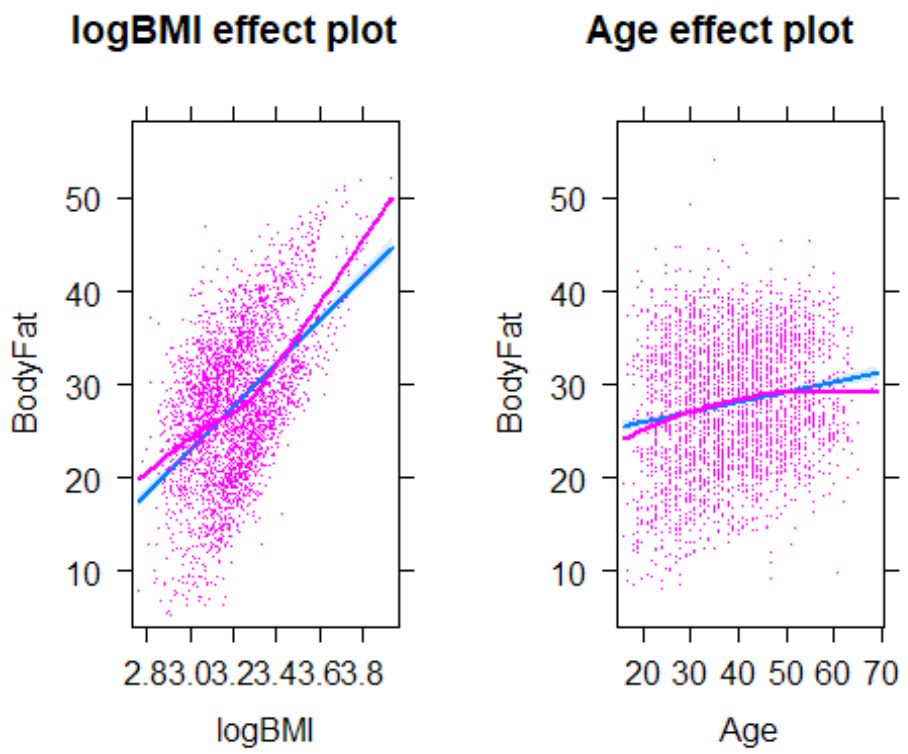
```
model2a_alt <- lm(BodyFat ~ logBMI + Age, data= Fuster_data1)
summary(model2a_alt)

##
## Call:
## lm(formula = BodyFat ~ logBMI + Age, data = Fuster_data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -20.5289  -5.2277   0.2235   5.5054  26.4731
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -50.8303     2.2321 -22.773   <2e-16
## logBMI       23.1919     0.7225  32.099   <2e-16
## Age           0.1075     0.0117   9.187   <2e-16
##
## Residual standard error: 6.796 on 3197 degrees of freedom
## Multiple R-squared:  0.3076, Adjusted R-squared:  0.3071
## F-statistic:   710 on 2 and 3197 DF,  p-value: < 2.2e-16
```

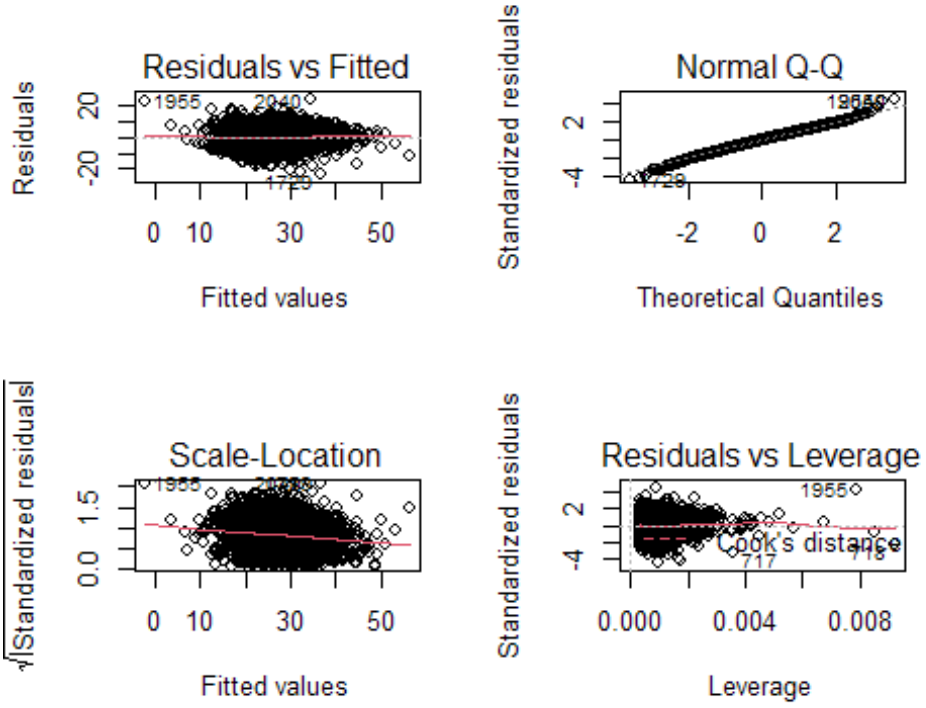
```
par(mfrow=c(2,2))
plot(model2a_alt)
```



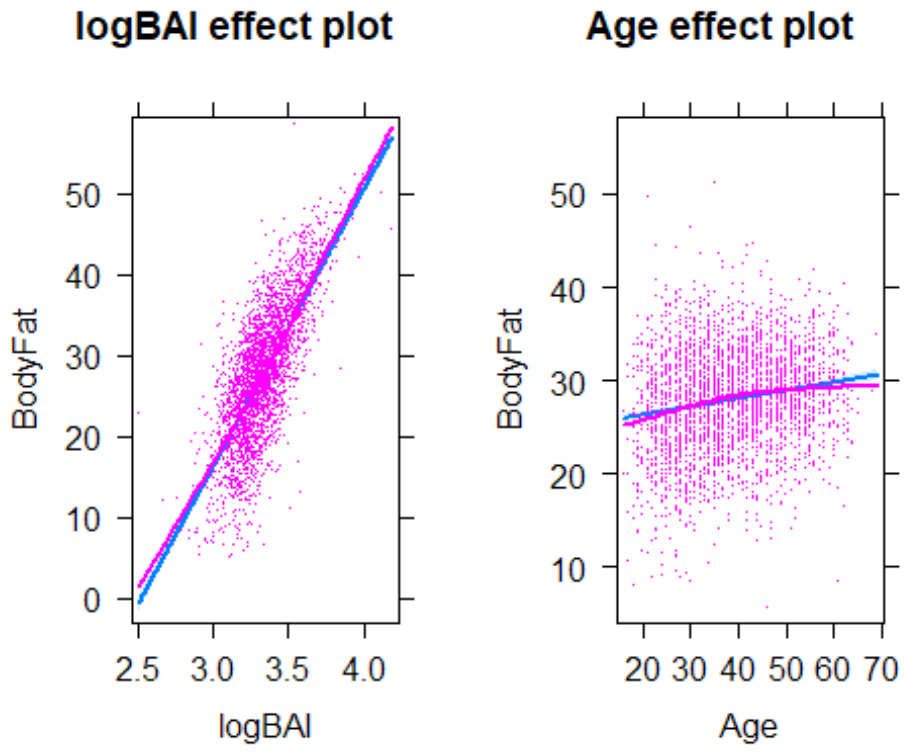

```
plot(allEffects(model2a_alt, residuals=T), partial.residual=list(pch="."))
```




```
model2b_alt <- lm(BodyFat ~ logBAI + Age, data= Fuster_data1)
summary(model2b_alt)

##
## Call:
## lm(formula = BodyFat ~ logBAI + Age, data = Fuster_data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.9704  -3.3475   0.1023   3.6548  23.6260
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -90.072032   1.863409  -48.34   <2e-16
## logBAI       34.304466   0.573445   59.82   <2e-16
## Age           0.087250   0.009089    9.60   <2e-16
##
## Residual standard error: 5.368 on 3197 degrees of freedom
## Multiple R-squared:  0.568,  Adjusted R-squared:  0.5677
## F-statistic:  2102 on 2 and 3197 DF,  p-value: < 2.2e-16

par(mfrow=c(2,2))
plot(model2b_alt)
```




```
plot(allEffects(model2b_alt, residuals=T), partial.residual=list(pch="."))
```



6) Using `model2b_alt`, interpret the coefficient associated with `Age` in context.

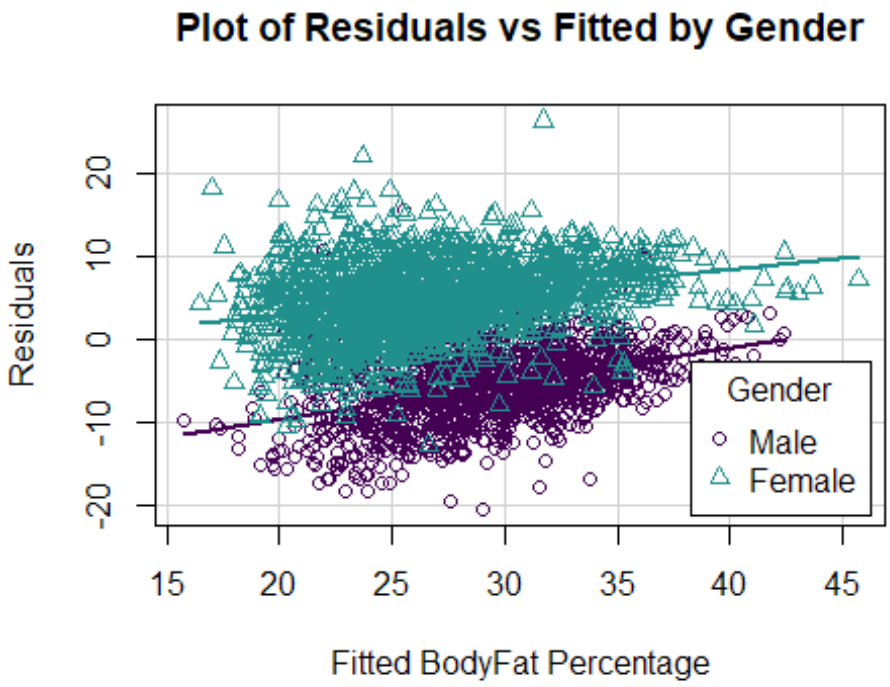
For a one year increase in age, we estimate the BodyFat percentage to go up by 0.087% after accounting for the log of BAI.

7) Fill in the blanks below to provide a conclusion for the Age component from `model2b_alt` in context. Leave the asterisks around each of your answers.

We have **Strong** (weak/moderate/strong) evidence **against** (for/against) the null hypothesis that there is no linear relationship between **Age** (variable name) and **BodyFat** (variable name), after accounting for **logBAI** (variable name) (t(**3197**) = **9.60**, p-value = **< 0.0001**). Thus, we conclude **there is a linear relationship between Age and the true mean BodyFat percentage after accounting for the log of BAI** (alternative in context).

8) These models have one potentially large issue - they fail to account for differences by Gender. Examine the code for the scatterplot below. Give the plot an appropriate title and axis labels to describe what is being plotted. Then provide a discussion of what is being plotted and what this tells us about the Gender variable.

```
scatterplot(residuals(model2a_alt) ~ fitted(model2a_alt)|Gender, data=
Fuster_data1,
            col=viridis(3)[-3], xlab="Fitted BodyFat Percentage",
ylab="Residuals", main="Plot of Residuals vs Fitted by Gender",
            legend=c(title="Gender", coords="bottomright"), smooth=F)
```



The plot above shows the residuals vs fitted BodyFat Percentages by Gender. It shows that the model overestimates the BodyFat Percentage for Males and underestimates the BodyFat Percentage for Females. The Male group is generally higher than the Female group.

9) It seems that we are missing something related to gender in these models. Below is code that will fit an interaction model. Replace the ? with **either** `logBAI` or `logBMI`, you may choose which (there is not a right or wrong choice). Fit the interaction model and write out the estimated model, defining all variables.

```
modelint<- lm(BodyFat ~ logBAI*Gender, data= Fuster_data1)
summary(modelint)
```

```
##
## Call:
## lm(formula = BodyFat ~ logBAI * Gender, data = Fuster_data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.360  -3.149   0.259   3.523  22.255
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -86.0817     3.0061 -28.636  < 2e-16
## logBAI                33.5511     0.9178  36.556  < 2e-16
## GenderFemale          15.3749     3.9077   3.935 8.51e-05
## logBAI:GenderFemale   -3.4617     1.1749  -2.946  0.00324
##
## Residual standard error: 5.135 on 3196 degrees of freedom
## Multiple R-squared:  0.6048, Adjusted R-squared:  0.6045
## F-statistic:  1631 on 3 and 3196 DF,  p-value: < 2.2e-16
```

$\widehat{BodyFat} = -86.08 + 33.55(logBAI) + 15.37(I_F) - 3.46(I_F)(logBAI)$

$I_F = \begin{cases} 1 & \text{if Female} \\ 0 & \text{else} \end{cases}$

10) Simplify the estimated interaction model you just wrote out down to the estimated models for male and female subjects (you should have two simplified models). Show your work.

$\widehat{BodyFatMale} = -86.08 + 33.55(logBAI) + 15.37(0) - 3.46(0)(logBAI)$

$\widehat{BodyFatMale} = -86.08 + 33.55(logBAI)$

$\widehat{BodyFatFemale} = -86.08 + 33.55(logBAI) + 15.37(1) - 3.46(1)(logBAI)$

$\widehat{BodyFatFemale} = -70.71 + 30.09(logBAI)$

11) True or false: The test of the interaction term suggests that the interaction is not needed in the model and we should fit the additive model instead.

False

12) Multiple choice (select the best answer): Something that the authors never seemed to try was to incorporate both BAI and BMI into the same model. Fit a model for `BodyFat`

that incorporates `logBAI`, `logBMI`, `Age`, and `Gender` (no interactions). Find the VIFs for this model and then select the best answer from the multiple choice options below.

```
fullModel <- lm(BodyFat ~logBAI + logBMI + Age + Gender, data=
Fuster_data1)
vif(fullModel)

##   logBAI   logBMI      Age   Gender
## 3.521607 3.376341 1.106238 2.396151
```

A) There are some concerns about multicollinearity in this model because the VIFs for two of the variables are over 3 and another is over 2.

B) There are no concerns about multicollinearity in this model because the VIFs for two of the variables are over 3 and another is over 2.

C) There is no multicollinearity in this model because none of the VIFs are over 5.

D) There is no multicollinearity in this model because the p-values are all small.

E) There is clear multicollinearity in this model because of the overall F-test has a small p-value.

A

13) Please list anyone that you worked on any part of this Project with. For tutors, give their name and affiliation (MLC, SmartyCats, etc.). For fellow students, give their name and section number. If you did not utilize any outside resources, simply enter "NA."

NA