My grades for **In-Class Lab 2**

Q1

1 / 1

Please upload your group's PDF lab report here.

Module 2 In-Class Lab #2

due Thursday September 30 by 11:59PM ET

General Instructions

1. Take the first 5 minutes to introduce yourselves, telling everyone your name, your degree program, and one thing you're excited to do on the weekend.

2. Decide amongst yourselves on the role that each student will perform and add the names to the role below:
   - Timekeeper: Wei-Han Wang
   - Submission Manager: Erica Zhou
   - Live Coder: Emi Yasuda
   - Moderator: Yiwen Gao
   - (if needed) Help finder: ADD NAME HERE

3. Get set up for conducting your roles by making sure everyone can see the shared screen (shared by the LIVE CODER) and that everyone knows who will be contributing verbally and/or by chat and can see/hear each other (MODERATOR should keep track).
   - The SUBMISSION MANAGER should use this time to access the Crowdmark link and create a group submission link by adding the group members to the group.
4. To get help at any time, anyone in the group (or the HELP FINDER if there is one) can tag the instructor by typing @Katherine Daignault in the chat with a question. The TIMEKEEPER should keep track of how long the group spends on each part so that the group will be able to finish the lab during class time.

Submission Instructions

All students will receive an email from Crowdmark which will be used to submit the knitted PDF you produce in your group. YOU WILL NEED TO CREATE YOUR GROUP BEFORE SUBMITTING. To do this:

1. The SUBMISSION MANAGER on the team should use the emailed link to access the assignment page on Crowdmark.

2. There will be an option to add group members to the submission.

3. Using the names you've entered for the group roles above, search for your teammates and add them to your group.

4. All teammembers will receive an email from Crowdmark stating they have been added to the group.

5. At the end of the lab (or before the submission deadline), the SUBMISSION MANAGER should upload the PDF you create from this document to Crowdmark using the group submission link that was created. This will submit the lab for everyone :)

1

Lab 2 - Conditional Nature of Regression

Summary

In this lab, we will be working on two different aspects of regression models. First, we will look more deeply into the conditional interpretation of the regression coefficients by looking into subsets of a dataset. Second, we will practice interpreting a parameter of a regression model by trying to write out the interpretation using extremely simple words.

The dataset for this lab are various measurements taken on 250 men to better understand how Percent Body Fat is related to physical attributes of an individual. While the dataset contains many variables, we will only be working with:

- Pct.BF = percent body fat (Y)
- Waist = waist size of the individual (measured in inches)
- Height = height of the individual (measured in inches)

Part 1: Illustrating the Conditional Nature of Multiple Regression

Step 1 - Load your data and perform an exploratory data analysis.

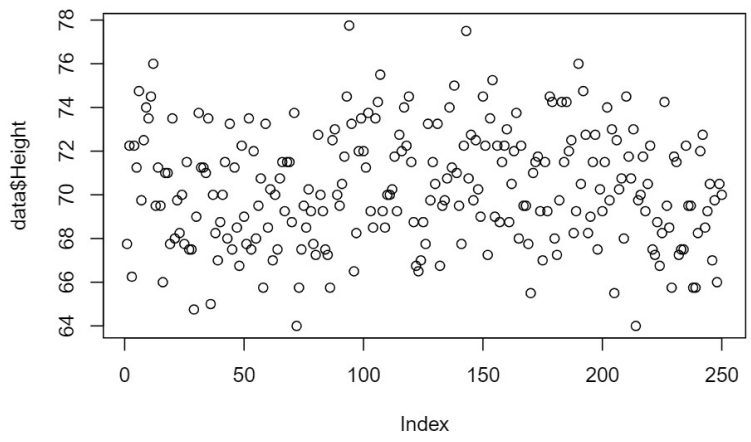
In this step, you'll want 1) to read in the body_fat_complete.csv file, 2) find numerical summaries to describe the centre and spread of these variables, 3) plot the 3 variables listed above, and 4) make two scatterplots (1 between Pct.BF and waist, and 1 between Pct.BF and height).

```
# load you data and find numerical summaries of the variables
data <- read.csv('body_fat_complete.csv', header = T)
summary(data)
```

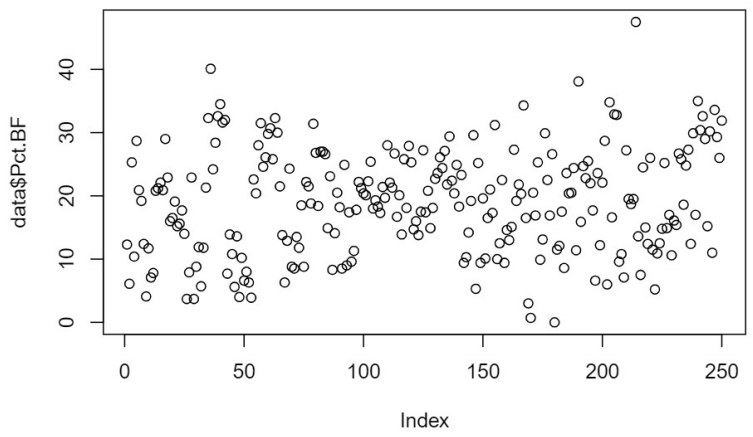
```
##      Pct.BF           Age            Weight          Height     
##  Min.   : 0.00   Min.   :22.00   Min.   :118.5   Min.   :64.00  
##  1st Qu.:12.43   1st Qu.:35.25   1st Qu.:158.5   1st Qu.:68.25  
##  Median :19.20   Median :43.00   Median :176.1   Median :70.00  
##  Mean   :19.03   Mean   :44.88   Mean   :178.1   Mean   :70.30  
##  3rd Qu.:25.20   3rd Qu.:54.00   3rd Qu.:196.8   3rd Qu.:72.25  
##  Max.   :47.50   Max.   :81.00   Max.   :262.8   Max.   :77.75  
##       Neck            Chest           Waist            Hip       
##  Min.   :31.10   Min.   : 79.30   Min.   :27.32   Min.   : 85.00  
##  1st Qu.:36.40   1st Qu.: 94.25   1st Qu.:33.28   1st Qu.: 95.50  
##  Median :38.00   Median : 99.60   Median :35.79   Median : 99.30  
##  Mean   :37.94   Mean   :100.66   Mean   :36.33   Mean   : 99.65  
##  3rd Qu.:39.40   3rd Qu.:105.30   3rd Qu.:39.05   3rd Qu.:103.17  
##  Max.   :43.90   Max.   :128.30   Max.   :49.69   Max.   :125.60  
##      Thigh            Knee            Ankle           Bicep      
##  Min.   :47.20   Min.   :33.00   Min.   :19.10   Min.   :24.80  
##  1st Qu.:56.00   1st Qu.:36.92   1st Qu.:22.00   1st Qu.:30.20  
##  Median :58.95   Median :38.45   Median :22.80   Median :32.00  
##  Mean   :59.25   Mean   :38.53   Mean   :23.07   Mean   :32.22  
##  3rd Qu.:62.25   3rd Qu.:39.88   3rd Qu.:24.00   3rd Qu.:34.30  
##  Max.   :74.40   Max.   :46.00   Max.   :33.90   Max.   :39.10  
##     Forearm          Wrist      
##  Min.   :21.00   Min.   :15.80  
##  1st Qu.:27.30   1st Qu.:17.60  
##  Median :28.70   Median :18.30  
##  Mean   :28.66   Mean   :18.22  
##  3rd Qu.:30.00   3rd Qu.:18.80
```

```
##  Max.   :34.90   Max.   :21.40
```

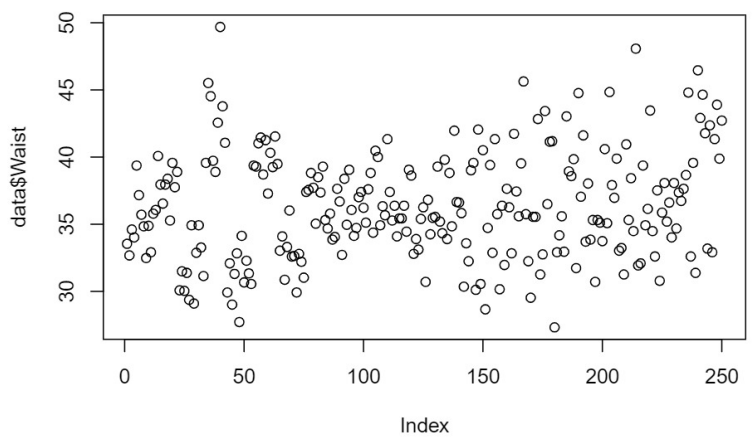
```
# make your plots here
# try to arrange them in grids that group similar plots together
# par(mfrow=c(2,2))  # would create a 2x2 grid of plots if using base R plot functions
plot(data$Height)
```



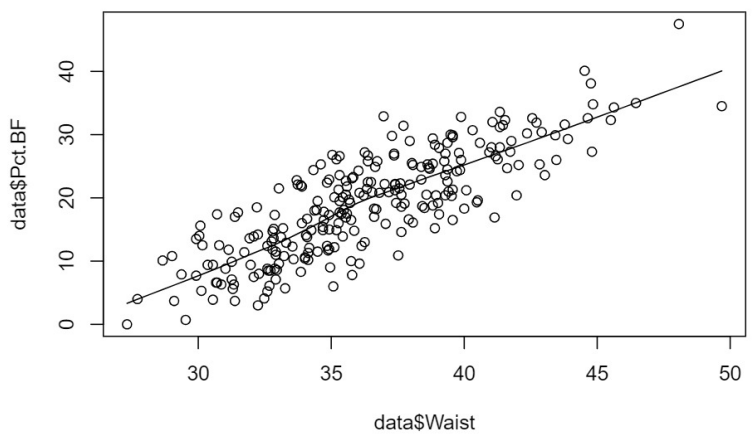

```
plot(data$Pct.BF)
```



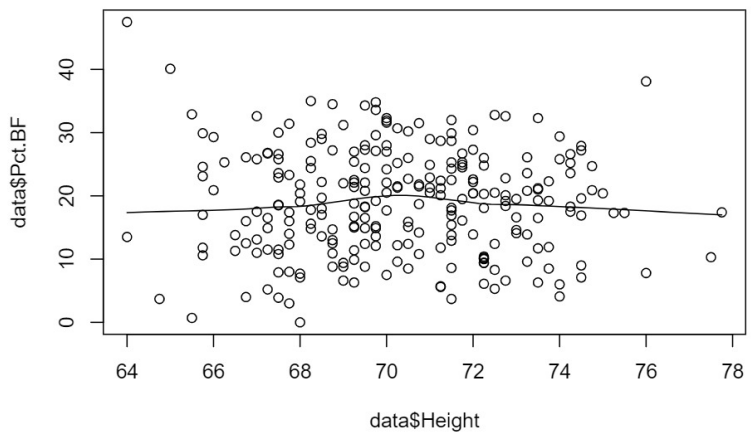

```
plot(data$Waist)
```




```
scatter.smooth(data$Waist, data$Pct.BF)
```




```
scatter.smooth(data$Height, data$Pct.BF)
```



What do you see in the univariate plots of each variable? What can you say about each predictor's relationship with the response, Percent Body Fat? They are all randomly distributed. The waitst and the percent body fat have a strong positive linear relationship. There's no relationship between height and percent body fat.

Step 2 - Fit some linear regression models

In this step, we will look at some regression relationships. You should fit 3 different linear regression models to Percent Body Fat:

- a simple linear model using Height as a predictor
- a simple linear model using Waist as a predictor
- a two-predictor model using both Height and Waist as predictors.

```
# fit your first linear model here
lm(Pct.BF ~ Height, data=data)
```

```
##
## Call:
## lm(formula = Pct.BF ~ Height, data = data)
##
## Coefficients:
## (Intercept)       Height
##    25.58078     -0.09316
```

```
# fit your second linear model here
lm(Pct.BF ~ Waist, data=data)
```

```
##
## Call:
```

```
## lm(formula = Pct.BF ~ Waist, data = data)
##
## Coefficients:
## (Intercept)        Waist
##      -42.73         1.70
```

```
# fit your two predictor model here
lm(Pct.BF ~ Waist + Height, data=data)
```

```
##
## Call:
## lm(formula = Pct.BF ~ Waist + Height, data = data)
##
## Coefficients:
## (Intercept)        Waist       Height
##     -3.1009       1.7731      -0.6015
```

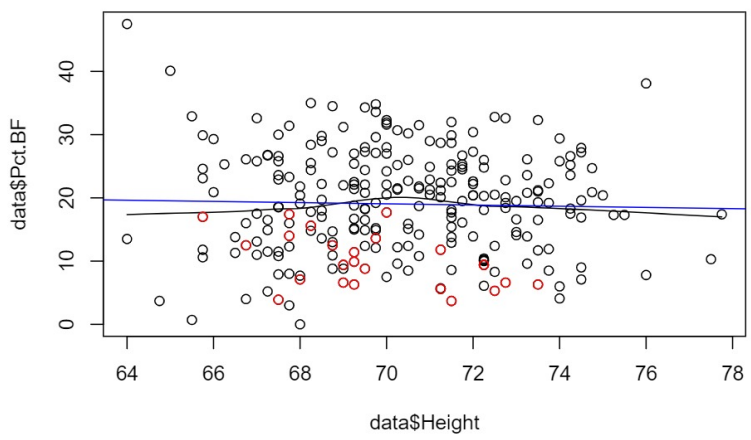
What do you notice to be different between the coefficients of the models?

The percent body fat and height have a negative relatinoship, and the waist and the percent body fat have a positive slope. However, when you fit a two-predictor model, there appeares a positive relationship.

Step 3 - Looking at simple relationships among a subset

This step will let us see why regression has this conditional aspect to it when we include two or predictors in a model. First you will need to plot a scatterplot of Percent Body Fat versus Height, adding a line representing the simple regression relationship from step 2. Then, get a subset of your dataset that correspond to only individuals with a Waist size between 30-32 (inclusive). Add these points to your plot in a different color.

```
# make your scatterplot with the added regression line
# if using base R to plot, you can add the line using
# abline(a = intercept, b = slope, col="blue")
subset <- data[which(data$Waist >= 30 & data$Waist <= 32), ]
scatter.smooth(data$Height, data$Pct.BF) +
  abline(a = 25.58078, b = -0.09316, col="blue") +
  points(subset$Height, subset$Pct.BF, col='red')
```




```
## integer(0)
```

```
# to get the subset of individuals with these waist sizes, you may use an if statement
# you may also try the which() function, used as below, or another way if you know one

# by using the data in subset, add points in color to your plot above (see lab 1)
```

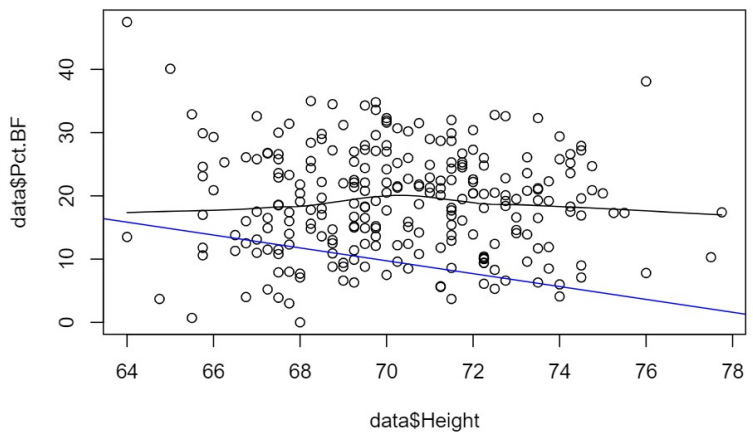
Now, let's fit the relationship between Percent Body Fat and Height among only men with waist sizes between 30 and 32 inches.

```
# fit your model here
lm(Pct.BF ~ Height, data=subset)
```

```
##
## Call:
## lm(formula = Pct.BF ~ Height, data = subset)
##
## Coefficients:
## (Intercept)       Height
##      81.073       -1.019
```

Copy-paste your plot code from the beginning of this step, and add this regression relationship to your plot using the same colour as your points.

```
# copy paste your code from above
scatter.smooth(data$Height, data$Pct.BF) +
  abline(a = 81.073, b = -1.019, col="blue")
```




```
## integer(0)
```

```
# add a line with your regression relationship on the subset
# if using base R to plot, you can do this using abline(a = intercept, b = slope, col="blue")
```

(OPTIONAL) Repeat this process by creating a new subset of your data for men with waist sizes between 40-42 inches (inclusive) and fitting a simple linear relationship on this subset between Pct. BF and Height. Copy-paste your plot code from the previous code chunk, highlight these new points in a different colour (in addition to the ones before) and add a line representing this relationship in a colour to match your points.

```
# OPTIONAL PART
# copy-paste plot code from previous chunk

# create the new subset and fit the linear model

# update plot with new coloured points and line for this subset
```

Using these plots to guide your thinking, explain why the simple linear model using height as a predictor showed a very flat slope, but the 2-predictor model estimated a much stronger decreasing relationship for Height.

ADD COMMENTS HERE

Part 2: Interpreting your slope in extreme simple language

The only task for this part is to write out a correct interpretation of the coefficient for Height in the two-predictor model you fit in Part 1, step 2, but using only extremely simple words. Using this XKCD word

checker https://xkcd.com/simplewriter/, you will write out an interpretation for this coefficient in the text box at this site. Words that are not extremely simple will be highlighted in red. You should attempt to write out a complete and correct interpretation of this coefficient with no red words (i.e. using extremely simple language). Once finished, copy your interpretation from the XKCD site to here.

ADD XKCD INTERPRETATION HERE