Group Project 1: Body Fat Modeling

Catherine Zheng, Peter Kryspin, Jon Starfeldt

10/10/2021

# Step 1: Analyzing Raw Data

BodyFat=read.csv("Bodyfat.csv")  
attach(BodyFat)  
library(car)

## Loading required package: carData

dim(BodyFat)

## [1] 252 17

252 means that there are 252 rows (men) in this dataset, 17 means that there are 17 parameters measured.

colnames(BodyFat)

## [1] "IDNO" "BODYFAT" "DENSITY" "AGE" "WEIGHT" "HEIGHT"   
## [7] "ADIPOSITY" "NECK" "CHEST" "ABDOMEN" "HIP" "THIGH"   
## [13] "KNEE" "ANKLE" "BICEPS" "FOREARM" "WRIST"

head(BodyFat)

## IDNO BODYFAT DENSITY AGE WEIGHT HEIGHT ADIPOSITY NECK CHEST ABDOMEN HIP  
## 1 1 12.6 1.0708 23 154.25 67.75 23.7 36.2 93.1 85.2 94.5  
## 2 2 6.9 1.0853 22 173.25 72.25 23.4 38.5 93.6 83.0 98.7  
## 3 3 24.6 1.0414 22 154.00 66.25 24.7 34.0 95.8 87.9 99.2  
## 4 4 10.9 1.0751 26 184.75 72.25 24.9 37.4 101.8 86.4 101.2  
## 5 5 27.8 1.0340 24 184.25 71.25 25.6 34.4 97.3 100.0 101.9  
## 6 6 20.6 1.0502 24 210.25 74.75 26.5 39.0 104.5 94.4 107.8  
## THIGH KNEE ANKLE BICEPS FOREARM WRIST  
## 1 59.0 37.3 21.9 32.0 27.4 17.1  
## 2 58.7 37.3 23.4 30.5 28.9 18.2  
## 3 59.6 38.9 24.0 28.8 25.2 16.6  
## 4 60.1 37.3 22.8 32.4 29.4 18.2  
## 5 63.2 42.2 24.0 32.2 27.7 17.7  
## 6 66.0 42.0 25.6 35.7 30.6 18.8

tail(BodyFat)

## IDNO BODYFAT DENSITY AGE WEIGHT HEIGHT ADIPOSITY NECK CHEST ABDOMEN HIP  
## 247 247 29.1 1.0308 69 215.50 70.50 30.5 40.8 113.7 107.6 110.0  
## 248 248 11.5 1.0736 70 134.25 67.00 21.1 34.9 89.2 83.6 88.8  
## 249 249 32.3 1.0236 72 201.00 69.75 29.1 40.9 108.5 105.0 104.5  
## 250 250 28.3 1.0328 72 186.75 66.00 30.2 38.9 111.1 111.5 101.7  
## 251 251 25.3 1.0399 72 190.75 70.50 27.0 38.9 108.3 101.3 97.8  
## 252 252 30.7 1.0271 74 207.50 70.00 29.8 40.8 112.4 108.5 107.1  
## THIGH KNEE ANKLE BICEPS FOREARM WRIST  
## 247 63.3 44.0 22.6 37.5 32.6 18.8  
## 248 49.6 34.8 21.5 25.6 25.7 18.5  
## 249 59.6 40.8 23.2 35.2 28.6 20.1  
## 250 60.3 37.3 21.5 31.3 27.2 18.0  
## 251 56.0 41.6 22.7 30.5 29.4 19.8  
## 252 59.3 42.2 24.6 33.7 30.0 20.9

summary(BodyFat)

## IDNO BODYFAT DENSITY AGE   
## Min. : 1.00 Min. : 0.00 Min. :0.995 Min. :22.00   
## 1st Qu.: 63.75 1st Qu.:12.80 1st Qu.:1.041 1st Qu.:35.75   
## Median :126.50 Median :19.00 Median :1.055 Median :43.00   
## Mean :126.50 Mean :18.94 Mean :1.056 Mean :44.88   
## 3rd Qu.:189.25 3rd Qu.:24.60 3rd Qu.:1.070 3rd Qu.:54.00   
## Max. :252.00 Max. :45.10 Max. :1.109 Max. :81.00   
## WEIGHT HEIGHT ADIPOSITY NECK   
## Min. :118.5 Min. :29.50 Min. :18.10 Min. :31.10   
## 1st Qu.:159.0 1st Qu.:68.25 1st Qu.:23.10 1st Qu.:36.40   
## Median :176.5 Median :70.00 Median :25.05 Median :38.00   
## Mean :178.9 Mean :70.15 Mean :25.44 Mean :37.99   
## 3rd Qu.:197.0 3rd Qu.:72.25 3rd Qu.:27.32 3rd Qu.:39.42   
## Max. :363.1 Max. :77.75 Max. :48.90 Max. :51.20   
## CHEST ABDOMEN HIP THIGH   
## Min. : 79.30 Min. : 69.40 Min. : 85.0 Min. :47.20   
## 1st Qu.: 94.35 1st Qu.: 84.58 1st Qu.: 95.5 1st Qu.:56.00   
## Median : 99.65 Median : 90.95 Median : 99.3 Median :59.00   
## Mean :100.82 Mean : 92.56 Mean : 99.9 Mean :59.41   
## 3rd Qu.:105.38 3rd Qu.: 99.33 3rd Qu.:103.5 3rd Qu.:62.35   
## Max. :136.20 Max. :148.10 Max. :147.7 Max. :87.30   
## KNEE ANKLE BICEPS FOREARM WRIST   
## Min. :33.00 Min. :19.1 Min. :24.80 Min. :21.00 Min. :15.80   
## 1st Qu.:36.98 1st Qu.:22.0 1st Qu.:30.20 1st Qu.:27.30 1st Qu.:17.60   
## Median :38.50 Median :22.8 Median :32.05 Median :28.70 Median :18.30   
## Mean :38.59 Mean :23.1 Mean :32.27 Mean :28.66 Mean :18.23   
## 3rd Qu.:39.92 3rd Qu.:24.0 3rd Qu.:34.33 3rd Qu.:30.00 3rd Qu.:18.80   
## Max. :49.10 Max. :33.9 Max. :45.00 Max. :34.90 Max. :21.40

Some findings:  
1. The first column is just the index number of each individual  
2.The age of the participants are mainly around 40  
3. In height, weight, abdomen circumference, adiposity, and age, they have relatively larger range than other measurements  
4. Someone has a Bodyfat of 0 %. This might be an incorrect value.

##Take a look at the first ten rows##  
BodyFat[1:10,]

## IDNO BODYFAT DENSITY AGE WEIGHT HEIGHT ADIPOSITY NECK CHEST ABDOMEN HIP  
## 1 1 12.6 1.0708 23 154.25 67.75 23.7 36.2 93.1 85.2 94.5  
## 2 2 6.9 1.0853 22 173.25 72.25 23.4 38.5 93.6 83.0 98.7  
## 3 3 24.6 1.0414 22 154.00 66.25 24.7 34.0 95.8 87.9 99.2  
## 4 4 10.9 1.0751 26 184.75 72.25 24.9 37.4 101.8 86.4 101.2  
## 5 5 27.8 1.0340 24 184.25 71.25 25.6 34.4 97.3 100.0 101.9  
## 6 6 20.6 1.0502 24 210.25 74.75 26.5 39.0 104.5 94.4 107.8  
## 7 7 19.0 1.0549 26 181.00 69.75 26.2 36.4 105.1 90.7 100.3  
## 8 8 12.8 1.0704 25 176.00 72.50 23.6 37.8 99.6 88.5 97.1  
## 9 9 5.1 1.0900 25 191.00 74.00 24.6 38.1 100.9 82.5 99.9  
## 10 10 12.0 1.0722 23 198.25 73.50 25.8 42.1 99.6 88.6 104.1  
## THIGH KNEE ANKLE BICEPS FOREARM WRIST  
## 1 59.0 37.3 21.9 32.0 27.4 17.1  
## 2 58.7 37.3 23.4 30.5 28.9 18.2  
## 3 59.6 38.9 24.0 28.8 25.2 16.6  
## 4 60.1 37.3 22.8 32.4 29.4 18.2  
## 5 63.2 42.2 24.0 32.2 27.7 17.7  
## 6 66.0 42.0 25.6 35.7 30.6 18.8  
## 7 58.4 38.3 22.9 31.9 27.8 17.7  
## 8 60.0 39.4 23.2 30.5 29.0 18.8  
## 9 62.9 38.3 23.8 35.9 31.1 18.2  
## 10 63.1 41.7 25.0 35.6 30.0 19.2

##Take a random look at 67th,135th,and 220th individual##  
BodyFat[c(67,135,220),]

## IDNO BODYFAT DENSITY AGE WEIGHT HEIGHT ADIPOSITY NECK CHEST ABDOMEN HIP  
## 67 67 21.1 1.0499 54 151.50 70.75 21.3 35.6 90.0 83.9 93.9  
## 135 135 23.8 1.0435 41 168.25 69.50 24.5 36.5 98.4 87.2 98.4  
## 220 220 15.1 1.0646 53 154.50 69.25 22.7 37.6 93.9 88.7 94.5  
## THIGH KNEE ANKLE BICEPS FOREARM WRIST  
## 67 55.0 36.1 21.7 29.6 27.4 17.4  
## 135 56.0 36.9 23.0 34.0 29.8 18.1  
## 220 53.7 36.2 22.0 28.5 25.7 17.1

We noticed that 135th and 220th have similar abdomen circumference(87.2 and 88.7), but their body fat are quite different (23.8 and 15.1), we think this may be influenced by their age (41 and 53).

Since we noticed that the range of age, abdomen, height, and weight are relatively large.  
Then we take a subset of the first ten men and look at body fat, age, abdomen, height,weight,and adiposity.

BodyFat[1:10,c(2,4,5,6,7,10)]

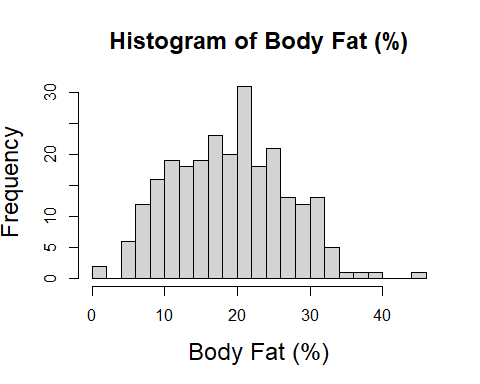
## BODYFAT AGE WEIGHT HEIGHT ADIPOSITY ABDOMEN  
## 1 12.6 23 154.25 67.75 23.7 85.2  
## 2 6.9 22 173.25 72.25 23.4 83.0  
## 3 24.6 22 154.00 66.25 24.7 87.9  
## 4 10.9 26 184.75 72.25 24.9 86.4  
## 5 27.8 24 184.25 71.25 25.6 100.0  
## 6 20.6 24 210.25 74.75 26.5 94.4  
## 7 19.0 26 181.00 69.75 26.2 90.7  
## 8 12.8 25 176.00 72.50 23.6 88.5  
## 9 5.1 25 191.00 74.00 24.6 82.5  
## 10 12.0 23 198.25 73.50 25.8 88.6

From some of the background readings we read, we noticed that with the increase of male’s body fat, their abdomen circumference increase a lot.

# Step2:Visualizing Data

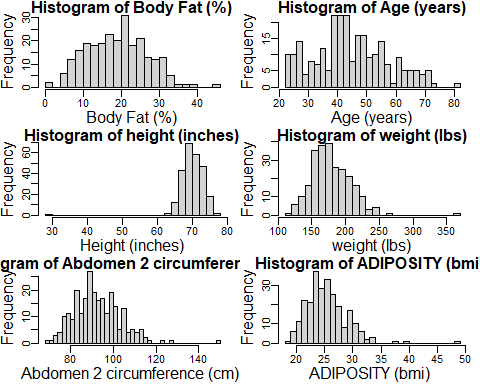
We started with histograms of BodyFat %

hist(BODYFAT,breaks=30,cex.lab=1.5,cex.main=1.5,  
 main="Histogram of Body Fat (%)",xlab="Body Fat (%)")



Then we created multiple plots contain Body Fat %, Age, Height, Weight, and Abdomen 2 circumference

par(mfrow=c(3,2))  
par(mgp=c(1.8,.5,0), mar=c(3,3,1,1))  
hist(BODYFAT, breaks=30,cex.lab=1.5,cex.main=1.5,  
 main="Histogram of Body Fat (%)", xlab="Body Fat (%)")  
hist(AGE, breaks=30,cex.lab=1.5,cex.main=1.5,  
 main="Histogram of Age (years)", xlab="Age (years)")  
hist(HEIGHT, breaks=30,cex.lab=1.5,cex.main=1.5,  
 main="Histogram of height (inches)", xlab="Height (inches)")  
hist(WEIGHT, breaks=30,cex.lab=1.5,cex.main=1.5,  
 main="Histogram of weight (lbs)", xlab="weight (lbs)")  
hist(ABDOMEN, breaks=30,cex.lab=1.5,cex.main=1.5,  
 main="Histogram of Abdomen 2 circumference (cm) ", xlab="Abdomen 2 circumference (cm) ")  
hist(ADIPOSITY, breaks=30,cex.lab=1.5,cex.main=1.5,  
 main="Histogram of ADIPOSITY (bmi) ", xlab="ADIPOSITY (bmi) ")



It seems like there is one person that has the highest weight, abdomen circumference, and adiposity. So we were thinking that this person could be an outlier.

## So we used some subsets to find the possible outlier##  
BodyFat[WEIGHT > 350,]

## IDNO BODYFAT DENSITY AGE WEIGHT HEIGHT ADIPOSITY NECK CHEST ABDOMEN HIP  
## 39 39 33.8 1.0202 46 363.15 72.25 48.9 51.2 136.2 148.1 147.7  
## THIGH KNEE ANKLE BICEPS FOREARM WRIST  
## 39 87.3 49.1 29.6 45 29 21.4

## For the Abdomen subset##  
BodyFat[ABDOMEN > 140,]

## IDNO BODYFAT DENSITY AGE WEIGHT HEIGHT ADIPOSITY NECK CHEST ABDOMEN HIP  
## 39 39 33.8 1.0202 46 363.15 72.25 48.9 51.2 136.2 148.1 147.7  
## THIGH KNEE ANKLE BICEPS FOREARM WRIST  
## 39 87.3 49.1 29.6 45 29 21.4

## For the Adiposity##  
BodyFat[ADIPOSITY > 45,]

## IDNO BODYFAT DENSITY AGE WEIGHT HEIGHT ADIPOSITY NECK CHEST ABDOMEN HIP  
## 39 39 33.8 1.0202 46 363.15 72.25 48.9 51.2 136.2 148.1 147.7  
## THIGH KNEE ANKLE BICEPS FOREARM WRIST  
## 39 87.3 49.1 29.6 45 29 21.4

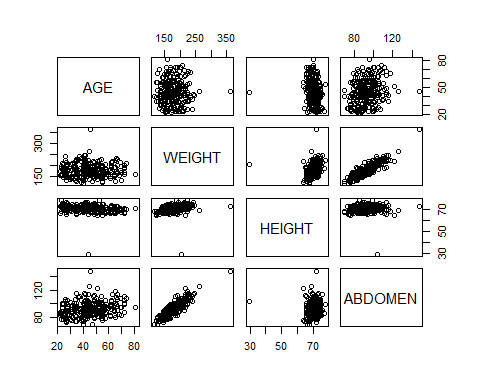
Since subject 39 has such a large weight, abdomen, and adiposity, we think he could be an outlier and we will determine that in leverage values and cook’s distance.

Then we try to show that age, weight, height, and abdomen could have linear relationship with age, weight, height,abdomen

specific = BodyFat[c(4,5,6,10)]  
summary(specific)

## AGE WEIGHT HEIGHT ABDOMEN   
## Min. :22.00 Min. :118.5 Min. :29.50 Min. : 69.40   
## 1st Qu.:35.75 1st Qu.:159.0 1st Qu.:68.25 1st Qu.: 84.58   
## Median :43.00 Median :176.5 Median :70.00 Median : 90.95   
## Mean :44.88 Mean :178.9 Mean :70.15 Mean : 92.56   
## 3rd Qu.:54.00 3rd Qu.:197.0 3rd Qu.:72.25 3rd Qu.: 99.33   
## Max. :81.00 Max. :363.1 Max. :77.75 Max. :148.10

pairs(specific)

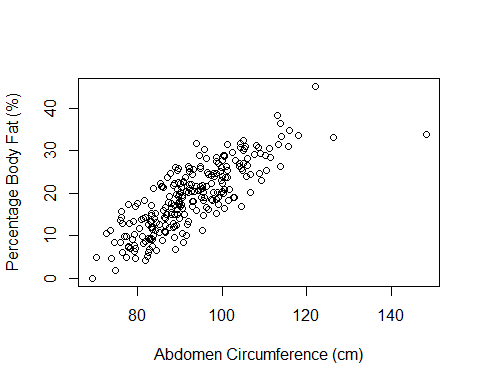
 We observe a positive linear relationship between body fat and abdomen circumference as well as body fat and weight. The body fat versus height scatter plot may be linear, however an obvious slope unequal to zero must be observed in order to suggest a linear relationship.

# Step3: Statistical Modeling and Analysis

From the description of data and our background readings, we noticed that adiposity is just bmi, which is calculated by weight/(height)^2  
After the first two steps, we decided our predictors to be: age, adiposity,and abdomen.

Plot of Body Fat by Abdomen Circumference and summary statistics

plot(ABDOMEN, BODYFAT, xlab = "Abdomen Circumference (cm)", ylab = "Percentage Body Fat (%)")

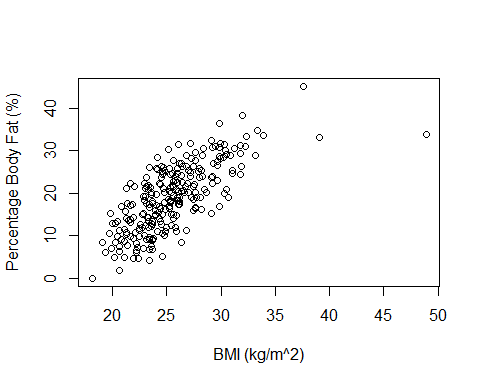


AbLinearModel = lm(BODYFAT~ABDOMEN)  
summary(AbLinearModel)

##   
## Call:  
## lm(formula = BODYFAT ~ ABDOMEN)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.6257 -3.4672 0.0111 3.1415 11.9754   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -35.19661 2.46229 -14.29 <2e-16 \*\*\*  
## ABDOMEN 0.58489 0.02643 22.13 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.514 on 250 degrees of freedom  
## Multiple R-squared: 0.6621, Adjusted R-squared: 0.6608   
## F-statistic: 489.9 on 1 and 250 DF, p-value: < 2.2e-16

Plot of Body Fat by Adiposity (BMI) and summary statistics

plot(BodyFat$ADIPOSITY, BodyFat$BODYFAT, xlab = "BMI (kg/m^2)", ylab = "Percentage Body Fat (%)")

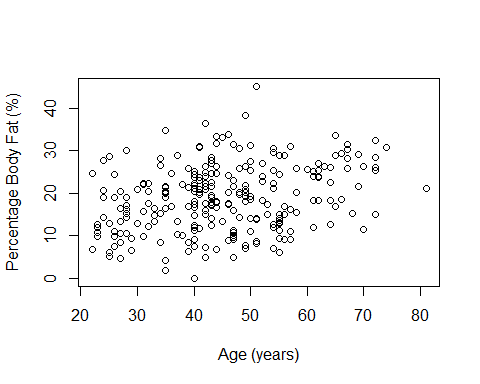


BMILinearModel = lm(BodyFat$BODYFAT~BodyFat$ADIPOSITY)  
summary(BMILinearModel)

##   
## Call:  
## lm(formula = BodyFat$BODYFAT ~ BodyFat$ADIPOSITY)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.4292 -3.4478 0.2113 3.8663 11.7826   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -20.40508 2.36723 -8.62 7.78e-16 \*\*\*  
## BodyFat$ADIPOSITY 1.54671 0.09212 16.79 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.324 on 250 degrees of freedom  
## Multiple R-squared: 0.53, Adjusted R-squared: 0.5281   
## F-statistic: 281.9 on 1 and 250 DF, p-value: < 2.2e-16

Plot of Body Fat by Age and summary statistics

plot(BodyFat$AGE, BodyFat$BODYFAT, xlab = "Age (years)", ylab = "Percentage Body Fat (%)")



AgeLinearModel = lm(BODYFAT~AGE)  
summary(AgeLinearModel)

##   
## Call:  
## lm(formula = BODYFAT ~ AGE)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.0697 -5.7025 0.2846 4.8301 25.0739   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.95546 1.73576 6.312 1.25e-09 \*\*\*  
## AGE 0.17786 0.03724 4.776 3.04e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.435 on 250 degrees of freedom  
## Multiple R-squared: 0.08362, Adjusted R-squared: 0.07996   
## F-statistic: 22.81 on 1 and 250 DF, p-value: 3.045e-06

From the above three SLR models, we think ADPOSITY,AGE,and ABDOMEN have a linear relationship with body fat.

guessmodel = lm(BODYFAT ~ ABDOMEN + ADIPOSITY + AGE)  
summary(guessmodel)

##   
## Call:  
## lm(formula = BODYFAT ~ ABDOMEN + ADIPOSITY + AGE)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.5026 -3.1891 0.1558 3.4636 12.3348   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -37.13370 2.54301 -14.602 <2e-16 \*\*\*  
## ABDOMEN 0.63498 0.07177 8.847 <2e-16 \*\*\*  
## ADIPOSITY -0.21192 0.20790 -1.019 0.3090   
## AGE 0.05996 0.02367 2.533 0.0119 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.449 on 248 degrees of freedom  
## Multiple R-squared: 0.6744, Adjusted R-squared: 0.6705   
## F-statistic: 171.2 on 3 and 248 DF, p-value: < 2.2e-16

According to our guess MLR model, if we want to use Abdomen circumference, Adiposity, and age to predict body fat %, we can use the equation  
Body Fat Pct = -37.13 + 0.635Abdomen Circumference(cm) - 0.212Adiposity + 0.06Age

Now we will run some f-tests to figure out if our variables are significant after accounting for other variables.

1. Test whether one of the three variables abdomen, adiposity, and age is important after accounting for the other two variables. Hypothesis here would be:

summary(guessmodel)

##   
## Call:  
## lm(formula = BODYFAT ~ ABDOMEN + ADIPOSITY + AGE)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.5026 -3.1891 0.1558 3.4636 12.3348   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -37.13370 2.54301 -14.602 <2e-16 \*\*\*  
## ABDOMEN 0.63498 0.07177 8.847 <2e-16 \*\*\*  
## ADIPOSITY -0.21192 0.20790 -1.019 0.3090   
## AGE 0.05996 0.02367 2.533 0.0119 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.449 on 248 degrees of freedom  
## Multiple R-squared: 0.6744, Adjusted R-squared: 0.6705   
## F-statistic: 171.2 on 3 and 248 DF, p-value: < 2.2e-16

From here, we noticed that Abdomen and Age both have p-value less than 0.05, the associated p-value for adiposity is 0.3090, so we fail to reject the null hypothesis that beta3 =0. In this case, we thought that we need to take out Adiposity as our predictors since it’s not so significant.

1. Test whether taking out a variable is significant  
   H0: taking out variable is insignificant;  
   H1: taking out variable is significant  
   i)For Adiposity;

noBMI= lm(BODYFAT~ABDOMEN+AGE,data=BodyFat)  
anova(noBMI,guessmodel)

## Analysis of Variance Table  
##   
## Model 1: BODYFAT ~ ABDOMEN + AGE  
## Model 2: BODYFAT ~ ABDOMEN + ADIPOSITY + AGE  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 249 4930.3   
## 2 248 4909.7 1 20.57 1.039 0.309

ii)For Age:

noAge= lm(BODYFAT~ABDOMEN+ADIPOSITY,data=BodyFat)  
anova(noAge,guessmodel)

## Analysis of Variance Table  
##   
## Model 1: BODYFAT ~ ABDOMEN + ADIPOSITY  
## Model 2: BODYFAT ~ ABDOMEN + ADIPOSITY + AGE  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 249 5036.7   
## 2 248 4909.7 1 127.04 6.4173 0.01192 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

iii)For Abdomen

noAbdomen= lm(BODYFAT~AGE+ADIPOSITY,data=BodyFat)  
anova(noAbdomen,guessmodel)

## Analysis of Variance Table  
##   
## Model 1: BODYFAT ~ AGE + ADIPOSITY  
## Model 2: BODYFAT ~ ABDOMEN + ADIPOSITY + AGE  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 249 6459.4   
## 2 248 4909.7 1 1549.7 78.278 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

We noticed that if we take out Adiposity, the corresponding p-vlue is 0.309, greater than 0.05, meaning that Adiposity is insignificant in out guess model.

In this case, we decided to only use Age and Abdomen in our MLR model as predictors.

reducedModel <- lm(BODYFAT~ABDOMEN+AGE)  
summary(reducedModel)

##   
## Call:  
## lm(formula = BODYFAT ~ ABDOMEN + AGE)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.7114 -3.2622 0.0285 3.2248 12.0577   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -36.51507 2.46972 -14.785 < 2e-16 \*\*\*  
## ABDOMEN 0.56710 0.02677 21.187 < 2e-16 \*\*\*  
## AGE 0.06605 0.02290 2.884 0.00427 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.45 on 249 degrees of freedom  
## Multiple R-squared: 0.673, Adjusted R-squared: 0.6704   
## F-statistic: 256.3 on 2 and 249 DF, p-value: < 2.2e-16

So now we make our guess model to a reduced one.  
According to our reduced model, we can interpret the MLR model as  
Body Fat Pct = -36.52 +0.567 Abdomen + 0.066Age

Interpretation of coefficients:  
1. If we keep Abdomen fixed, we increase age by 1 year, the body fat will increase by 0.066 %  
2. If we keep Age fixed, we increase Abdomen by 1 cm, the body fat will increase by 0.57

fullmodel <- lm(BODYFAT ~AGE+WEIGHT+HEIGHT+ADIPOSITY+NECK+CHEST+ABDOMEN+HIP+THIGH+KNEE+ANKLE+BICEPS+FOREARM+WRIST)  
reducedModel <- lm(BODYFAT~ABDOMEN+AGE)

3.Test whether at least one of abdomen circumference or age is significant in predicting body fat percentage.  
Hypothesis test here would be:

anova(reducedModel,fullmodel)

## Analysis of Variance Table  
##   
## Model 1: BODYFAT ~ ABDOMEN + AGE  
## Model 2: BODYFAT ~ AGE + WEIGHT + HEIGHT + ADIPOSITY + NECK + CHEST +   
## ABDOMEN + HIP + THIGH + KNEE + ANKLE + BICEPS + FOREARM +   
## WRIST  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 249 4930.3   
## 2 237 3784.4 12 1145.9 5.9803 4.212e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Here, we get the p-value is 4.212e-09, so we reject the null hypothesis and conclude that at least one predictor is significant in predicting body fat percentage.

1. Test if each variable is important after accounting for the other

i)For Abdomen

anova(AgeLinearModel,reducedModel)

## Analysis of Variance Table  
##   
## Model 1: BODYFAT ~ AGE  
## Model 2: BODYFAT ~ ABDOMEN + AGE  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 250 13818.1   
## 2 249 4930.3 1 8887.8 448.87 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Here, we get the p-value is < 2.2e-16, so we reject the null hypothesis and conclude that Abdomen is significant in predicting Bodayfat after accounting for Age.

ii)For Age

anova(AbLinearModel,reducedModel)

## Analysis of Variance Table  
##   
## Model 1: BODYFAT ~ ABDOMEN  
## Model 2: BODYFAT ~ ABDOMEN + AGE  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 250 5094.9   
## 2 249 4930.3 1 164.66 8.3163 0.004273 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Here, we get the p-value is 0.004273, so we reject the null hypothesis and conclude that Abdomen is significant in predicting Body fat after accounting for Age.

Now that we have decided on a model, we want to use our model repair the data point with 0% body fat that we observed earlier.

zeroBodyFat = which(BodyFat$BODYFAT == 0)  
##Want to confirm this is the correct data point  
BodyFat[182,]

## IDNO BODYFAT DENSITY AGE WEIGHT HEIGHT ADIPOSITY NECK CHEST ABDOMEN HIP  
## 182 182 0 1.1089 40 118.5 68 18.1 33.8 79.3 69.4 85  
## THIGH KNEE ANKLE BICEPS FOREARM WRIST  
## 182 47.2 33.5 20.2 27.7 24.6 16.5

predictedValue = -36.52+.567\*69.4+.066\*40  
predictedValue

## [1] 5.4698

BodyFat[182,2] = 5.5  
  
##Update our model to include this data point  
reducedModel = lm(BodyFat$BODYFAT~BodyFat$ABDOMEN+BodyFat$AGE)  
summary(reducedModel)

##   
## Call:  
## lm(formula = BodyFat$BODYFAT ~ BodyFat$ABDOMEN + BodyFat$AGE)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.4882 -3.2802 0.0653 3.2106 12.0420   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -36.09328 2.46201 -14.660 < 2e-16 \*\*\*  
## BodyFat$ABDOMEN 0.56269 0.02668 21.088 < 2e-16 \*\*\*  
## BodyFat$AGE 0.06625 0.02283 2.901 0.00405 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.436 on 249 degrees of freedom  
## Multiple R-squared: 0.6712, Adjusted R-squared: 0.6686   
## F-statistic: 254.1 on 2 and 249 DF, p-value: < 2.2e-16

The equation for this updated MLR model is:  
Body Fat Percentage = -36.09 + 0.563Abdomen Circumference(cm) + 0.066Age

confint(reducedModel)

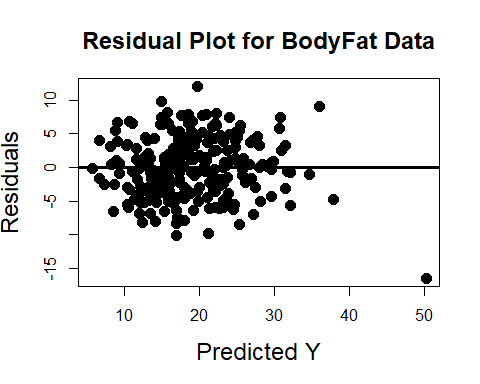
## 2.5 % 97.5 %  
## (Intercept) -40.94229483 -31.2442593  
## BodyFat$ABDOMEN 0.51013407 0.6152423  
## BodyFat$AGE 0.02127766 0.1112147

Our model contains a 5% error rate, so here are shown the boundaries for 95% percent confidence intervals for the true value of the intercept and each of the predictor variables.

# Step4: Diagnostics

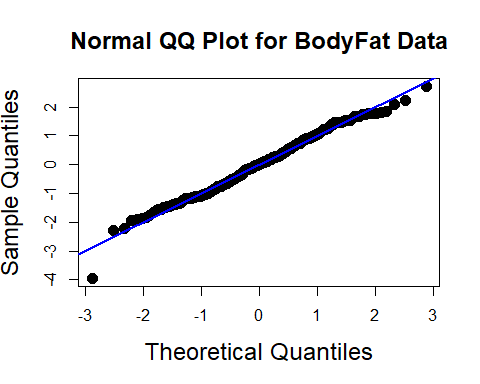
1.Residual Plot

r\_bf =residuals(reducedModel)  
yhat\_bf=predict(reducedModel)  
plot(yhat\_bf,r\_bf,main="Residual Plot for BodyFat Data",  
 xlab="Predicted Y", ylab="Residuals",  
 pch=19,cex=1.5,cex.main=1.5,cex.lab=1.5)  
abline(a=0,b=0,col="black",lwd=3)

 1.We don’t see any non-linear trends in the residual plot  
2.We don’t see any < or > out pattern; more generally, I don’t see the width of the points (in the y-direction) changing as I move across the x-axis.  
As such, I think linearity and homoscedasticity seem reasonable.

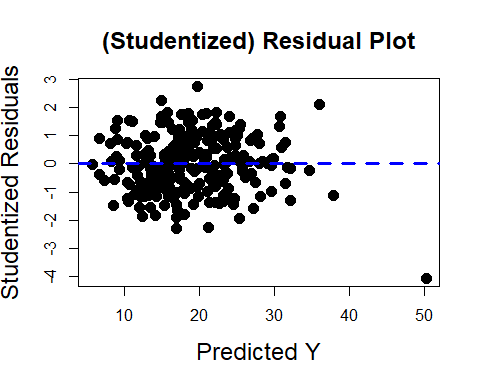
2.QQ Plot

qqnorm(rstandard(reducedModel),main="Normal QQ Plot for BodyFat Data",  
 pch=19,cex=1.5,cex.main=1.5,cex.lab=1.5)  
abline(a=0,b=1,lwd=2,col="blue")

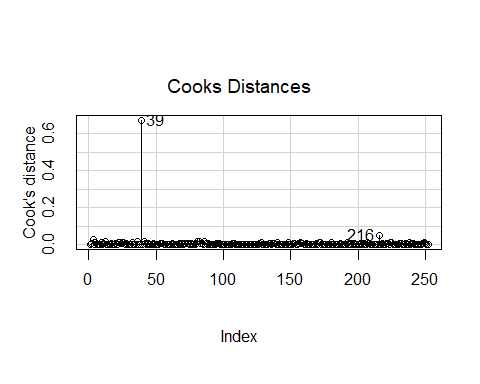
 From the QQ plot, Normality looks reasonable; the points are reasonably close to the 45 degree line.

1. Outliers in BodyFat data

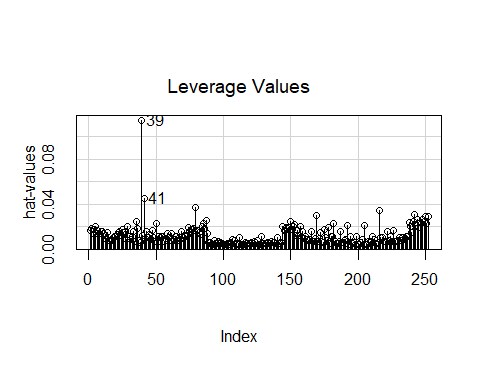
plot(yhat\_bf,rstudent(reducedModel),main="(Studentized) Residual Plot",  
 xlab="Predicted Y", ylab="Studentized Residuals",  
 pch=19,cex=1.5,cex.main=1.5,cex.lab=1.5); abline(h=0,col="blue",lwd=3,lty=2)

 4.leverage and influential points

#Cook's Distance  
infIndexPlot(reducedModel, vars = c("Cook"), main = "Cooks Distances")



#Leverage Values  
infIndexPlot(reducedModel, vars = c("Hat"), main = "Leverage Values")

 Point at index 39 appears to be influential as it is an outlier in both plots so we try removing it from data set.

newBodyFat <- BodyFat[c(1:38, 40:252),]  
newReducedModel <- lm(newBodyFat$BODYFAT~newBodyFat$ABDOMEN+newBodyFat$AGE)  
summary(newReducedModel)

##   
## Call:  
## lm(formula = newBodyFat$BODYFAT ~ newBodyFat$ABDOMEN + newBodyFat$AGE)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.1082 -3.2879 -0.1107 3.1423 11.9120   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -39.16391 2.50460 -15.637 < 2e-16 \*\*\*  
## newBodyFat$ABDOMEN 0.59997 0.02746 21.851 < 2e-16 \*\*\*  
## newBodyFat$AGE 0.05942 0.02221 2.675 0.00797 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.303 on 248 degrees of freedom  
## Multiple R-squared: 0.6872, Adjusted R-squared: 0.6846   
## F-statistic: 272.4 on 2 and 248 DF, p-value: < 2.2e-16

Taking out point 39 has lowered the intercept beta0 by 3 (body fat percentage) points and increased the beta1 slope with abdomen size by .037(body fat percentage) points while the linear relationship with age remained roughly similar.

confint(newReducedModel)

## 2.5 % 97.5 %  
## (Intercept) -44.09690526 -34.2309157  
## newBodyFat$ABDOMEN 0.54589570 0.6540543  
## newBodyFat$AGE 0.01566618 0.1031659

The new model contains a 5% error rate, so here are shown the boundaries for 95% percent confidence intervals for the true value of the intercept and each of the predictor variables.

5.Multicollinearity

vif(newReducedModel)

## newBodyFat$ABDOMEN newBodyFat$AGE   
## 1.062087 1.062087

For each predictor, the VIFs are below 5 and thus, we don’t have to worry about issues of multicollinearity.

We decided to leave the measurement of abdomen circumference in cm because when measuring abdominal circumference, a person would likely use a tape measure, which are almost always in units of meters/centimeters.

# Summary of Analysis

Our proposed MLR model if we want to use abdomen circumference and age to predict body fat % is:  
Body Fat Percentage = -39.16 + 0.60 Abdomen circumference (cm) + 0.06 Age

Possible rule of thumb: Multiply your abdomen circumference (in centimeters) by 0.6 and your age by 0.1 and subtract 40.  
A caveat of this rule of thumb is that it will always overestimate body fat percentage compared to the non-rounded model because the slope value for age is higher. However, this rule of thumb is within the 95% confidence intervals for both of the predictors?? slopes and the intercept so it is reasonable.

Example Usage:  
For a man who is 20 and has abdomen circumference of 90.0 cm, their predicted body fat percentage would be 16.04% using the model and 16.0% using the rule of thumb.