CPSC375HW6

Kenn Son, Hamid Suda, Vivian Truong

3/17/2022

Body fat percentage refers to the relative proportions of body weight in terms of lean body mass (muscle, bone, internal organs, and connective tissue) and body fat. The most accurate means of estimating body fat percentage are cumbersome and require specialized equipment. Instead, we can estimate bodyfat percentage from other measurements.

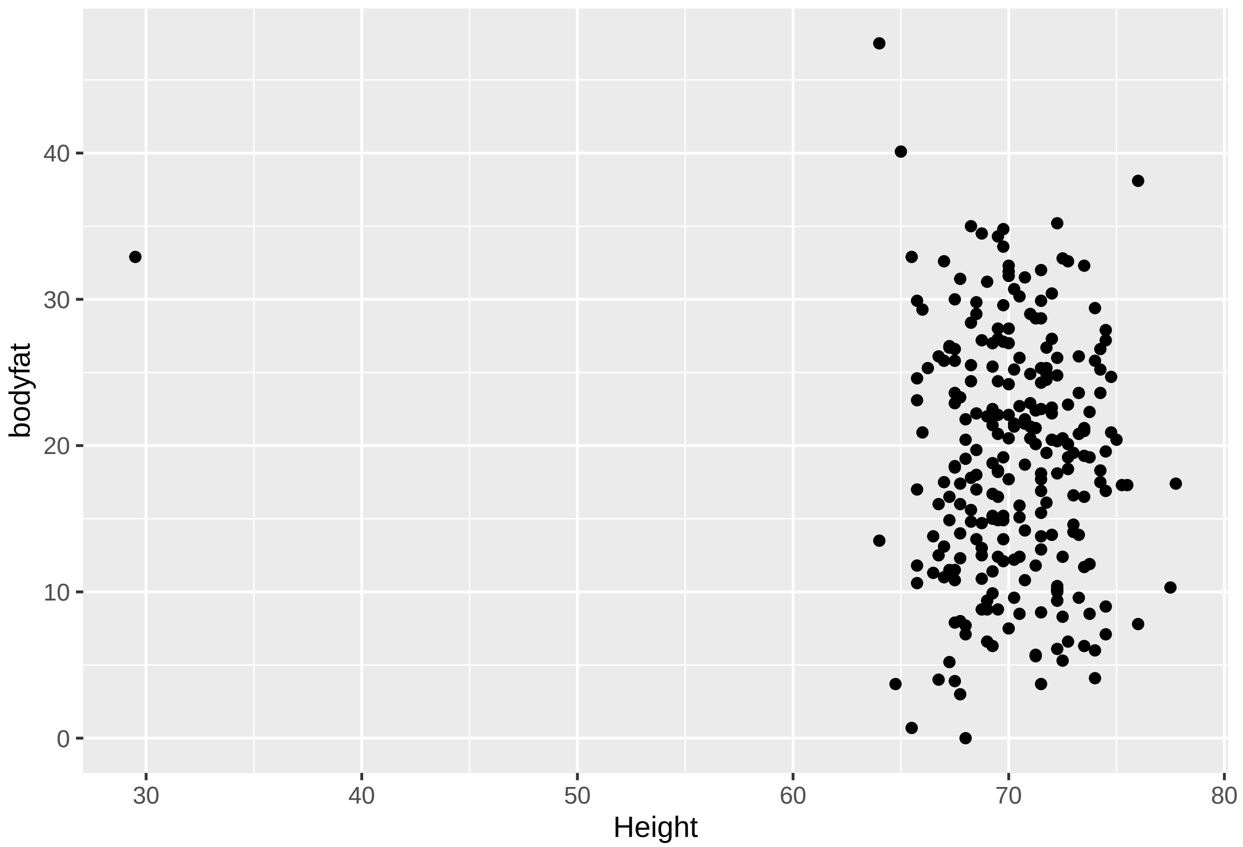
Consider this dataset of 13 measurements from subjects (all men) along with their bodyfat percentage: http://staff.pubhealth.ku.dk/~tag/Teaching/share/data/Bodyfat.csv Note that you can read from the URL directly, like so: read_csv("http://staff.pubhealth.ku.dk/~tag/Teaching/share/data/Bodyfat.csv")

Read the data file and answer the following questions.

```
health <- read.csv("http://staff.pubhealth.ku.dk/~tag/Teaching/share/data/Bodyfat.csv")
#health
```

a) Plot bodyfat vs. Height (code, plot) Which is the dependent variable? Which is the independent variable?

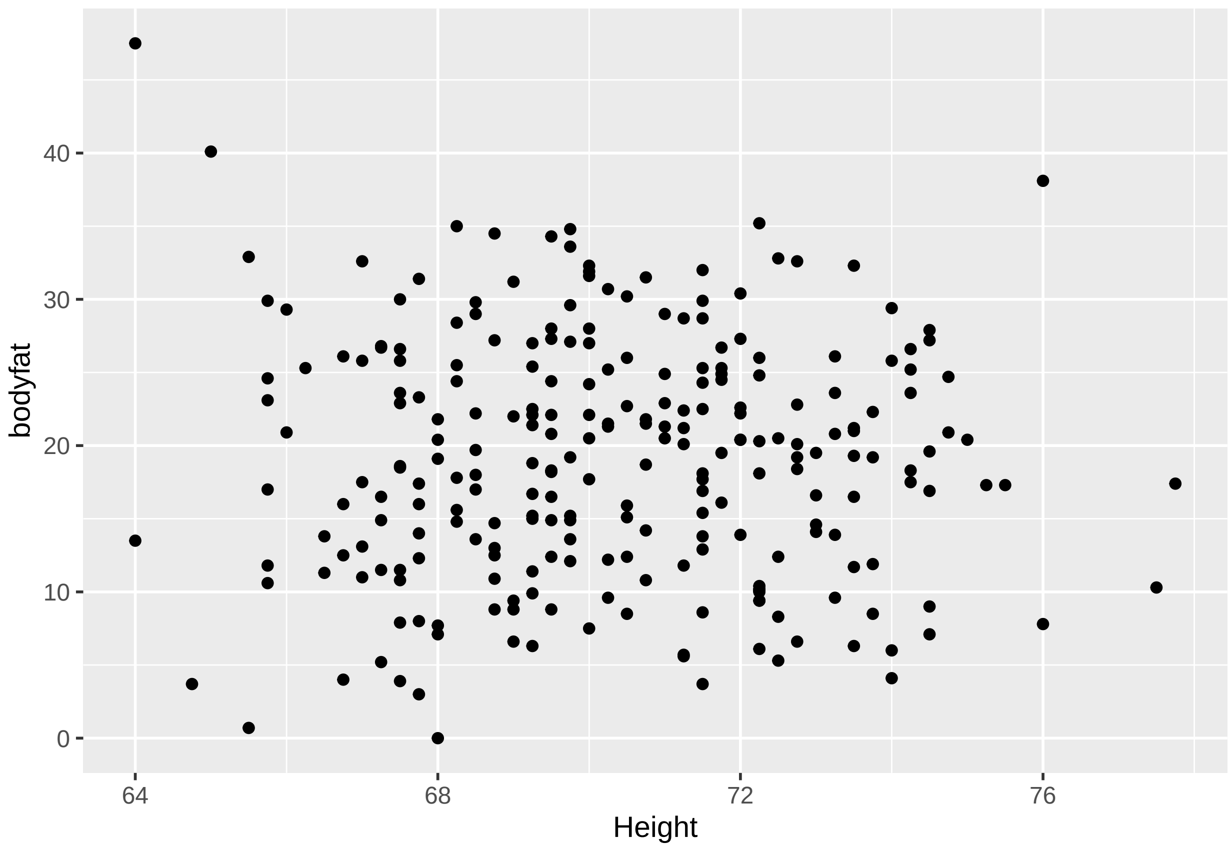
```
ggplot(data=health) + geom_point(mapping = aes(x = Height, y = bodyfat))
```



According to the graph *bodyfat* is the independent as height would be constant, and *Height* is the dependent variable since it is continuous.

b) There is one obvious outlier in the Height column. Remove the corresponding row from the data. (Show: plot, code to remove the row). This will be the data used for the following questions. Confirm that the mean Height is now 70.31076.

```
health <- health %>% filter(Height > 30)
ggplot(data=health) + geom_point(mapping = aes(x = Height, y = bodyfat))
```



```r
health %>% summarise(mean(Height))
```

```
##   mean(Height)
## 1     70.31076
```

c) Create a linear model of bodyfat vs. Height. (code, output of summary(model))

```r
m <- lm(bodyfat~Height, data=health)
summary(m)
```

```
##
## Call:
## lm(formula = bodyfat ~ Height, data = health)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.268  -6.697   0.286   6.162  27.933
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.3412    14.2206   1.712   0.0882 .
## Height       -0.0746     0.2021  -0.369   0.7124
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 8.355 on 249 degrees of freedom
## Multiple R-squared:  0.0005468,  Adjusted R-squared:  -0.003467
## F-statistic: 0.1362 on 1 and 249 DF,  p-value: 0.7124
```

I) What is the R2 value?

R2 value = 0.0005468

II) Is this a "good" model? Why or why not?

By looking at the scatterplot and R2 we can conclude that there is no correlation between Height and bodyfat

III) What is the linear equation relating bodyfat and Height according to this model?

Height = 24.3412 - 0.0746*bodyfat

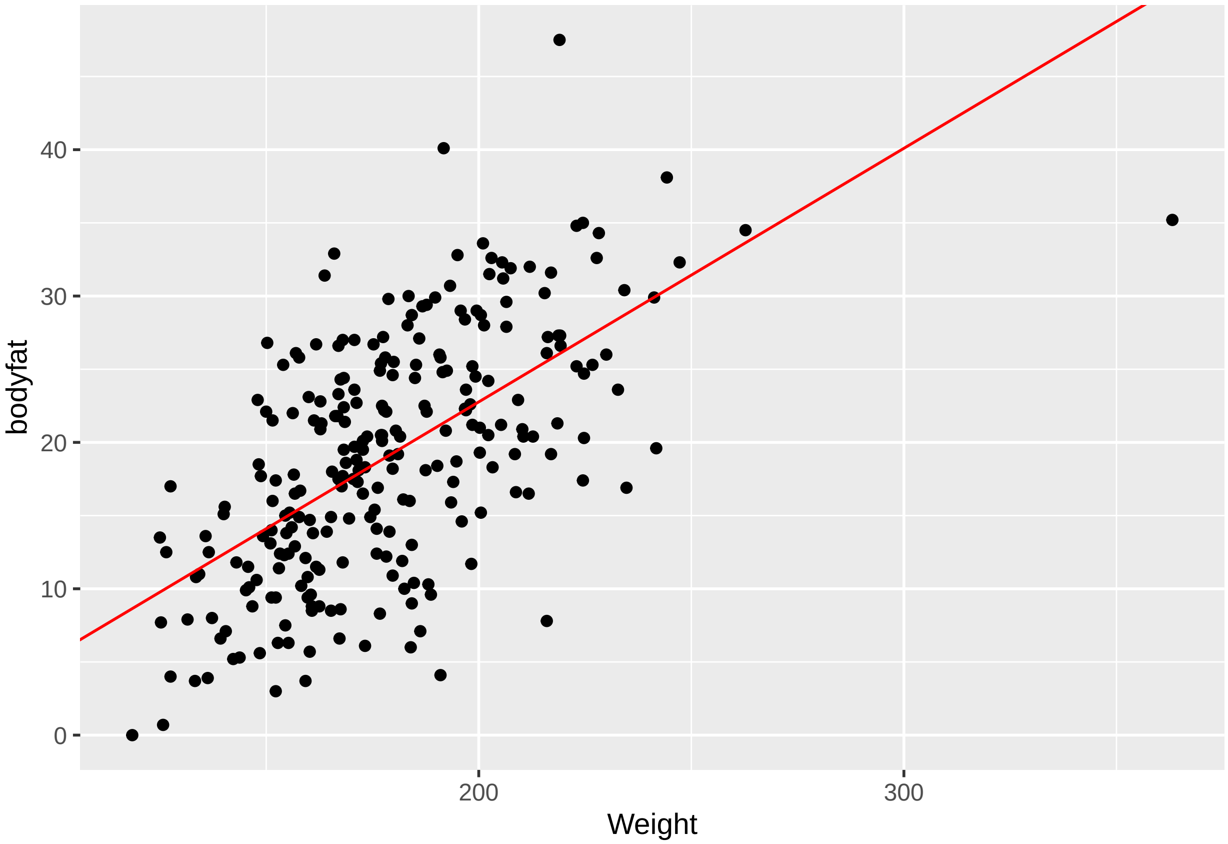
d) Create a linear model of bodyfat vs. Weight. (code, output of summary(model))

```
m <- lm(bodyfat~Weight, data=health)
summary(m)
```

```
##
## Call:
## lm(formula = bodyfat ~ Weight, data = health)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.7382  -4.7052   0.0973   4.9305  21.4419
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.88891    2.57914   -4.61 6.45e-06 ***
## Weight        0.17327    0.01423   12.17  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 6.616 on 249 degrees of freedom
## Multiple R-squared:  0.3731, Adjusted R-squared:  0.3706
## F-statistic: 148.2 on 1 and 249 DF,  p-value: < 2.2e-16
```

I) What is the R2 value?
R2 = 6.616
II) Is this a better model than that based on Height? Why or why not?
Yes this is a better model than the one based on Height. The closer R2 is to 1 it means that there is a stronger correlation between the two variables.
III)What is the linear equation relating bodyfat and Weight according to this model?
bodyfat = -11.88891 + 0.17327*Weight
IV) Plot bodyfat vs. Weight and overlay the best fit line. Use a different color for the line. (plot, code)

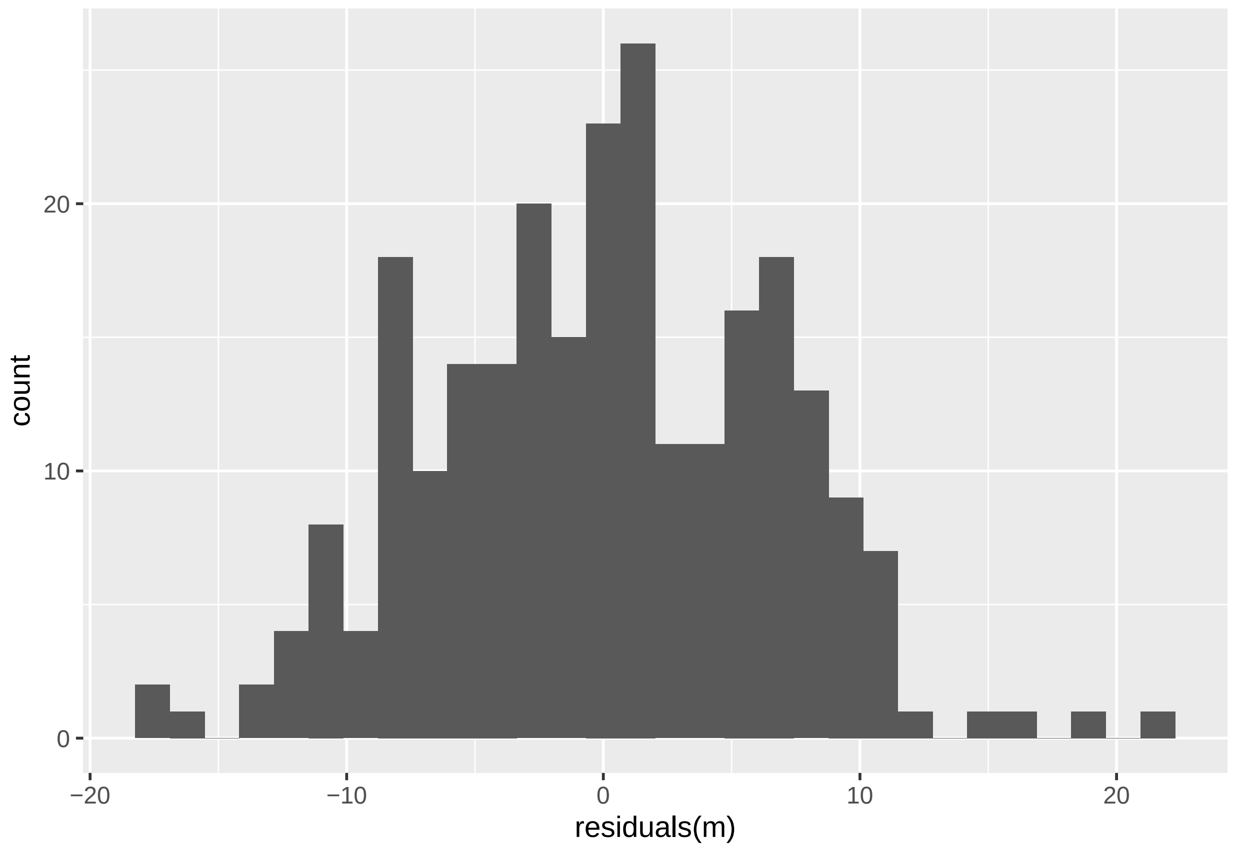
```
cf <- coef(m)
ggplot(data=health) + geom_point(mapping = aes(x = Weight, y = bodyfat)) +
  geom_abline(slope=cf[2], intercept=cf[1], color = "red")
```



V) Plot the histogram of residuals (plot, code). Does this show an approximately normal distribution?

```
ggplot(data=health) + geom_histogram(mapping = aes(x = residuals(m)))
```

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



VI) From the model, predict the bodyfat for two persons: Person A weighs 150 lbs, Person B weighs 300lbs. Include the 99% confidence intervals for the predictions. In which prediction (for Person A or B), are you more confident? Why?

```
predx <- data.frame(Weight = c(150, 300))
predict(m, predx, interval="confidence", level=0.99)
```

```
##        fit      lwr      upr
## 1 14.10217 12.58268 15.62166
## 2 40.09325 35.48700 44.69950
```

I am more confident in Person A because the interval range is smaller in the model compared to Person B.

e) Create a linear model of bodyfat vs. Weight and Height. (code, output of summary(model))

```
m = lm(bodyfat~Weight+Height, data = health)
summary(m)
```

```
##
## Call:
## lm(formula = bodyfat ~ Weight + Height, data = health)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.0328  -3.6411   0.0281   4.3236  13.2125
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 72.52439   10.42582   6.956 3.09e-11 ***
## Weight       0.23195    0.01446  16.037  < 2e-16 ***
## Height      -1.34979    0.16265  -8.299 6.81e-15 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 5.865 on 248 degrees of freedom
## Multiple R-squared:  0.5094, Adjusted R-squared:  0.5054
## F-statistic: 128.7 on 2 and 248 DF,  p-value: < 2.2e-16
```

I) What is the R2 value?
R2 = 0.5094
II) Is this a better model than that based only on Weight or Height? Why or why not?
Yes is it better since it takes in more variables that can correlate with the bodyfat.
III)What is the linear equation relating bodyfat, Weight, and Height according to this model? bodyfat = 72.52439 + 0.23195 * Weight - 1.34979 * Height
IV) From the model, predict the bodyfat for two persons: Person A weighs 150 lbs, Person B weighs 300lbs. Both persons have height=70”. Include the 99% confidence intervals for the predictions. In which prediction (for Person A or B), are you more confident? Why?

```
predx <- data.frame(Weight = c(150, 300), Height = c(70,70))
predict(m, predx, interval="confidence", level=0.99)
```

```
##        fit      lwr      upr
## 1 12.83068 11.42618 14.23519
## 2 47.62251 42.90860 52.33643
```

I am more confident in Person A because the interval is a smaller range with a 99% chance that the data falls in this interval.

f) Add a new transformed variable BMI = Weight/Height2 to the dataset. Create a linear model of bodyfat vs. BMI.

```
health <- health %>% mutate(BMI = (Weight/(Height*Height)))
m <- lm(bodyfat~BMI, data = health)
```

I) Give R code, output of summary(model)

```
summary(m)
```

```
##
## Call:
## lm(formula = bodyfat ~ BMI, data = health)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7769  -3.7061   0.1652   4.1546  12.8061
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -22.859      2.553  -8.955   <2e-16 ***
## BMI         1161.973     69.977  16.605   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.757 on 249 degrees of freedom
## Multiple R-squared:  0.5255, Adjusted R-squared:  0.5236
## F-statistic: 275.7 on 1 and 249 DF,  p-value: < 2.2e-16
```

R2 = 0.5255

II) Is this a better model than the previous models? Why or why not?

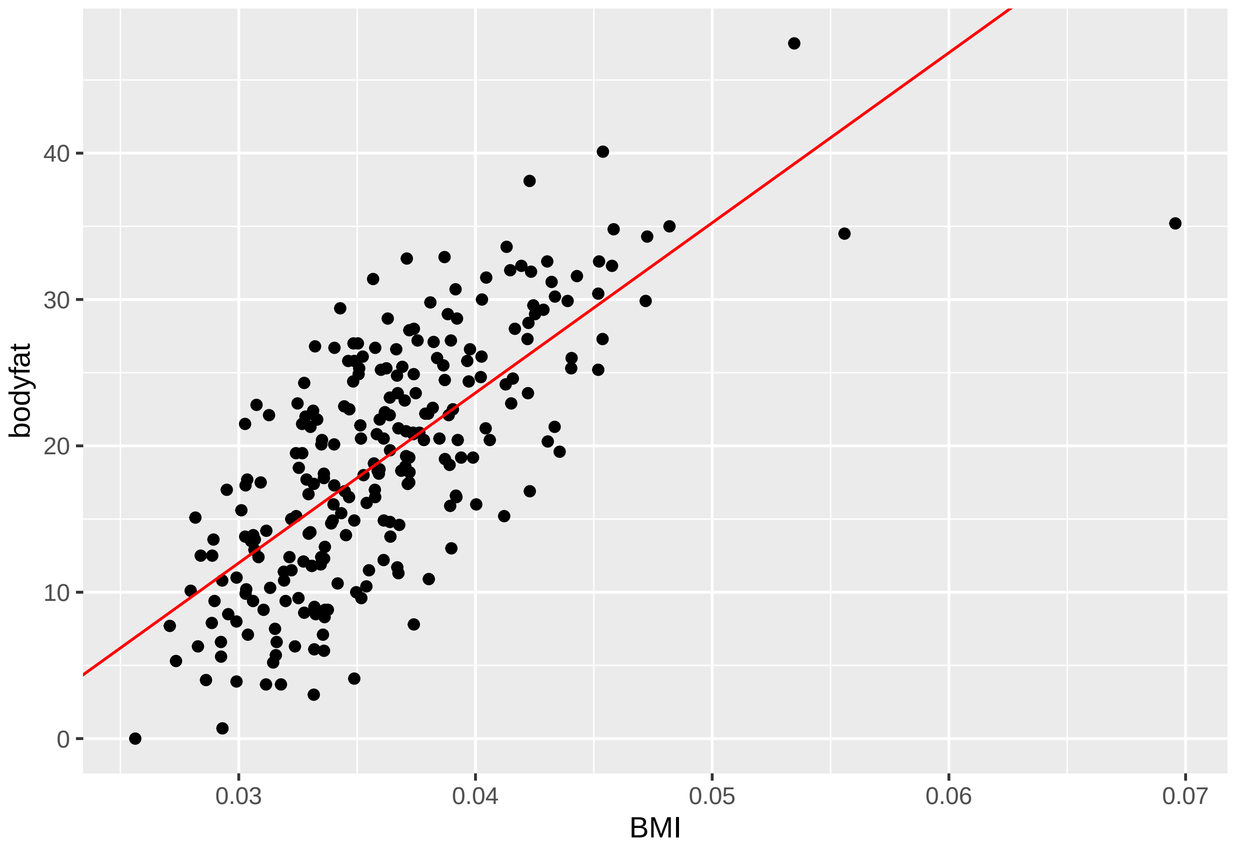
This is a better model because the R2 is closer to one meaning that theres a higher correlation in the model.

III)What is the equation relating bodyfat, Weight, and Height according to this model? Is this a linear or nonlinear equation?

This model is going to be linear even though we took Height and squared it. It became on variable and therefore it is now linear. The equation is "bodyfat = -22.859 + 1161.973*BMI"

IV) Plot bodyfat vs. BMI and overlay the best fit model as a straight line. (code, plot)

```
cf <- coef(m)
ggplot(data=health) + geom_point(mapping = aes(x = BMI, y = bodyfat)) +
  geom_abline(slope=cf[2], intercept=cf[1], color = "red")
```



V) From the model, predict the bodyfat for two persons: Person A weighs 150 lbs, Person B weighs 300lbs. Both persons have height=70”. Include the 99% confidence intervals for the predictions.

```
PersonA_BMI <- (150/(70*70))
PersonB_BMI <- (300/(70*70))
predx <- data.frame(BMI = c(PersonA_BMI, PersonB_BMI))
predict(m, predx, interval="confidence", level=0.99)
```

```
##        fit      lwr      upr
## 1 12.71124 11.33803 14.08446
## 2 48.28185 43.62305 52.94065
```

VI) Body Mass Index (BMI) is actually defined as a person’s weight in kilograms divided by the square of height in meters but your data has Weight in pounds and Height in inches. Thus, the correct BMI transformation should have been BMI = (Weight/2.20)/(Height*0.0254)2. Would using this correct BMI transformation result in a different model from what was calculated? Why or why not?

Yes the model would be non-linear but different as the ratio of height and weight aren’t the same. The new kilo model would be accurate too.

g) Add a new categorical variable (factor) AgeGroup to the dataset. AgeGroup should have three values: “Young” for Age<40, “Middle” for Age between 40 and 60, and “Older” for Age>60.

I) Show R code that adds the AgeGroup variable. This can be done with mutate and the cut() function like so: cut (Age, breaks = c(-Inf,40,60,Inf), labels = c(“Young”, “Middle”, “Older”)[Code]

```r
col = cut(health$Age, breaks = c(-Inf,40,60,Inf), labels = c("Young", "Middle", "Older"))
health <- health %>% mutate(AgeGroup = col)
```

II) Create a linear model of bodyfat vs. BMI and AgeGroup.[Code, output of summary(model)]

```r
m <- lm(bodyfat~BMI+AgeGroup, data=health)
summary(m)
```

```
##
## Call:
## lm(formula = bodyfat ~ BMI + AgeGroup, data = health)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.4537  -3.9137  -0.1361   3.7127  12.0269
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -22.8344     2.4552  -9.301  < 2e-16 ***
## BMI            1105.0576    67.8315  16.291  < 2e-16 ***
## AgeGroupMiddle    2.6113     0.7607   3.433    7e-04 ***
## AgeGroupOlder     5.3074     1.1075   4.792 2.85e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.502 on 247 degrees of freedom
## Multiple R-squared:   0.57,  Adjusted R-squared:  0.5648
## F-statistic: 109.2 on 3 and 247 DF,  p-value: < 2.2e-16
```

bodyfat=-22.8344+1105.0576* BMI+2.6113 * AgeGroupMiddle + 5.3074 * AgeGroupOlder

III)How many dummy (i.e., 0-1) variables were created in the model?

Two dummy variables are made ([0,0][0,1][1,0])

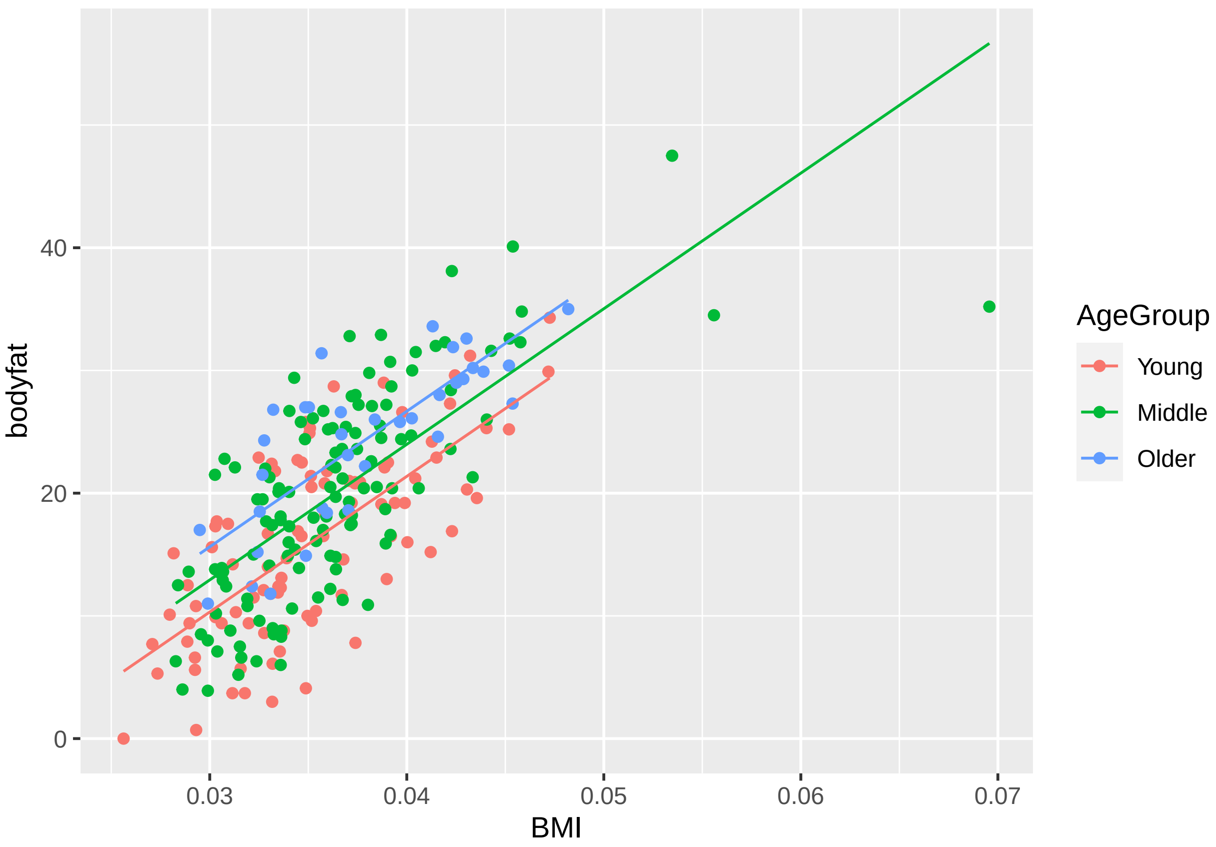
IV) Is this a better model than the previous models? Why or why not?
I would say yes as it is more specific on its target in the data for each AgeGroup also with a R2 of .57

V) What are the set of equations relating bodyfat, BMI, and AgeGroup according to this model?

bodyfat=-22.8344+1105.0576* BMI
bodyfat=-22.8344+1105.0576* BMI + 2.6113 * AgeGroupMiddle
bodyfat=-22.8344+1105.0576* BMI + 5.3074 * AgeGroupOlder

VI) Plot bodyfat vs. BMI and overlay the model predictions (multiple lines: one for each value of the discrete variable). [Code, plot]

```r
health <- health %>% add_predictions(m)%>% arrange(AgeGroup)
ggplot(data=health) + geom_point(aes(x = BMI, y = bodyfat, color = AgeGroup)) +
  geom_line(mapping=aes(x=BMI, y=pred, color=AgeGroup))
```


bodyfat
0
20
40
BMI
0.03
0.04
0.05
0.06
0.07
AgeGroup
Young
Middle
Older