Homework 7

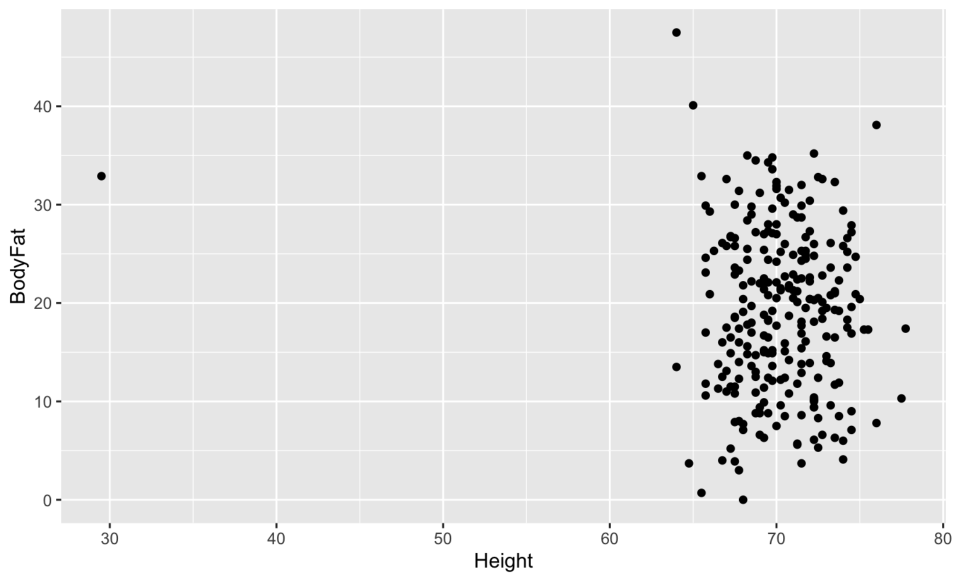
Mariia Nikitash
Prepare your answers as a **single PDF file**.
Group work: You may work in groups of 1-3. Include all group member names in the PDF file. You may work with students in both sections (375-01, -02). Only one person in the group should submit to Canvas.
Due: check on Canvas.

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.
You probably already know that body fat percentage is an important indicator of overall health - too little or too much body fat is associated with several health issues. This assignment is about *estimating body fat* percentage from *other body measurements*.

a. Why is there a need to *estimate* body fat percentage instead of directly *measuring* it (e.g., we can directly measure a person's weight, we don't have to calculate it)? Do an internet search and answer in 2-3 sentences.
   Body Mass Index (BMI) is not a perfect measure because it does not directly assess body fat. Muscle and bone are denser than fat, so BMI can overestimate body fat in athletes with high bone density and muscle mass or underestimate it in older people who have low bone density and muscle mass.

b. The **`bodyfat.csv`** file in the Datasets module on Canvas contains 13 measurements from subjects (all men) along with their body fat percentage[1]. Read the file using `read_csv()`.
   i. Plot `BodyFat` vs. `Height` (code, plot)
      ggplot(data=bodyfat) + geom_point(mapping = aes(x=Height, y=BodyFat))

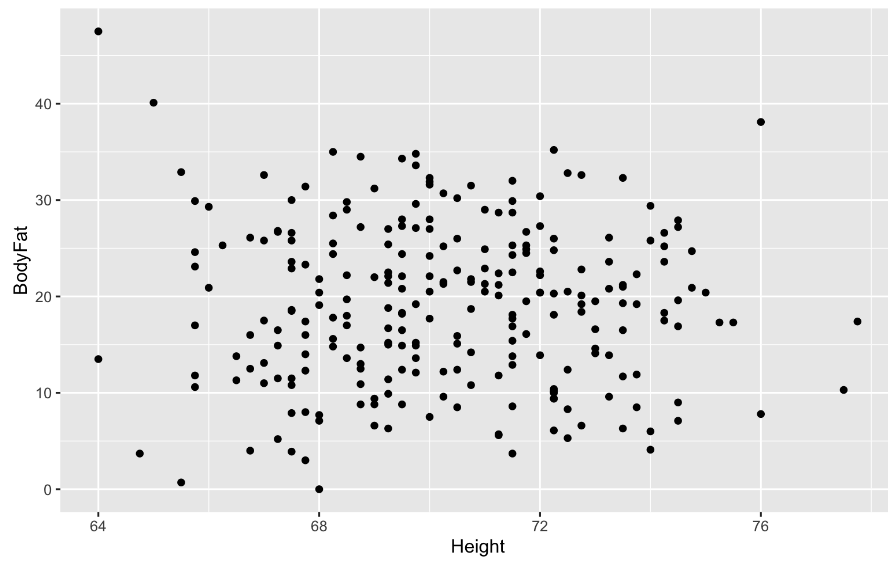
[1] https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset?resource=download



ii. Which should be the dependent variable?
Dependent variable should be BodyFat, since that is the variable we have to estimate from the other body measurements, which are independent
Which is the independent variable?
Independent variable is Height, since it is one of the measurements that are given and we have to estimate body fat by this given measurements .

c. There is one obvious outlier in the Height column. Remove the corresponding row from the data and plot again. This will be the data used for the following questions. Confirm that the mean Height is now 70.31076.
   i. Show code to remove the row, plot, and calculate mean;
   ii.

```
filteredbodyfat <- bodyfat %>% filter(Height > 30) %>% mutate(mean(Height))

ggplot(data=filteredbodyfat) + geom_point(mapping = aes(x=Height, y=BodyFat))
```



iii. Show plot

d. Create a linear model of `BodyFat` vs. `Height`.

i. Show code, output of summary(model)

```
m <- lm(BodyFat~Height, data=filteredbodyfat)

> summary(m)

Call:
lm(formula = BodyFat ~ Height, data = filteredbodyfat)

Residuals:
    Min      1Q  Median      3Q 
-19.268  -6.697   0.286   6.162 
    Max 
 27.933 

Coefficients:
            Estimate Std. Error
(Intercept)  24.3412    14.2206
Height       -0.0746     0.2021
            t value Pr(>|t|)
(Intercept)   1.712   0.0882 .
Height       -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
```

ii. What is the R2 value?

```
Multiple R-squared:  0.0005468
```

iii. Is this a “good” model? Why or why not?

R2 value of approximately 0.0005468, indicates that only a very small fraction of the variability in BodyFat can be explained by the Height variable. In general, a higher R2 value closer to 1 indicates a better fit of the model to the data, suggesting that more of the variability in the dependent variable is explained by the independent variable(s).

Given the very low R2 value in this case, it suggests that Height alone is not a good predictor of BodyFat. Therefore, this model may not be considered "good" for predicting BodyFat based solely on Height.

iv. What is the linear equation relating BodyFat and Height according to this model?

BodyFat = 24.3412 - 0.0746 x Height
Dependent var. Y = intercept + slope x Independent Var X

e. Create a linear model of `BodyFat` vs. **`Weight`**.

i. Show code, output of summary(model)

ggplot(data=filteredbodyfat) + geom_point(mapping = aes(x=Weight, y=BodyFat))

```
Call:
lm(formula = BodyFat ~ Weight, data = filteredbodyfat)

Residuals:
     Min       1Q   Median       3Q
-17.7382  -4.7052   0.0973   4.9305
     Max
 21.4419

Coefficients:
             Estimate Std. Error t value
(Intercept) -11.88891    2.57914   -4.61
```

```
Weight        0.17327    0.01423   12.17
            Pr(>|t|)
(Intercept) 6.45e-06 ***
Weight       < 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
```

ii. What is the R2 value?

```
Multiple R-squared:  0.3731
```

iii. Is this a better model than that based on Height? Why or why not?

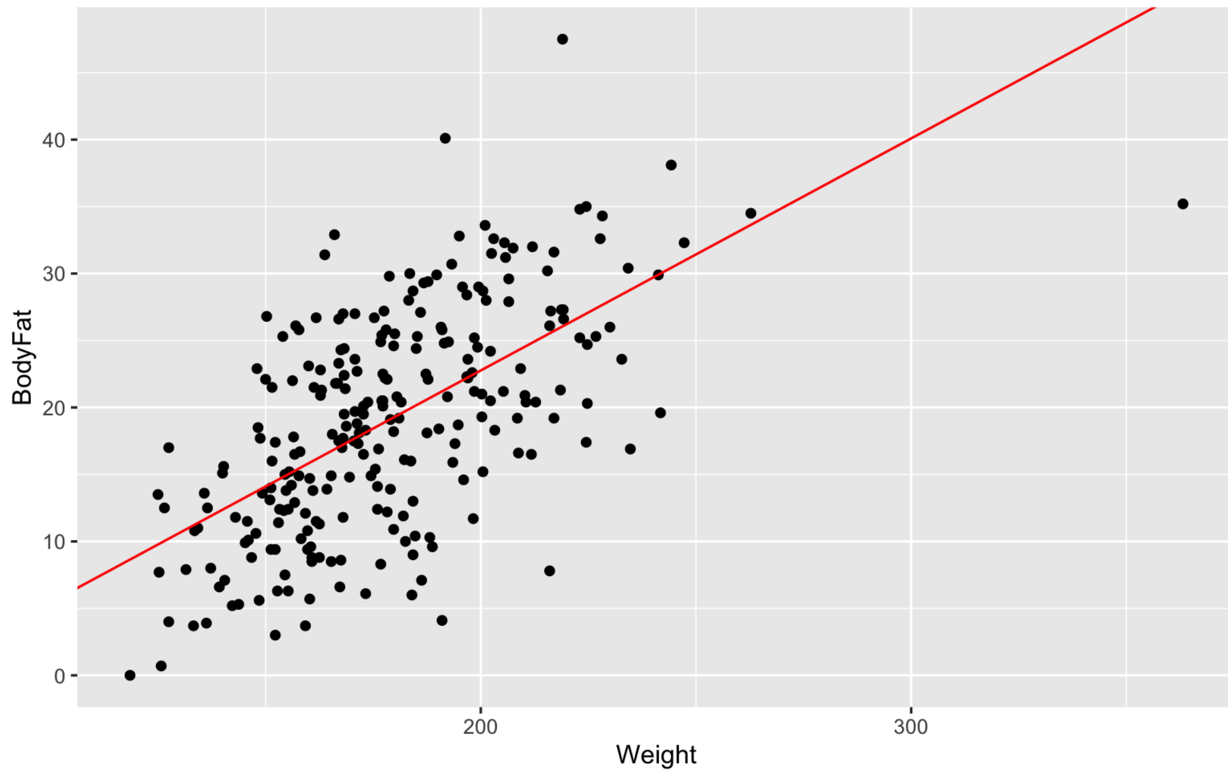
This model based on Weight is better, because R2 value is closer to 1, then in moel based on Height. It means that model based on Weight has better fit of the model to the data.

iv. What is the linear equation relating BodyFat and Weight according to this model?

BodyFat = -11.88891 + 0.17327 x Weight

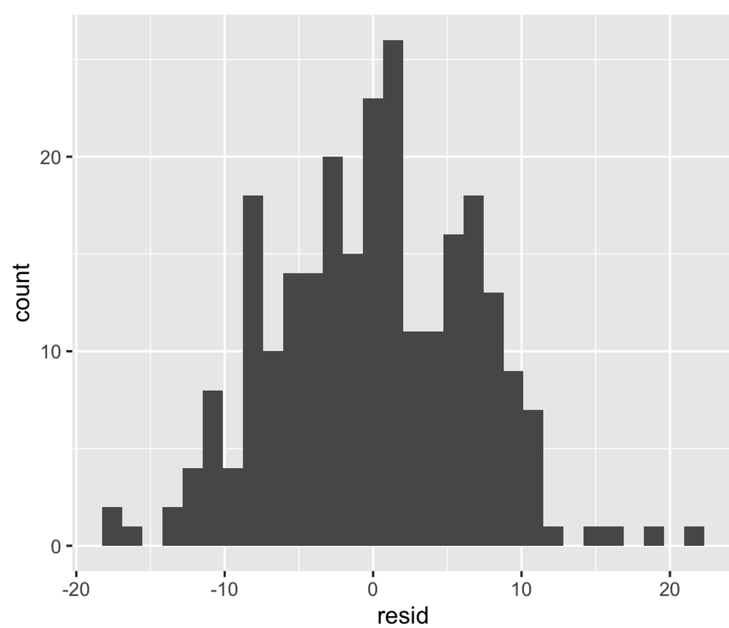
v. Plot BodyFat vs. Weight and overlay the best fit line. Use a different color for the line. (plot, code)

```
ggplot(data=filteredbodyfat) + geom_point(mapping = aes(x=Weight, y=BodyFat)) + geom_abline(slope = 0.17327, intercept = -11.88891, color="red")
```



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

```
ggplot(data=filteredbodyfat) + geom_histogram(mapping = aes(x=resid))
```



It shows an approximately normal distribution because it is bell-shaped distribution.

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

```
predA <- data.frame(Weight=c(150))
predict(m, predA, interval = "confidence", level = 0.99)
       fit      lwr      upr
1 14.10217 12.58268 15.62166

predB <- data.frame(Weight=c(300))

> predict(m, predB, interval = "confidence", level = 0.99)
       fit    lwr     upr
1 40.09325 35.487 44.6995
```

Comparing the widths of the confidence intervals, we observe that the confidence interval for Person A (150 lbs) is narrower than the confidence interval for Person B (300 lbs). A narrower confidence interval suggests that we are more confident in the prediction because it indicates less uncertainty about the true value of the prediction.

f. Create a linear model of `BodyFat` vs. **`Weight and Height`**.

i. Show code, output of summary(model)

```
m1 <- lm(BodyFat~Weight+Height, data=filteredbodyfat)
>
> summary(m1)

Call:
lm(formula = BodyFat ~ Weight + Height, data =
filteredbodyfat)

Residuals:
     Min       1Q   Median       3Q
-24.0328  -3.6411   0.0281   4.3236
     Max
 13.2125

Coefficients:
            Estimate Std. Error t value
```

```
(Intercept) 72.52439   10.42582   6.956
Weight       0.23195    0.01446  16.037
Height      -1.34979    0.16265  -8.299
            Pr(>|t|)
(Intercept) 3.09e-11 ***
Weight       < 2e-16 ***
Height      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
```

ii. What is the R2 value?

```
Multiple R-squared:  0.5094
```

iii. Is this a better model than that based only on Weight or Height? Why or why not?

Model based on Weight + Height is better then only based on Weight or Height since it’s R2 value is significantly larger. The higher its value, the ore accurate is the regression model.

iv. What is the linear equation relating BodyFat, Weight, and Height according to this model?

Dependent var. Y = intercept + slope x Independent Var X

BodyFat = 72.5243873 + 0.2319455 x Weight -1.3497934 x Height

v. From the model, predict the BodyFat for two persons: Person A weighs 150 lbs, Person B weighs 300 lbs. 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?

```
predictPersonA <- data.frame(Weight=150, Height=70)
> predict(m1,predictPersonA,interval="confidence",level=0.99)
     fit      lwr      upr
1 12.83068 11.42618 14.23519

> predictPersonB <- data.frame(Weight=300, Height=70)
```

```
> predict(m1,predictPersonB,interval="confidence",level=0.99)
       fit     lwr      upr
1 47.62251 42.9086 52.33643
```

A narrower confidence interval indicates less uncertainty in the prediction. Since predictPersonA CI is narrower then predictPersonB CI I can conclude that I am more confident in predictPersonA prediction.

g. Add a new transformed variable **BMI** = **Weight/Height²** to the dataset. Create a linear model of `BodyFat` vs. **`BMI`**.
   i. Show code, output of summary(model)

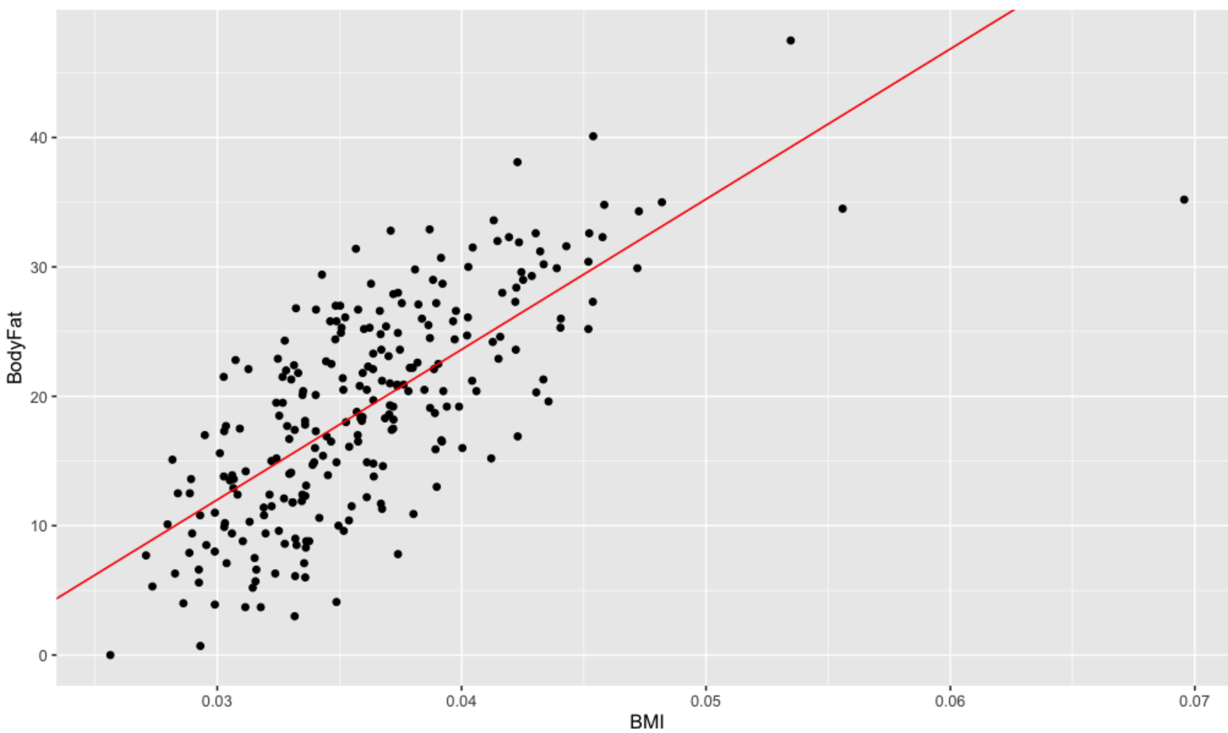
```
bmiset <- filteredbodyfat %>% mutate(BMI=(Weight/Height^2))
m3 <- lm(BodyFat~BMI, data=bmiset)
> summary(m3)

Call:
lm(formula = BodyFat ~ BMI, data = bmiset)

Residuals:
     Min       1Q   Median       3Q 
-22.7769  -3.7061   0.1652   4.1546 
     Max 
 12.8061 

Coefficients:
            Estimate Std. Error t value
(Intercept)  -22.859      2.553  -8.955
BMI         1161.973     69.977  16.605
            Pr(>|t|)    
(Intercept)   <2e-16 ***
BMI           <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
```

ii. Is this a better model than the previous models? Why or why not?
The R2 of this model is the highest => Multiple R-squared: 0.5255, so this model is a better model then all of the above. The reason is the R2 value is closer to 1, then in other models.

What is the equation relating BodyFat, Weight, and Height according to this model? Is this a linear or nonlinear equation?
This is a nonlinear equation.
BodyFat = -22.859 + 1161.973 * (Weight/Height^2).

iii. Plot `BodyFat` vs. `BMI` and overlay the best fit model as a straight line. (code, plot)

```
ggplot(data=bmiset)+ geom_point(mapping = aes(x=BMI, y=BodyFat)) +
geom_abline(intercept = coef(m3)[1], slope=coef(m3)[2], color="red")
```



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

```
predTwo<- data.frame(Weight=c(150, 300), Height=c(70,70)) %>%
mutate(BMI=Weight/Height^2)
>
```

```
> predTwo
  Weight Height        BMI
1    150     70 0.03061224
2    300     70 0.06122449

predict(m3, predTwo, interval = "confidence", level = 0.99)
       fit      lwr      upr
1 12.71124 11.33803 14.08446
2 48.28185 43.62305 52.94065
```

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

Model would be the same and only the units in linear transformation will have different calculation. If we would use Weight in pounds, but Height in cm that would be different, since measurement units are different.

h. 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 "Senior" for Age>60.

   i. Show 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", "Senior")`

```
bmiset <- bmiset %>% mutate(AgeGroup = cut (Age, breaks = c(-Inf,40,60,Inf),
labels = c("Young", "Middle", "Senior")))
```

   ii. Create a linear model of `BodyFat` vs. **`BMI and AgeGroup.`** [Code, output of summary(model)]

```
m4 <- lm(data=bmiset, formula = BodyFat~BMI+AgeGroup)
>
> summary(m4)

Call:
```

[2] https://www.cdc.gov/healthyweight/assessing/bmi/index.html

```
lm(formula = BodyFat ~ BMI + AgeGroup, data = bmiset)

Residuals:
     Min       1Q   Median       3Q      Max
-21.4537  -3.9137  -0.1361   3.7127  12.0269

Coefficients:
              Estimate Std. Error t value
(Intercept)    -22.8344     2.4552  -9.301
BMI           1105.0576    67.8315  16.291
AgeGroupMiddle    2.6113     0.7607   3.433
AgeGroupSenior    5.3074     1.1075   4.792
               Pr(>|t|)
(Intercept)     < 2e-16 ***
BMI            < 2e-16 ***
AgeGroupMiddle    7e-04 ***
AgeGroupSenior 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
```

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

2 dummy variables were created AgeGroupMiddle, AgeGroupSenior

iv. Is this a better model than the previous models? Why or why not?

This model is better then al the others, because it's R2 value is the closest to 1

v. What are the set of equations relating BodyFat, BMI, and AgeGroup according to this model?

BodyFat = -22.8344 + 1105.0576*BMI + 2.6113*AgeGroupMiddle + 5.3074*AgeGroupSenior

Middle age group: (dummy vars are AgeGroupMiddle == 1, AgeGroupSenior=0)
BodyFat = (-22.8344 + 2.6113*AgeGroupMiddle) + 1105.0576*BMI

Senior age Group: (dummy vars are AgeGroupMiddle == 0, AgeGroupSenior=1)
BodyFat = (-22.8344 + + 5.3074*AgeGroupSenior) + 1105.0576*BMI

Young Age Group: (dummy vars are AgeGroupMiddle == 1, AgeGroupSenior=0)
BodyFat = -22.8344 + 1105.0576*BMI

i. Plot `BodyFat` vs. `BMI` and overlay the model predictions (Hint: add a new column with predictions and plot the predictions using geom_line. You should see multiple lines, one for each value of the discrete variable). [Code, plot]

library(modelr)

bmiset <- bmiset %>% add_predictions(m4)

ggplot(data=bmiset)+geom_point(mapping = aes(x=BMI, y=BodyFat)) + geom_line(mapping = aes(x=BMI, y=pred, color=AgeGroup))