Homework 5

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.

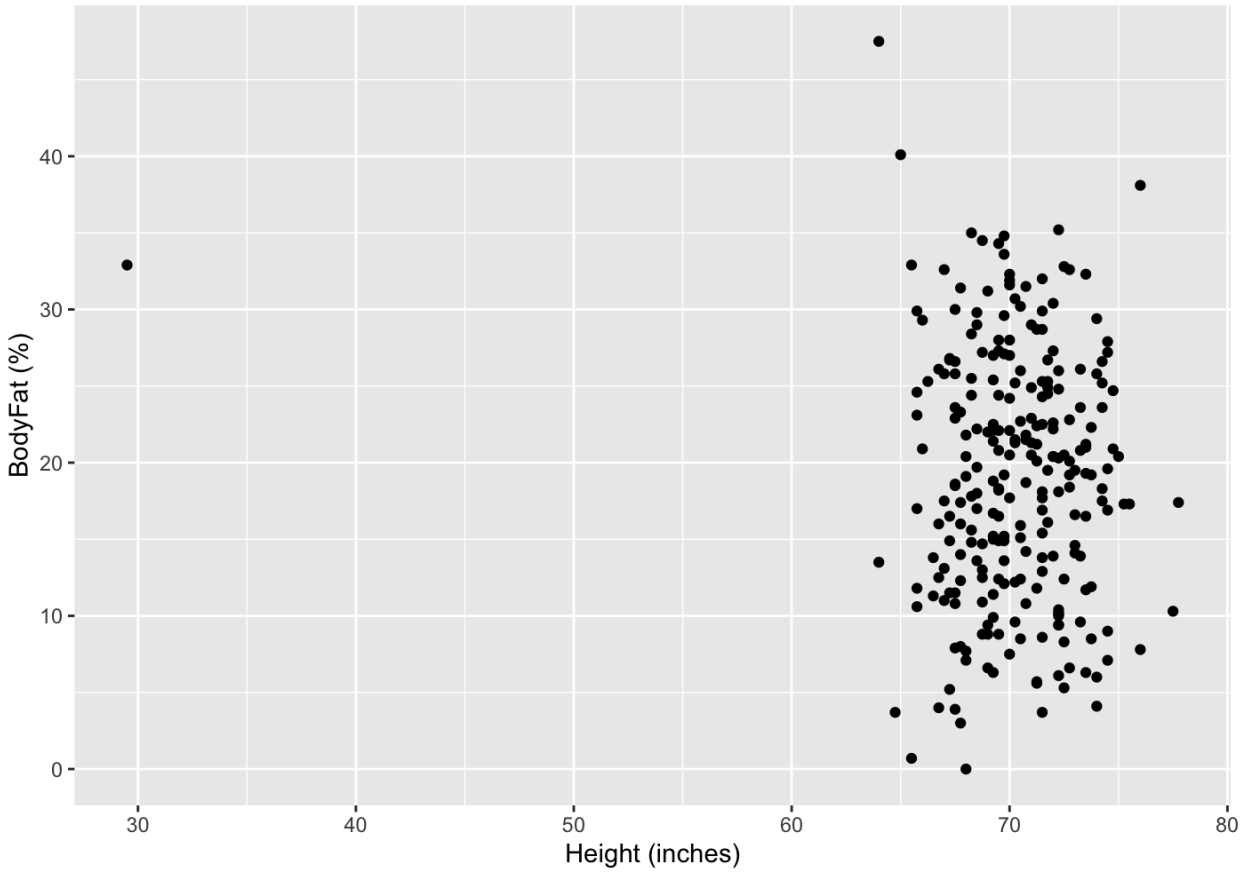
Direct measurement of body fat percentage often requires techniques like dual-energy X-ray absorptiometry (DXA), underwater weighing, or skinfold measurements, which can be invasive, time-consuming, and may not be suitable for regular health monitoring. Estimation methods can use easily accessible measurements like weight, height, and circumference.

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(). Plot `BodyFat` vs. `Height` (code, plot) Which should be the dependent variable? Which is the independent variable?

BodyFat should be the dependent variable, as we want to understand how body fat percentage changes concerning changes in height, which should be the independent variable.

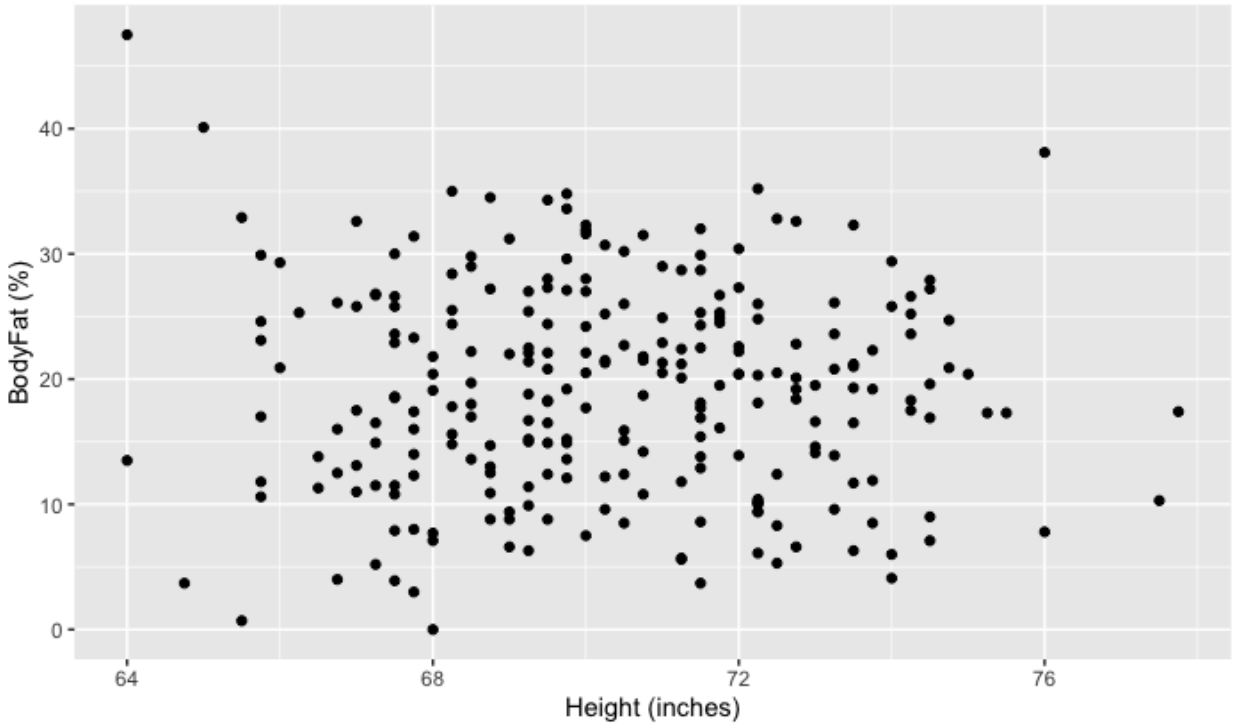
> ggplot(data=survey, aes(x = Height, y = BodyFat))+geom_point()+labs(x = "Height (inches)", y = "BodyFat (%)")

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



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. (Show: code to remove the row, plot, and calculate mean; plot).

```
> library(dplyr)
> surveyIQR <- IQR(survey$Height)
> lower_bound <- quantile(survey$Height)[2] - 1.5 * surveyIQR
> upper_bound <- quantile(survey$Height)[4] + 1.5 * surveyIQR
> filteredSurvey <- survey %>% filter(Height > lower_bound & Height < upper_bound)
> ggplot(data=filteredSurvey, aes(x = Height, y = BodyFat))+geom_point()+labs(x = "Height
(inches)", y = "BodyFat (%)")
```



d. Create a linear model of `BodyFat` vs. `Height`. (code, output of summary(model))

i. What is the R2 value?

```
> model <- lm(BodyFat~Height, data =filteredSurvey)
> summary(model)

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

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
```

R2 is 0.0005468

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

This model is not considered good because the p-value for the Height coefficient is much larger than commonly accepted significance levels, indicating that Height is not a statistically significant predictor of BodyFat.

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

BodyFat = 24.3412 - 0.0746(Height)

e. Create a linear model of `BodyFat` vs. **`Weight`**. (code, output of summary(model))

i. What is the R2 value?\

```
> model_weight <- lm(BodyFat ~ Weight, data = filteredSurvey)
> summary(model_weight)

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

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
```

R2 is 0.3731

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

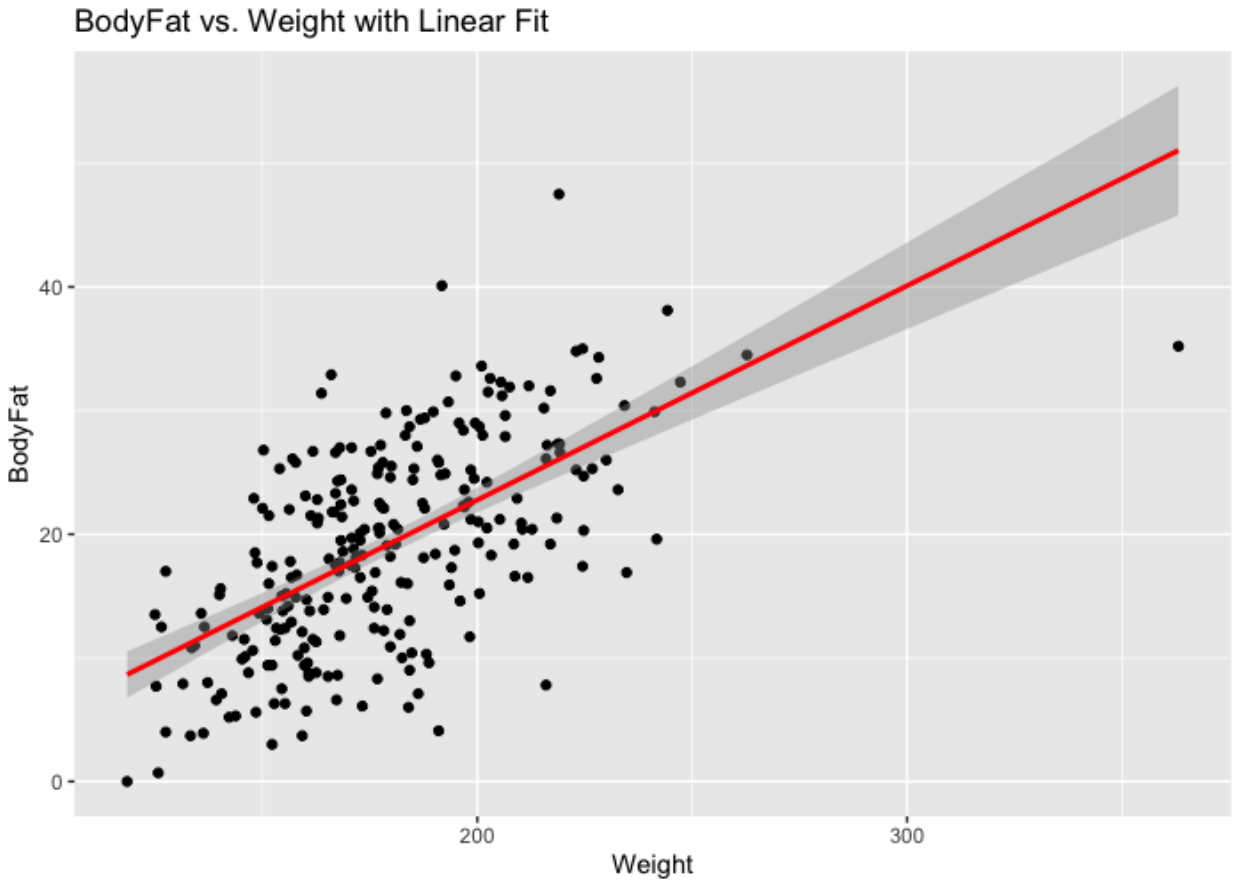
This model is considered better than the Height model because the R-squared value is higher, indicating that Weight explains a significant portion of the variance in BodyFat.

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

BodyFat = -11.88891 + 0.17327 * Weight

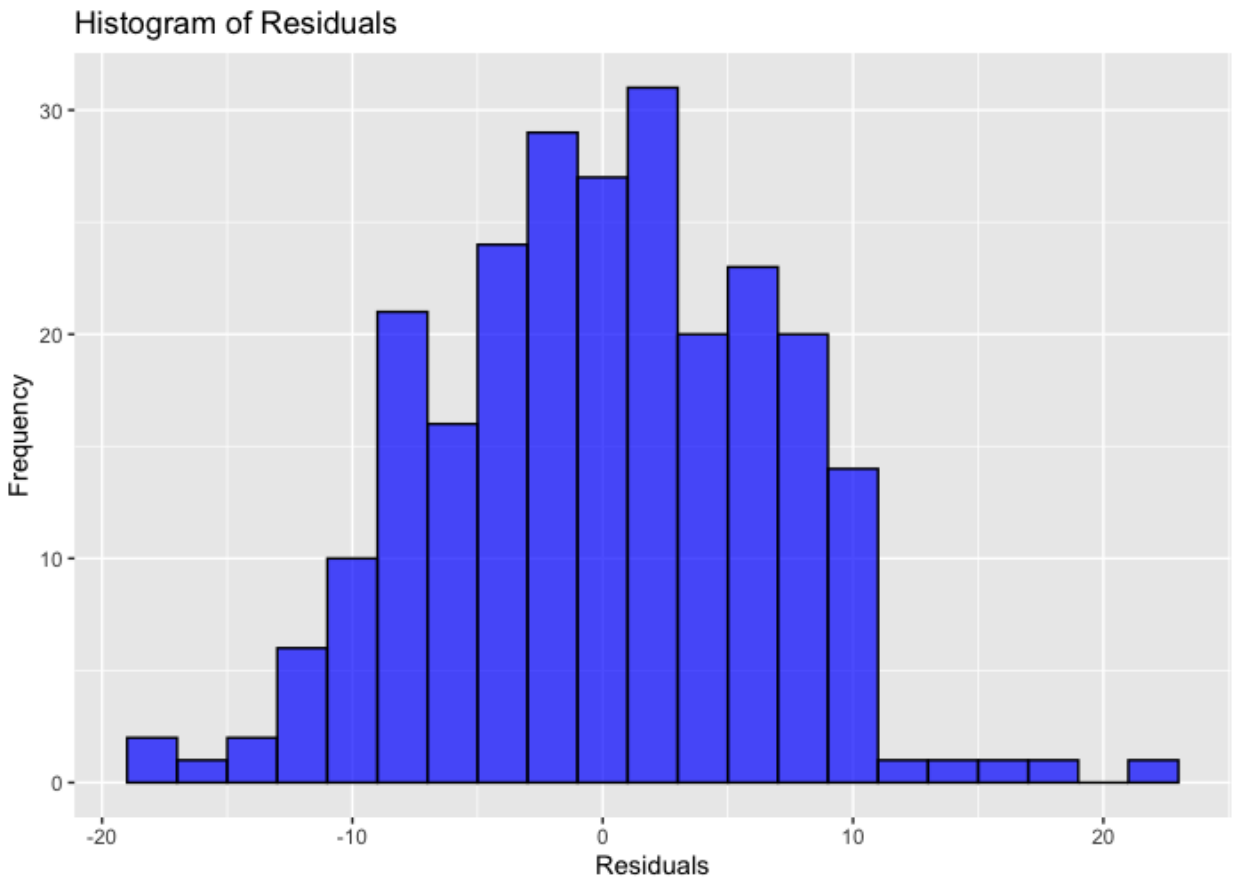
iv. Plot BodyFat vs. Weight and overlay the best fit line. Use a different color for the line. (plot, code)

```
> ggplot(filteredSurvey, aes(x = Weight, y = BodyFat)) +geom_point() + geom_smooth(method = "lm", color = "red") + labs(x = "Weight", y = "BodyFat") + ggtitle("BodyFat vs. Weight with Linear Fit")
```



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

```
> ggplot(residuals_df, aes(x = Residuals)) + geom_histogram(binwidth = 2, fill = "blue", color = "black", alpha = 0.7) + labs(title = "Histogram of Residuals",x = "Residuals",y = "Frequency")
```



Yes, it does show an approximately normal distribution

vi. 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?

```
> persons_val <- data.frame(Weight = c(150, 300))
> predictions <- predict(model_weight, newdata = persons_val,  interval = "confidence", level =
0.99)
> predictions
       fit      lwr      upr
1 14.10217 12.58268 15.62166
2 40.09325 35.48700 44.69950
```

Person A (Weight = 150 lbs) because the confidence interval is narrower, indicating a smaller range of uncertainty around the predicted BodyFat value. This means that the prediction for Person A is more precise compared to Person B.

f. Create a linear model of `BodyFat` vs. **`Weight and Height`**. (code, output of summary(model))
   i. What is the R2 value?

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

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

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
```

R2 is 0.5094

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

This model is better than the models based only on Weight or Height individually. The R-squared value of 0.5094 is higher than the R-squared values obtained for the individual models, suggesting that the combined model provides a better fit to the data and explains more of the variation in 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 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?

```
> persons_val <- data.frame(Weight = c(150, 300), Height = c(70, 70))
> predictions <- predict(model_combined, newdata = persons_val, interval = "confidence", level
= 0.99)
> predictions
       fit      lwr      upr
1 12.83068 11.42618 14.23519
2 47.62251 42.90860 52.33643
>
```

Person A because their confidence interval is narrower (indicating greater precision) compared to Person B's prediction, which has a wider confidence interval.

g. Add a new transformed variable **BMI** = **Weight/Height^2** to the dataset. Create a linear model of `BodyFat` vs. **`BMI`**.

   i. Give R code, output of summary(model)

```
> filteredSurvey <- filteredSurvey %>% mutate(BMI = Weight / (Height * Height))
> model_BMI <- lm(BodyFat ~ BMI, data = filteredSurvey)
> summary(model_BMI)

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

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 is 0.5255

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

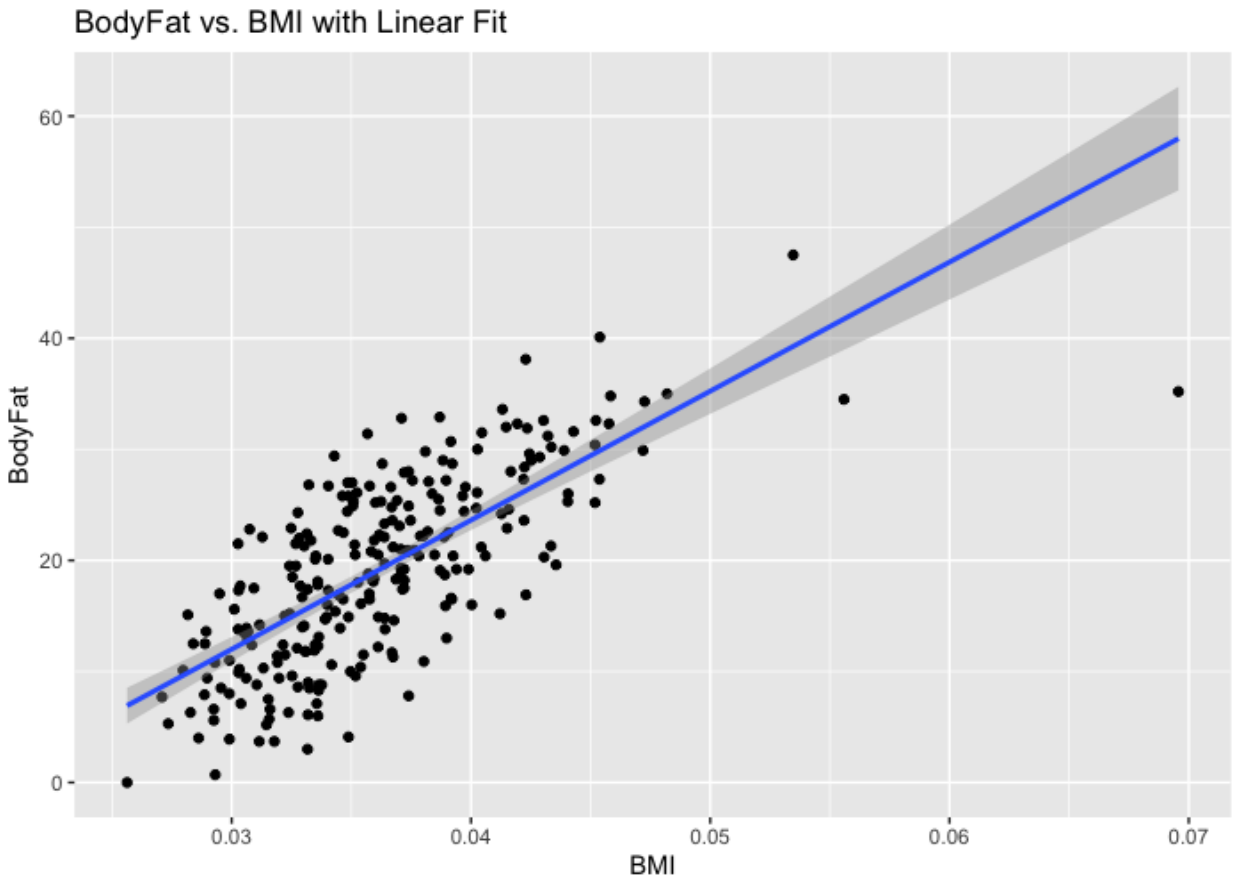
It has a higher R-squared value, indicating that it explains more of the variance in BodyFat.

   iii. What is the equation relating BodyFat, Weight, and Height according to this model? Is this a linear or nonlinear equation?

BodyFat = -22.859 + 1161.973 * BMI

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

```
> ggplot(data = filteredSurvey, aes(x = BMI, y = BodyFat)) +geom_point()
+geom_smooth(method = "lm", formula = y ~ x) +labs(x = "BMI", y = "BodyFat", title = "BodyFat vs. BMI with Linear Fit")
```



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.

```
> persons_val <- data.frame(Weight = c(150, 300), Height = c(70, 70))
> predictions <- predict(model_combined, newdata = persons_val, interval = "confidence", level
= 0.99)
> predictions
       fit      lwr      upr
1 12.83068 11.42618 14.23519
2 47.62251 42.90860 52.33643
```

vi. 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 = (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?

```
> corrected_model_BMI <- lm(BodyFat ~ Corrected_BMI, data = filteredSurvey)
> summary(corrected_model_BMI)

Call:
lm(formula = BodyFat ~ Corrected_BMI, data = filteredSurvey)

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)   -22.85937    2.55264  -8.955   <2e-16 ***
Corrected_BMI   1.65271    0.09953  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
```

.

It did not change the model. The unit conversions (from pounds to kilograms and inches to meters) had a minimal impact on the relationship between BMI and BodyFat.

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 "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]

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

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

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

```
> model_combined <- lm(BodyFat ~ Corrected_BMI + AgeGroup, data = filteredSurvey)
> View(filteredSurvey)
> summary(model_combined)

Call:
lm(formula = BodyFat ~ Corrected_BMI + AgeGroup, data = filteredSurvey)

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)    -22.83443    2.45518  -9.301  < 2e-16 ***
Corrected_BMI    1.57176    0.09648  16.291  < 2e-16 ***
AgeGroupMiddle   2.61129    0.76069   3.433    7e-04 ***
AgeGroupOlder    5.30741    1.10755   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
```

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

2 for AgeGroup Middle and Older

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

The R-squared value is 0.57, suggesting that this model explains a substantial portion of the variance in BodyFat. Compared to the previous models, this model seems to be performing better based on the higher R-squared value, indicating a better fit to the data.

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

For the Young AgeGroup:
BodyFat = (-22.83443) + (1.57176) * BMI
For the Middle AgeGroup:
BodyFat = (-22.83443 + 2.61129) + (1.57176) * BMI
For the Older AgeGroup:
BodyFat = (-22.83443 + 5.30741) + (1.57176) * 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]

```
> predictions <- data.frame(Corrected_BMI = rep(seq(min(filteredSurvey$Corrected_BMI), max(filteredSurvey$Corrected_BMI), length.out = 100), 2), AgeGroup = rep(c("Middle", "Older"), each = 100))
> predictions$BodyFat <- predict(model_combined, newdata = predictions)
> ggplot(data = filteredSurvey, aes(x = Corrected_BMI, y = BodyFat)) + geom_point() + geom_line(data = predictions, aes(group = AgeGroup, color = AgeGroup)) + labs(title = "BodyFat vs. Corrected_BMI by AgeGroup", x = "Corrected_BMI", y = "BodyFat") + theme_minimal()
```