Richard Gresham
CPSC 375-01
October 18, 2022

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**. Only one person in the group should submit to Canvas.
Due: check on Canvas.

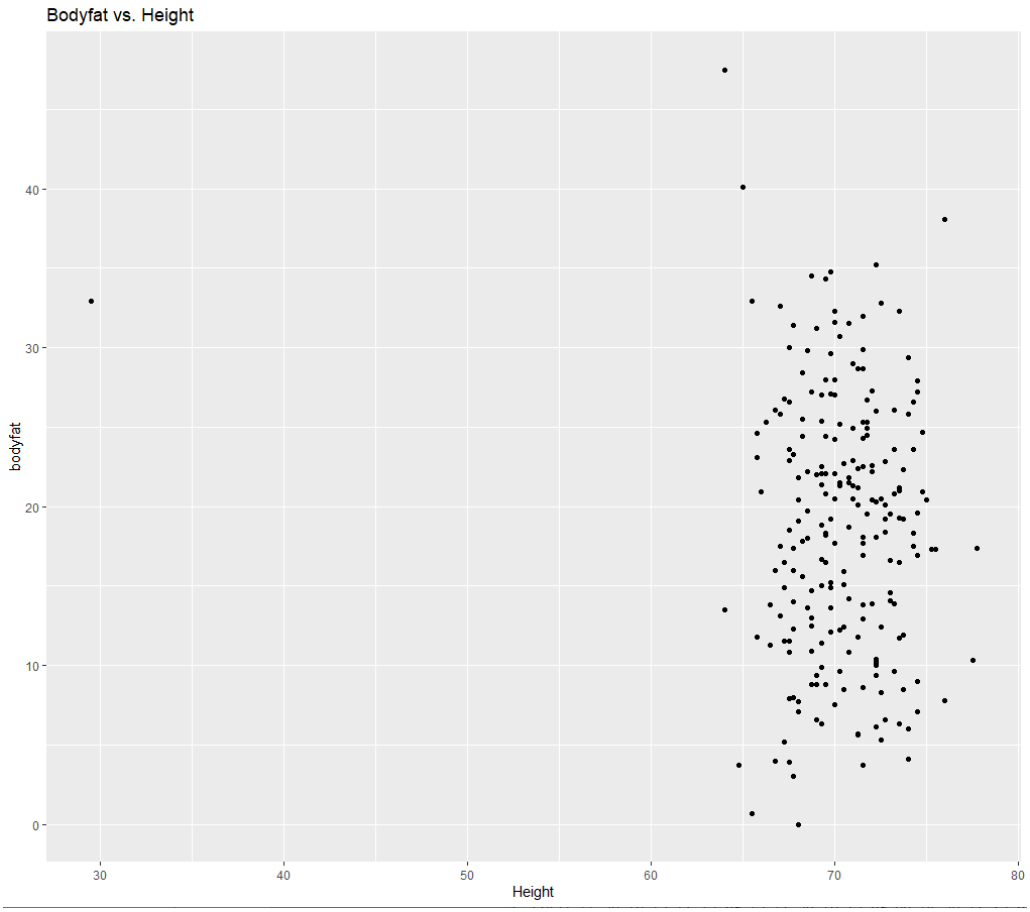
Body fat percentage refers to the relative proportions of body weight in terms of lean body mass (muscle, bone, internal organs, and connective tissue) and body fat. The most accurate means of estimating body fat percentage are cumbersome and require specialized equipment. Instead, we can estimate bodyfat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

Consider this dataset of 13 measurements from subjects (all men) along with their bodyfat percentage[1]:
`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")`

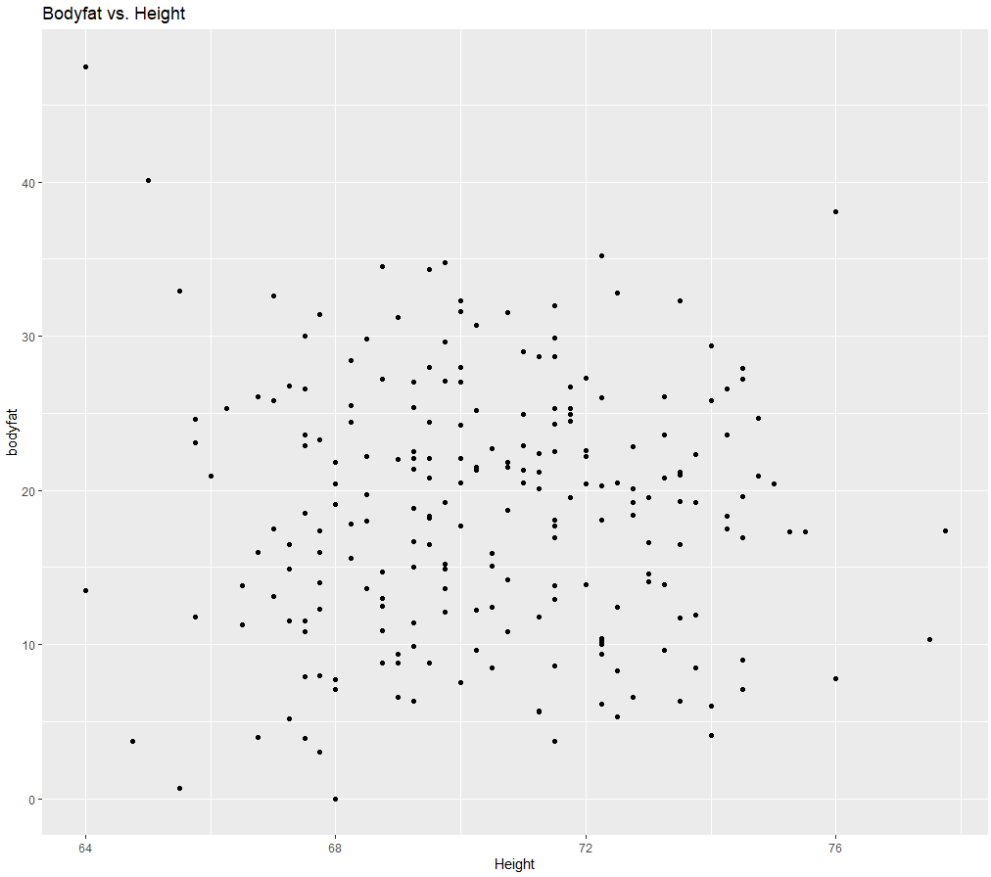
Read the data file and answer the following questions.

a. Plot `bodyfat` vs. `Height` (code, plot) Which is the dependent variable? Which is the independent variable?
   Body fat would be the dependent variable, and Height would be the independent varirable.
   ggplot(data = demo) + geom_point(mapping = aes(x = Height, y = bodyfat)) +
   ggtitle("Bodyfat vs. Height")

[1] More information about the dataset is here:
http://staff.pubhealth.ku.dk/~tag/Teaching/share/data/Bodyfat.html



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.
demo <- demo %>% filter(Height > 30)
ggplot(data = demo) + geom_point(mapping = aes(x = Height, y = bodyfat)) +
ggtitle("Bodyfat vs. Height")



Check mean:

demo %>% summarise(mean(Height))

[1] 70.456447

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

```
demo_regression <- lm(bodyfat ~ Height, data = demo)
summary(demo_regression)

Call:
lm(formula = bodyfat ~ Height, data = demo)

Residuals:
    Min      1Q  Median      3Q     Max
-18.8470  -6.5239   0.3695   5.6800  28.3489

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.01597   15.02399   1.599    0.111
Height      -0.07601    0.21309  -0.357    0.722

Residual standard error: 8.318 on 222 degrees of freedom
Multiple R-squared:  0.0005729,    Adjusted R-squared:  -0.003929
```

F-statistic: 0.1272 on 1 and 222 DF, p-value: 0.7216

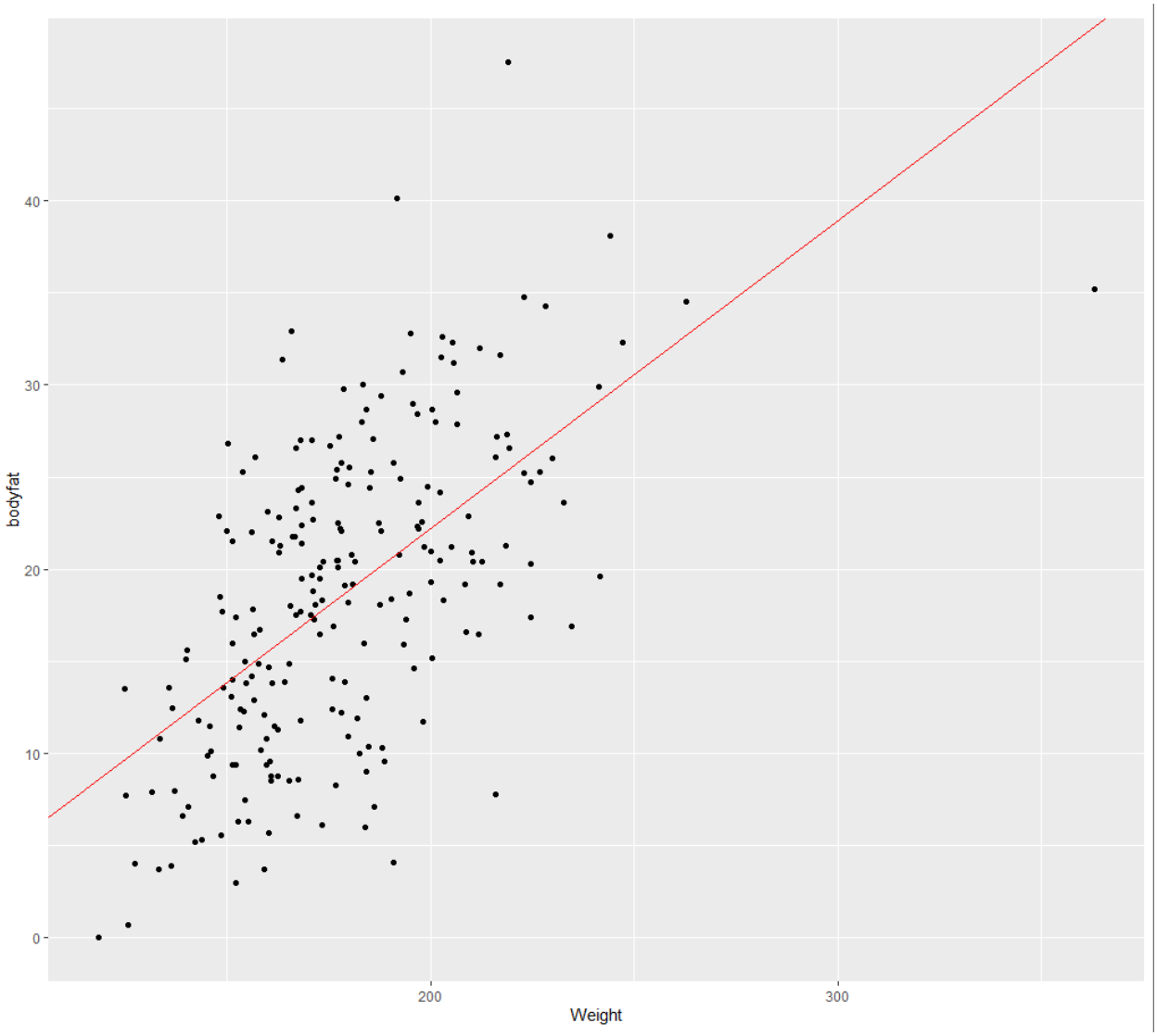
i. What is the R2 value?
The R2 value is .0005729
ii. Is this a “good” model? Why or why not?
By looking between the scatterplot and the R2 we can conclude that there is no good correlation between Height and Bodyfat.
iii. What is the linear equation relating bodyfat and Height according to this model?
bodyfat = 24.01597 – 0.07601*Height

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

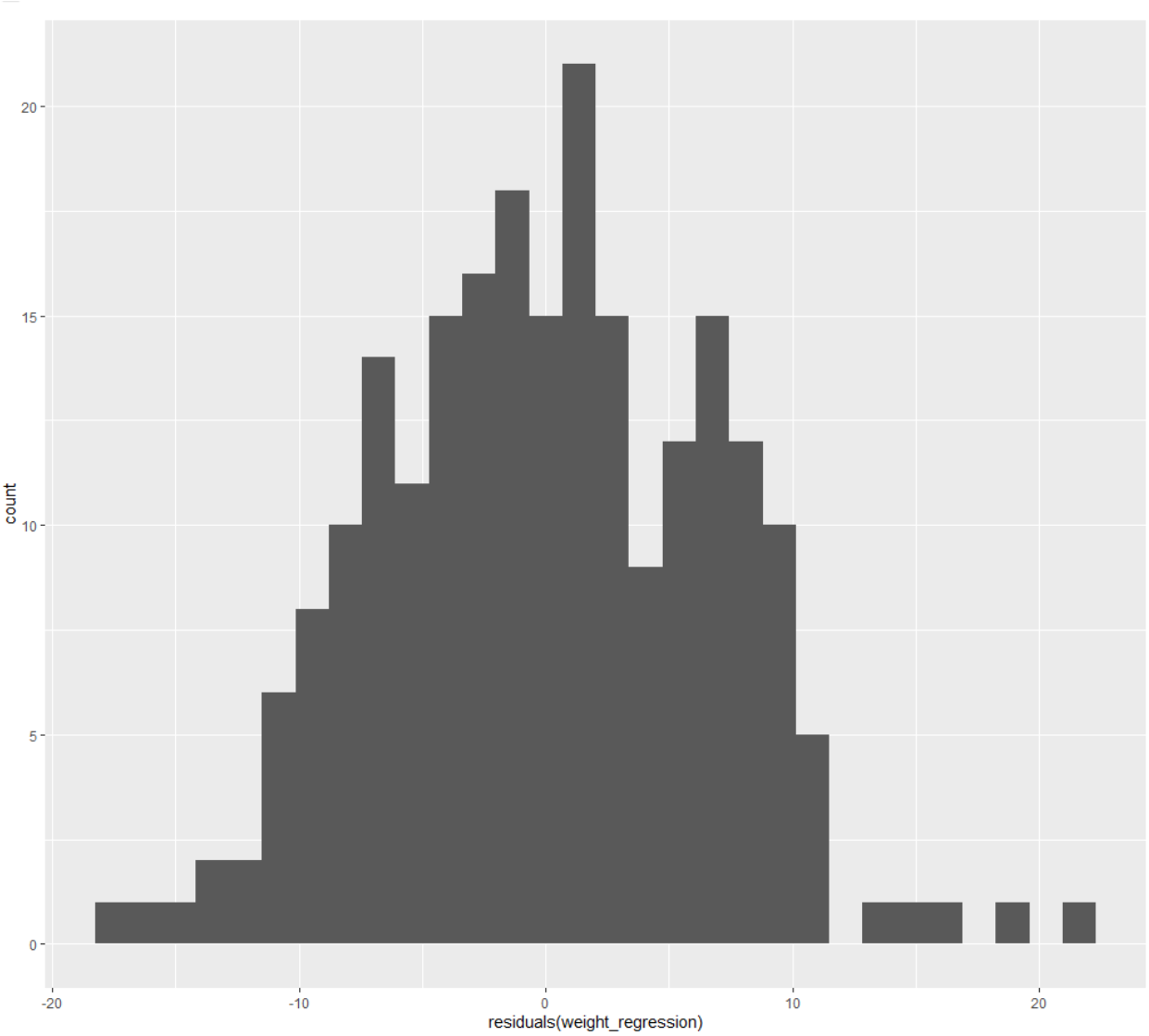
```
weight_regression <- lm(bodyfat ~ Weight, data = demo)
> summary(weight_regression)

Call:
lm(formula = bodyfat ~ Weight, data = demo)

Residuals:
    Min      1Q  Median      3Q     Max
-17.1195  -4.7496   0.0143   5.0708  22.0792

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.1727     2.7677  -4.037 7.46e-05 ***
Weight        0.1671     0.0153  10.923  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.711 on 222 degrees of freedom
Multiple R-squared:  0.3496,  Adjusted R-squared:  0.3466
F-statistic: 119.3 on 1 and 222 DF,  p-value: < 2.2e-16
```

i. What is the R2 value?
The R2 value is .3496 in this problem.
ii. Is this a better model than that based on Height? Why or why not?
This is a better model than that based on Height as the R2 value is much higher than on Height, since it is closer to 1. This 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 = -11.1727 + 0.1671 * Weight
iv. Plot bodyfat vs. Weight and overlay the best fit line. Use a different color for the line. (plot, code)

```
cf <- coef(weight_regression)
> ggplot(data = demo) + 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 = demo) + geom_histogram(mapping = aes(x = residuals(weight_regression)))



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(weight_regression, predx, interval = "confidence", level = .99)

	fit	lwr	upr
1	13.89134	12.26536	15.51733
2	38.95541	33.98944	43.92138

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))

WH_regression <- lm(bodyfat ~ Weight + Height, data = demo)

> summary(WH_regression)

```
Call:
lm(formula = bodyfat ~ Weight + Height, data = demo)

Residuals:
    Min      1Q  Median      3Q     Max
-22.564  -3.945   0.120   4.480  13.334

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 71.26119   11.35708   6.275 1.83e-09 ***
Weight       0.22460    0.01574  14.269  < 2e-16 ***
Height      -1.31572    0.17689  -7.438 2.24e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.015 on 221 degrees of freedom
Multiple R-squared:  0.4798,  Adjusted R-squared:  0.4751
F-statistic: 101.9 on 2 and 221 DF,  p-value: < 2.2e-16
```

i. What is the R2 value?
   The R2 value is .4798
ii. Is this a better model than that based only on Weight or Height? Why or why not?
   Yes it is better than basing it off of only Weight or Height as it takes in more variables that can correlate to our bodyfat dependent variable.
iii. What is the linear equation relating bodyfat, Weight, and Height according to this model?
   bodyfat = 71.26119 + .22460*Weight + -1.31572 * 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(WH_regression, predx, interval = "confidence", level = .99)
       fit       lwr       upr
1 12.85066 11.34853 14.35279
2 46.54079 41.36051 51.72107
```

   In this prediction I would be more confident in Person A vs B as they have the smaller interval range with a 99% chance that the data falls in this interval.

f. 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)

```
demo <- demo %>% mutate(BMI = (Weight/(Height*Height)))
```

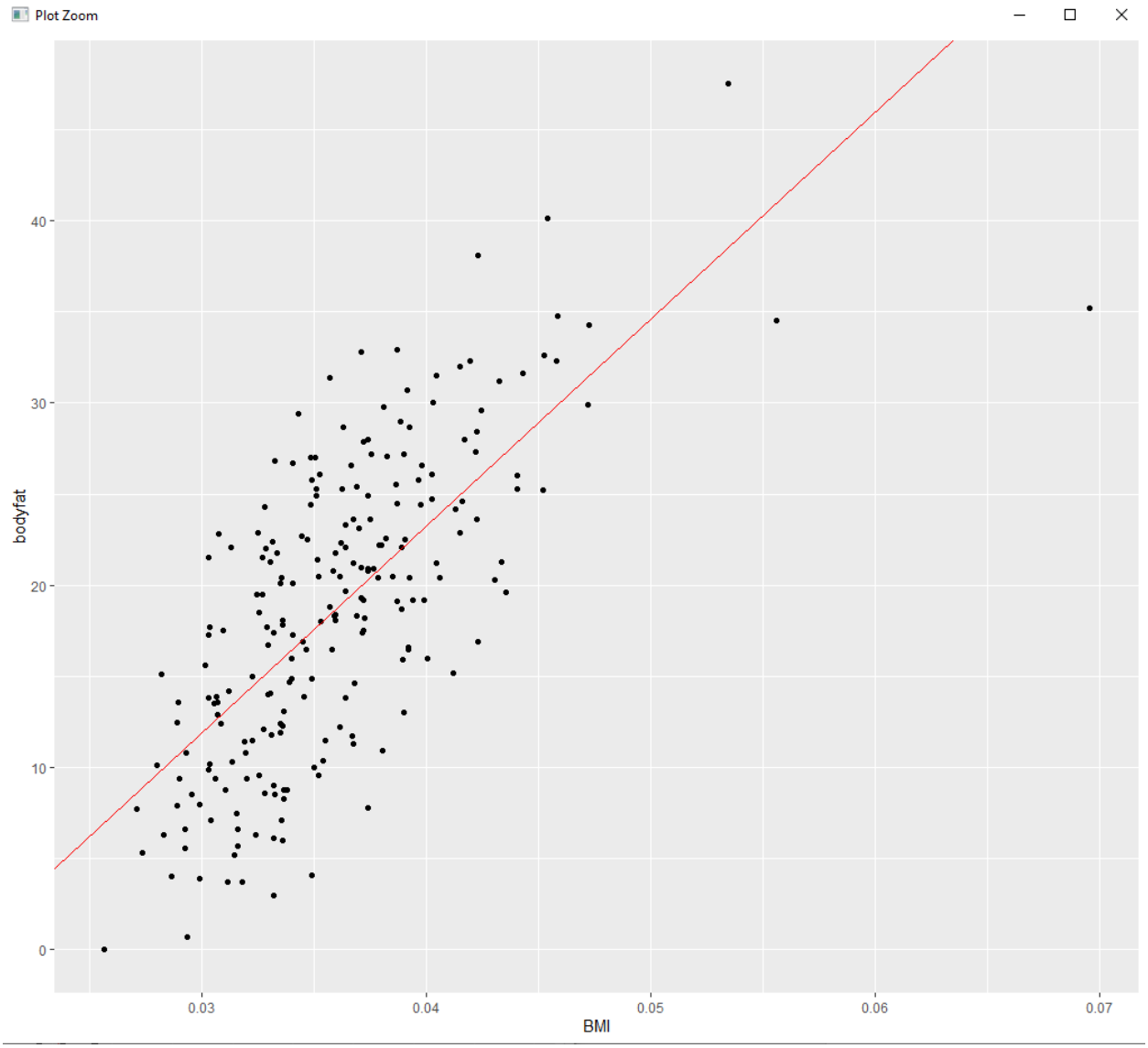
```
BMI_regression <- lm(bodyfat ~ BMI, data = demo)
> summary(BMI_regression)

Call:
lm(formula = bodyfat ~ BMI, data = demo)

Residuals:
    Min      1Q  Median      3Q     Max
-21.6522  -3.7217   0.1273   4.1130  12.9958

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -22.065      2.775  -7.951 9.32e-14 ***
BMI         1134.388     76.517  14.825  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.898 on 222 degrees of freedom
Multiple R-squared:  0.4975,  Adjusted R-squared:  0.4952
F-statistic: 219.8 on 1 and 222 DF,  p-value: < 2.2e-16
```

ii. Is this a better model than the previous models? Why or why not?
With a R2 value of .4975 this model is the best among the previous ones as it is the closest one to 1.

iii. What is the equation relating bodyfat, Weight, and Height according to this model? Is this a linear or nonlinear equation?
bodyfat = -22.065 + 1134.388 * BMI

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

```
cf <- coef(BMI_regression)
ggplot(data = demo) + 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.

```
Person1_BMI <- (150/(70*70))
 Person2_BMI <- (300/(70*70))
 predx <- data.frame(BMI = c(Person1_BMI, Person2_BMI))
 predict(BMI_regression,predx,interval = "confidence", level = .99)
       fit      lwr      upr
1 12.66115 11.19357 14.12874
2 47.38732 42.24980 52.52484
```

In this example I would predict that person A to be more confident as the intervals is smaller.

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?
Yes the model would be non-linear but different as the ratio of height and weight aren't the same. It might 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]
col <- cut(demo$Age, breaks = c(-Inf, 40,60,Inf), labels = c("Young", "Middle", "Older"))
demo <- demo %>% mutate(AgeGroup = col)

ii. Create a linear model of `bodyfat` vs. **`BMI and AgeGroup.`**[Code, output of summary(model)]
BMIage_regression <- lm(bodyfat ~ BMI + AgeGroup, data = demo)
> summary(BMIage_regression)

```
Call:
lm(formula = bodyfat ~ BMI + AgeGroup, data = demo)

Residuals:
    Min      1Q  Median      3Q     Max
-21.3977  -3.9927  -0.0714   4.1195  12.0094

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -22.7548     2.6645  -8.540 2.23e-15 ***
BMI            1102.7942    73.8328  14.936  < 2e-16 ***
AgeGroupMiddle    2.6332     0.7976   3.302  0.00112 **
AgeGroupOlder     6.0218     1.4589   4.128 5.20e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.654 on 220 degrees of freedom
Multiple R-squared:  0.5425,  Adjusted R-squared:  0.5363
F-statistic: 86.97 on 3 and 220 DF,  p-value: < 2.2e-16
```

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

bodyfat = -22.7548 + 1102.7942*BMI + 2.6332*AgeGroupMiddle + 6.0218*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]) takes 3 age groups, 3 -1 = 2.

iv. Is this a better model than the previous models? Why or why not?
I would probably say yes it is a better model than the previous as its more specific on its target in the data for each AgeGroup and R2 = .5654 which is the highest one here.

v. What are the set of equations relating bodyfat, BMI, and AgeGroup according to this model?
bodyfat = -22.7548 + 1102.7942*BMI
bodyfat = -22.7548 + 1102.7942*BMI + 2.6332*AgeGroupMiddle
bodyfat = -22.7548 + 1102.7942*BMI + 6.0218*AgeGroupolder

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

```
ggplot(data = predict2) + geom_point(mapping = aes(x = BMI, y = bodyfat, color = AgeGroup)) + geom_line(mapping =aes(x = BMI, y = pred, color = AgeGroup))
```


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