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

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.

c. Create a linear model of `bodyfat` vs. `Height`. (code, output of summary(model))
   1. What is the R2 value?
      The R2 value is 0.008009
   2. Is this a "good" model? Why or why not?
      This is not a good model as the best line of fit is not optimal.
   3. What is the linear equation relating bodyfat and Height according to this model?
      Bodyfat = 33.4945 + Height * -0.2045

d. Create a linear model of `bodyfat` vs. **`Weight`**. (code, output of summary(model))
   1. What is the R2 value?
      The R2 value is 0.3751
   2. Is this a better model than that based on Height? Why or why not?
      This model is better than the model based on Height as the best line of fit is optimal.
   3. What is the linear equation relating bodyfat and Weight according to this model?
      Bodyfat = -12.05158 + Weight * 0.17439
   4. Plot bodyfat vs. Weight and overlay the best fit line. Use a different color for the line. (plot, code)
   5. Plot the histogram of residuals (plot, code). Does this show an approximately normal distribution?
      This does show an approximate normal distribution
   6. 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?
      I'm more confident in Person B as the fit value is 40.3, lower value is 35.7 and upper value is 44.9. Whereas for Person a the fit value is 14.1, lower value is 12.58 and upper value is 15.6.

e. Create a linear model of `bodyfat` vs. **`Weight and Height`**. (code, output of summary(model))
   1. What is the R2 value?
      The R2 value is 0.4606.
   2. Is this a better model than that based only on Weight or Height? Why or why not?
      This model is better than the models only on weight or height because it is fitted much better
   3. What is the linear equation relating bodyfat, Weight, and Height according to this model?
      Bodyfat = 32.404646 + 0.20138*Weight – 0.70260*Height

4. 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?

I’m more confident in Person B as the fit value is 43.6, lower value is 39.1 and upper value is 48.1. Whereas for Person a the fit value is 13.4, lower value is 12 and upper value is 14.9.

f. Add a new transformed variable **BMI** = **Weight/Height**^{2} to the dataset. Create a linear model of `bodyfat` vs. **`BMI`**.

1. Give R code, output of summary(model)
2. Is this a better model than the previous models? Why or why not?
   This model is not better than the previous model as there is a systematic behavior in the residuals.
3. What is the equation relating bodyfat, Weight, and Height according to this model? Is this a linear or nonlinear equation?
   Bodyfat = 10.715 + BMI * 228.616. According to this model, this seems to be a linear equation
4. Plot `bodyfat` vs. `BMI` and overlay the best fit model as a straight line. (code, plot)
5. 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.

I’m more confident in Person B as the fit value is 43.6, lower value is 39.1 and upper value is 48.1. Whereas for Person a the fit value is 13.4, lower value is 12 and upper value is 14.9.

6. 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?
   No this would not result in a different model from what was calculated as it would still produce a linear equation and a relative best line of fit.

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.

1. 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]
2. Create a linear model of `bodyfat` vs. **`BMI and AgeGroup.`**[Code, output of summary(model)]
3. How many dummy (i.e., 0-1) variables were created in the model?
4. Is this a better model than the previous models? Why or why not?
5. What are the set of equations relating bodyfat, BMI, and AgeGroup according to this model?
6. Plot `bodyfat` vs. `BMI` and overlay the model predictions (**multiple** lines: one for each value of the discrete variable). [Code, plot]

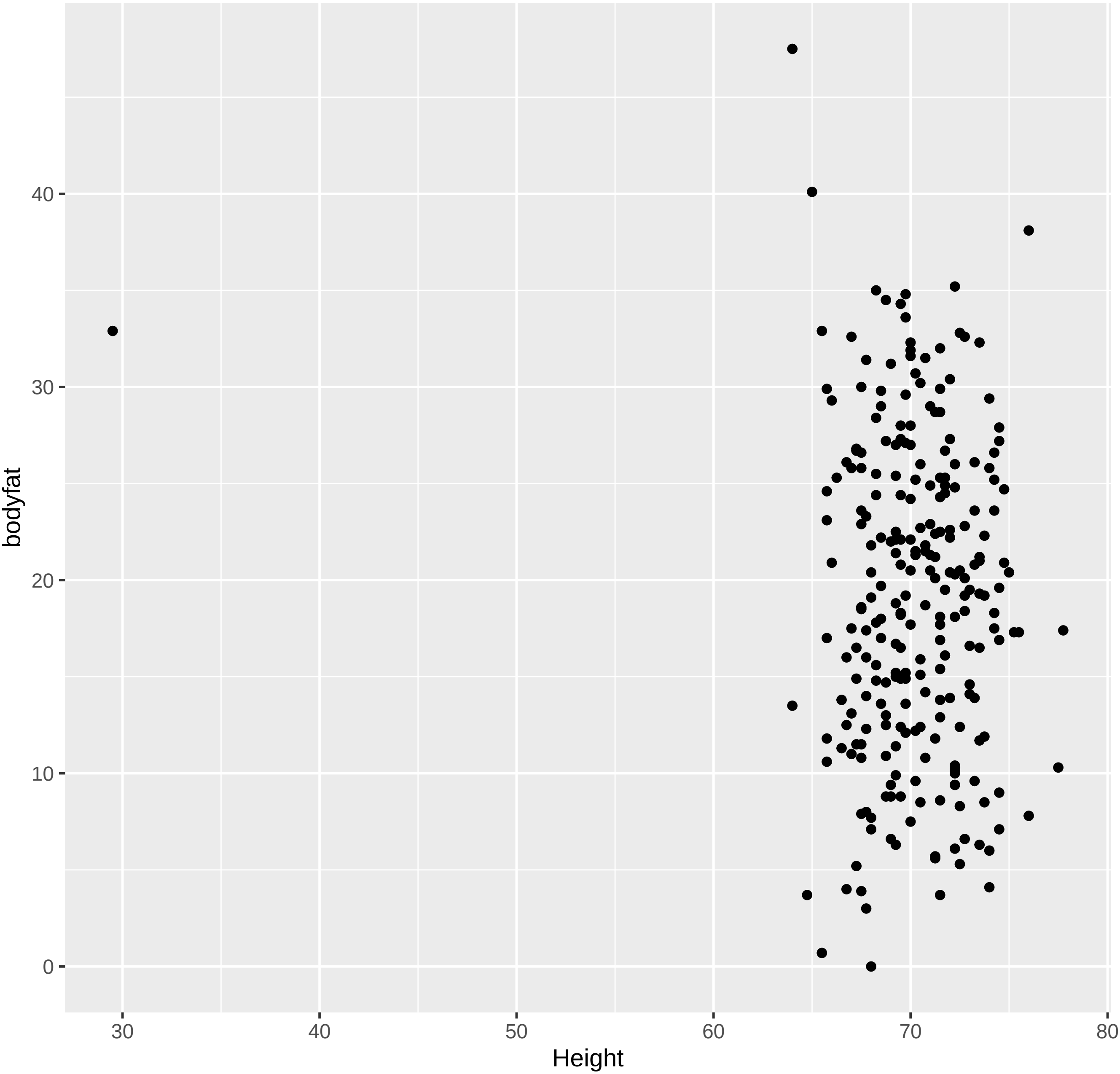
```
Call:
lm(formula = bodyfat ~ Height, data = Bodyfat)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.5902  -6.7124   0.3966   6.0716  27.0919 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  33.4945    10.1096   3.313  0.00106 **
Height       -0.2045     0.1439  -1.421  0.15664   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.352 on 250 degrees of freedom
Multiple R-squared:  0.008009,  Adjusted R-squared:  0.004041 
F-statistic: 2.019 on 1 and 250 DF,  p-value: 0.1566

Saving 7 x 7 in image

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

Residuals:
     Min       1Q   Median       3Q      Max 
-17.8164  -4.7430   0.0746   4.9283  21.3605 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -12.05158    2.58139  -4.669 4.95e-06 ***
Weight        0.17439    0.01424  12.249  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.629 on 250 degrees of freedom
Multiple R-squared:  0.3751,    Adjusted R-squared:  0.3726 
F-statistic:   150 on 1 and 250 DF,  p-value: < 2.2e-16

`geom_smooth()` using formula 'y ~ x'
Saving 7 x 7 in image
`geom_smooth()` using formula 'y ~ x'
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
Saving 7 x 7 in image
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
       fit      lwr      upr
1 14.10671 12.58441 15.62901
       fit      lwr      upr
1 40.26499 35.66127 44.86872

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

Residuals:
     Min       1Q   Median       3Q      Max 
-20.0617  -4.1356   0.0361   4.6262  15.9585 

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.40464    7.46992   4.338 2.09e-05 ***
Weight       0.20138    0.01393  14.455  < 2e-16 ***
Height      -0.70260    0.11178  -6.285 1.45e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.171 on 249 degrees of freedom
Multiple R-squared:  0.4606,    Adjusted R-squared:  0.4563
F-statistic: 106.3 on 2 and 249 DF,  p-value: < 2.2e-16

       fit      lwr      upr
1 13.43045 11.98608 14.87482
       fit      lwr      upr
1 43.63797 39.13165 48.14429
Saving 7 x 7 in image

Call:
lm(formula = bodyfat ~ BMI, data = Bodyfat.copy)

Residuals:
    Min      1Q  Median      3Q     Max
-31.669  -5.740  -0.169   5.968  24.561

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.715      1.421   7.540 8.66e-13 ***
BMI          228.616     36.147   6.325 1.16e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.786 on 250 degrees of freedom
Multiple R-squared:  0.1379,    Adjusted R-squared:  0.1345
F-statistic:    40 on 1 and 250 DF,  p-value: 1.162e-09

Saving 7 x 7 in image
`geom_smooth()` using formula 'y ~ x'
Saving 7 x 7 in image
`geom_smooth()` using formula 'y ~ x'
       fit      lwr      upr
1 13.43045 11.98608 14.87482
       fit      lwr      upr
1 43.63797 39.13165 48.14429
```

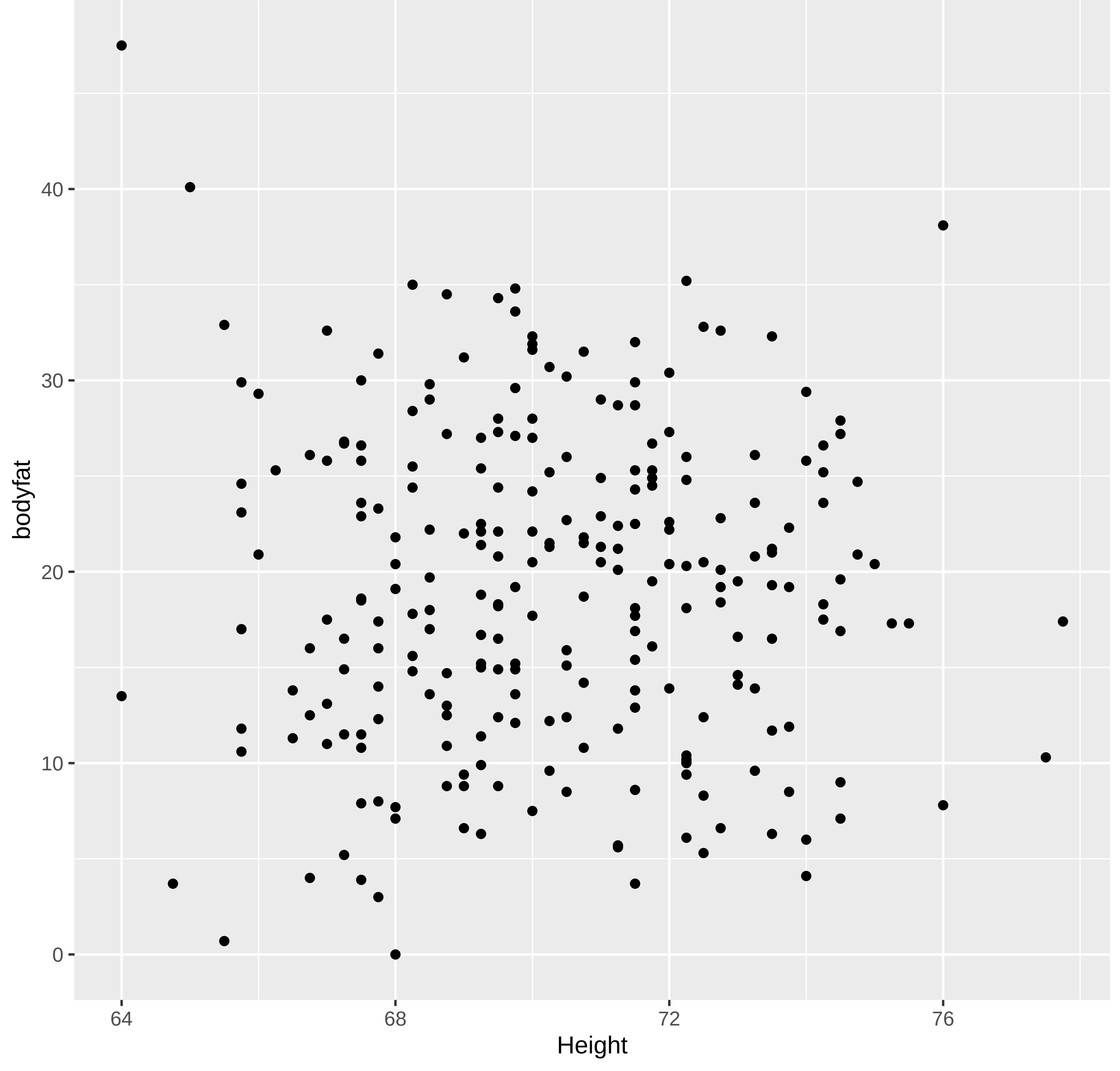
```r
library(ggplot2)
library(tidyverse)
library(modelr)

Bodyfat <- read_csv("Bodyfat.csv")

ggplot(data=Bodyfat)+geom_point(mapping=aes(y=bodyfat,x=Height)) +
  ggtitle('a. Plot bodyfat vs. Height')
ggsave("homework#6.1.pdf")

BodyfatRow <- Bodyfat[-c(42),]
ggplot(data=BodyfatRow)+geom_point(mapping=aes(y=bodyfat,x=Height)) +
  ggtitle('b. Remove the corresponding row from the data')
ggsave("homework#6.2.pdf")

m <- lm(data=Bodyfat, formula=bodyfat~Height)
summary(m)

cf <- coef(m)

ggplot(data=Bodyfat)+geom_point(mapping=aes(y=bodyfat,x=Weight)) +
  ggtitle('a. Plot bodyfat vs. Weight')
ggsave("homework#6.3.pdf")

m1 <- lm(data=Bodyfat, formula=bodyfat~Weight)
summary(m1)

ggplot(data = Bodyfat, aes(y = bodyfat,x = Weight)) +
  geom_point(aes(color = "red")) +
  geom_smooth(method = "lm") +
  labs(title = "bodyfat vs Weight best line of it",)
ggsave("homework#6.4.pdf")

hello <- residuals(m1)
ggplot(data = Bodyfat) + geom_histogram(aes(x = hello))
ggsave("homework#6.5.pdf")

new.dat <- data.frame(Weight=150)
predict(m1, newdata = new.dat, interval = 'confidence', level=0.99)

new.dat <- data.frame(Weight=300)
predict(m1, newdata = new.dat, interval = 'confidence', level=0.99)

m2 <- lm(data=Bodyfat, formula=bodyfat~Weight+Height)
summary(m2)

ggplot(data=BodyfatRow)+geom_point(mapping=aes(y=Weight,x=Height)) +
  ggtitle('e.2 Bodyfat vs Weight and Height')

new.dat <- data.frame(Weight=150, Height=70)
predict(m2, newdata = new.dat, interval = 'confidence', level=0.99)

new.dat <- data.frame(Weight=300, Height=70)
predict(m2, newdata = new.dat, interval = 'confidence', level=0.99)
ggsave("homework#6.6.pdf")

Bodyfat.copy <- Bodyfat %>% mutate(BMI = Weight/Height^2)
```

```r

m3 <- lm(data=Bodyfat.copy, formula=bodyfat~BMI)
summary(m3)

ggplot(data=Bodyfat.copy)+geom_point(mapping=aes(y=bodyfat,x=BMI)) +
  ggtitle('e.2 Bodyfat VS BMI')
ggsave("homework#6.7.pdf")

ggplot(data = Bodyfat.copy, aes(y = bodyfat,x = BMI)) +
  geom_point(aes(color = "red")) +
  geom_smooth(method = "lm") +
  labs(title = "Bodyfat VS BMI best line of it",)
ggsave("homework#6.8.pdf")

new.dat <- data.frame(Weight=150, Height=70)
predict(m2, newdata = new.dat, interval = 'confidence', level=0.99)

new.dat <- data.frame(Weight=300, Height=70)
predict(m2, newdata = new.dat, interval = 'confidence', level=0.99)

Bodyfat.copy <- Bodyfat %>% mutate(cut(Age, breaks = c(-Inf,40,60,Inf))) %>% mutate(BMI = Weight/Height^2)

m4 <- lm(data=Bodyfat.copy, formula=BMI~breaks)
summary(m4)
```

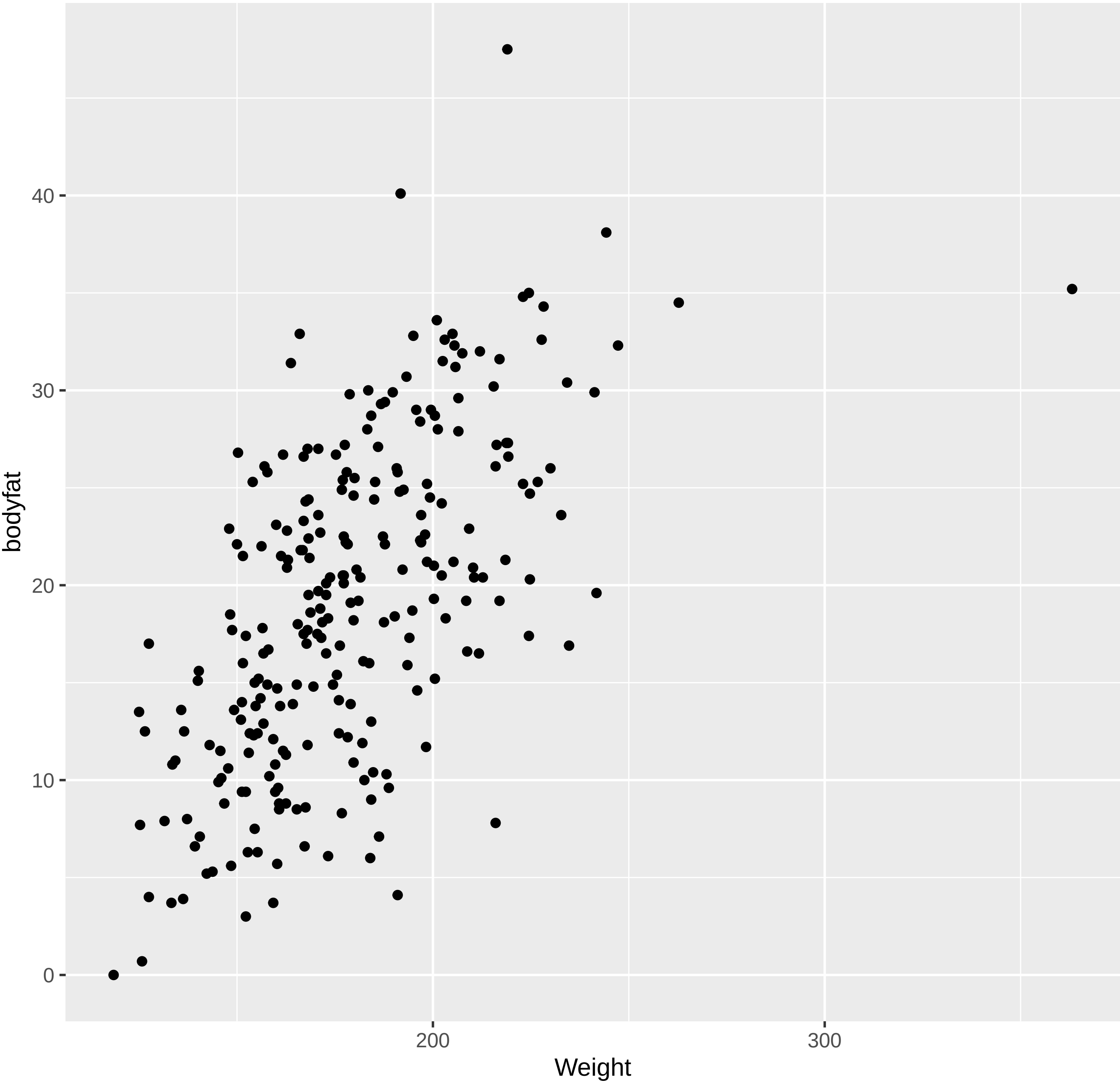
a. Plot bodyfat vs. Height

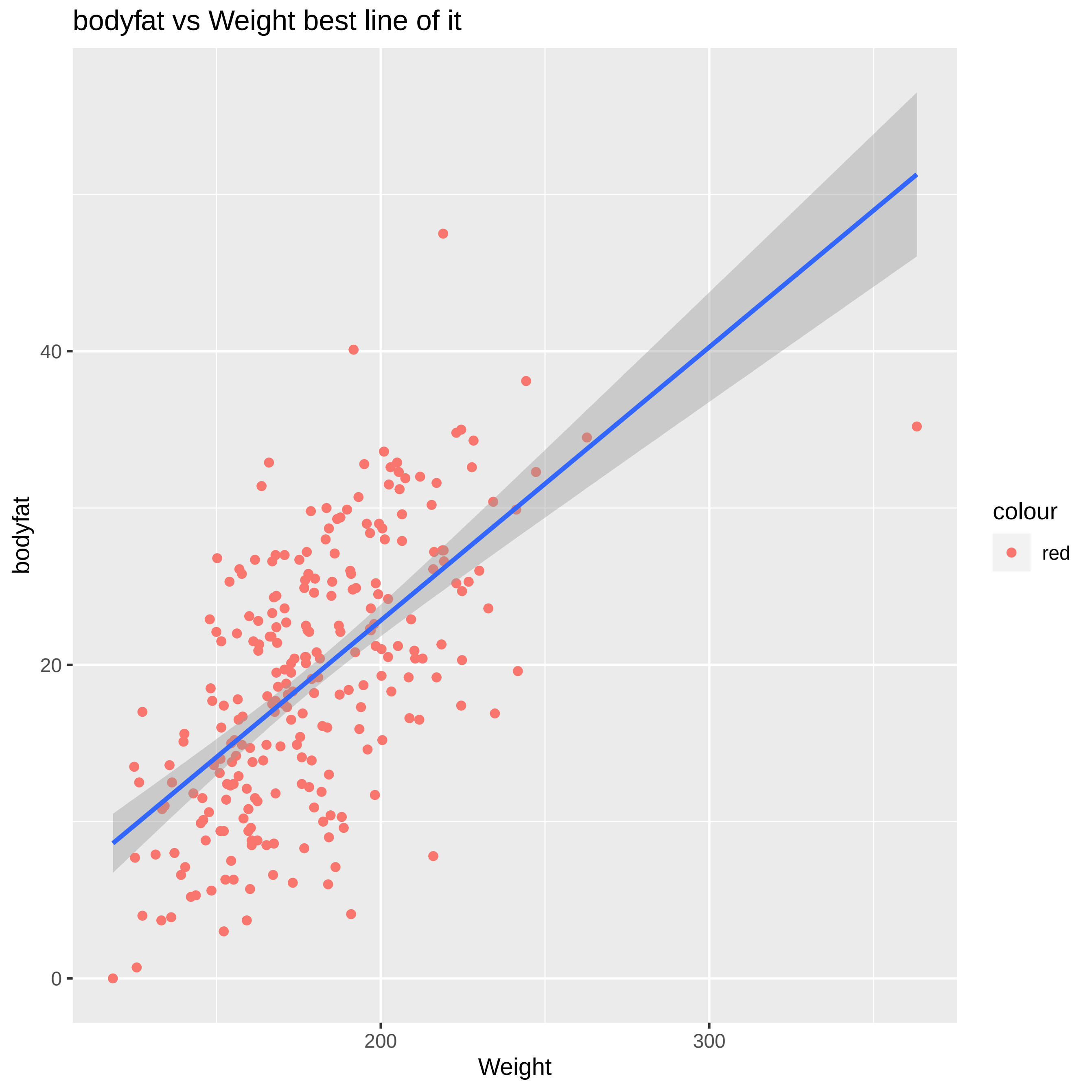


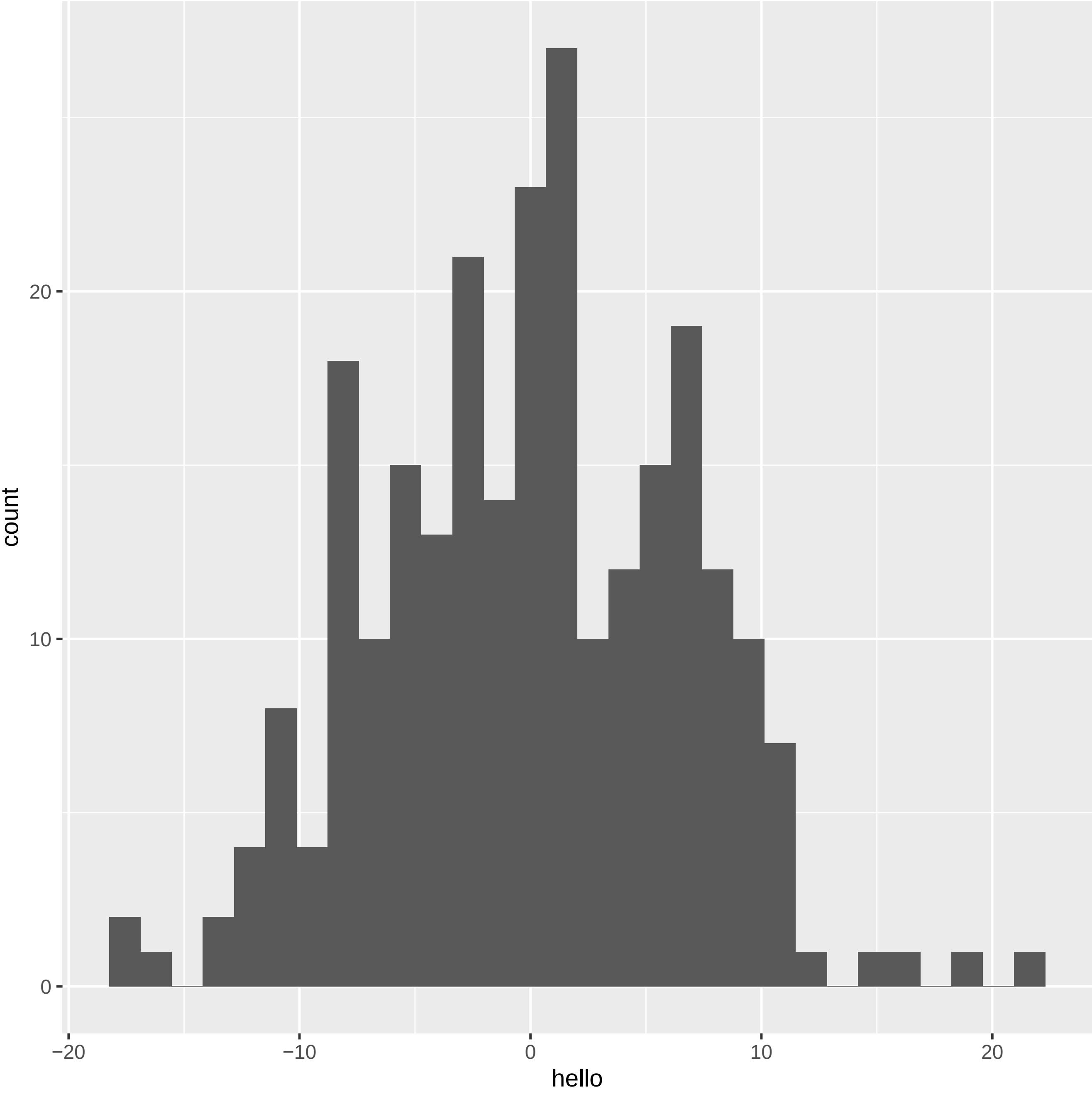
b. Remove the corresponding row from the data



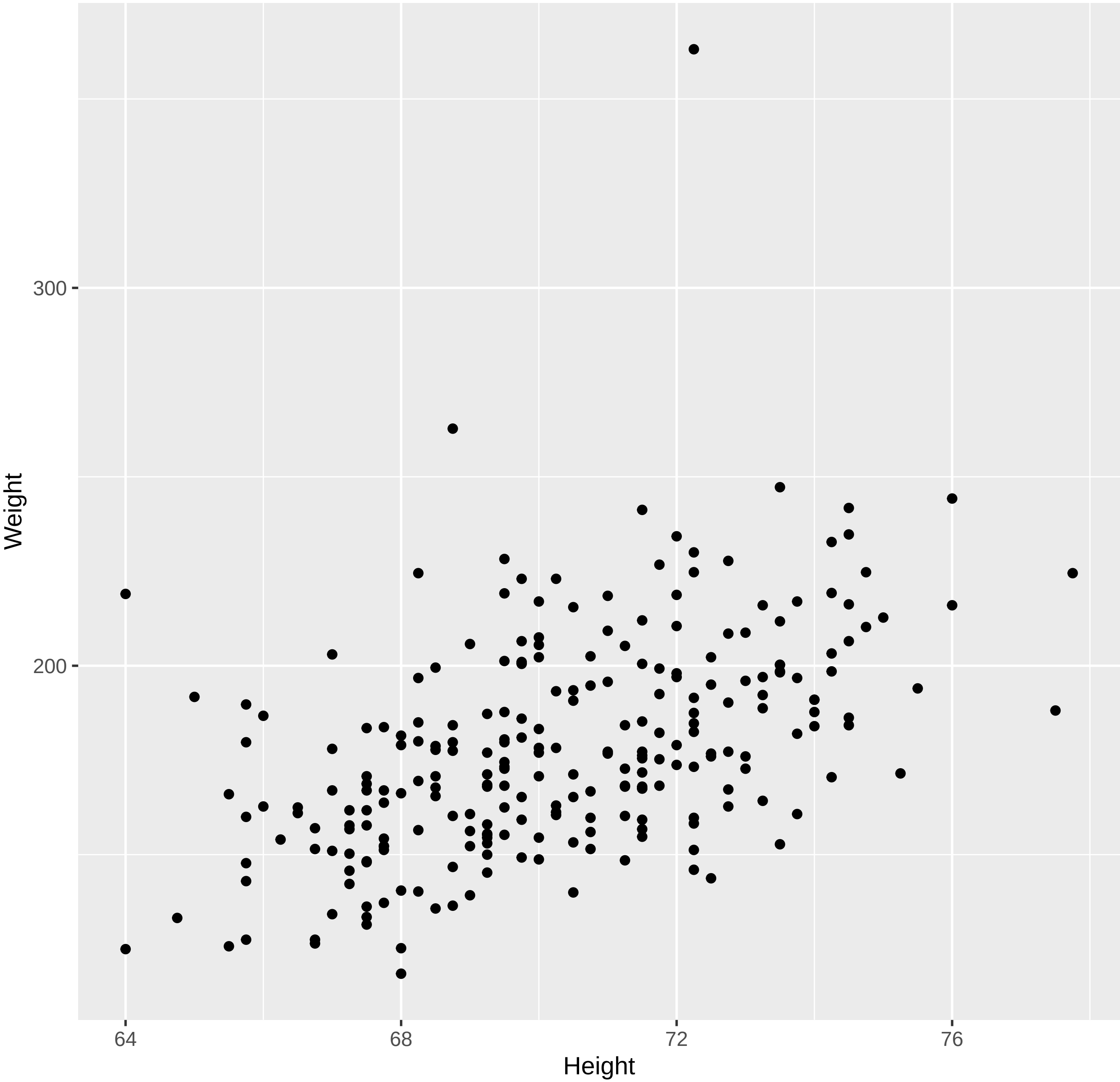
a. Plot bodyfat vs. Weight



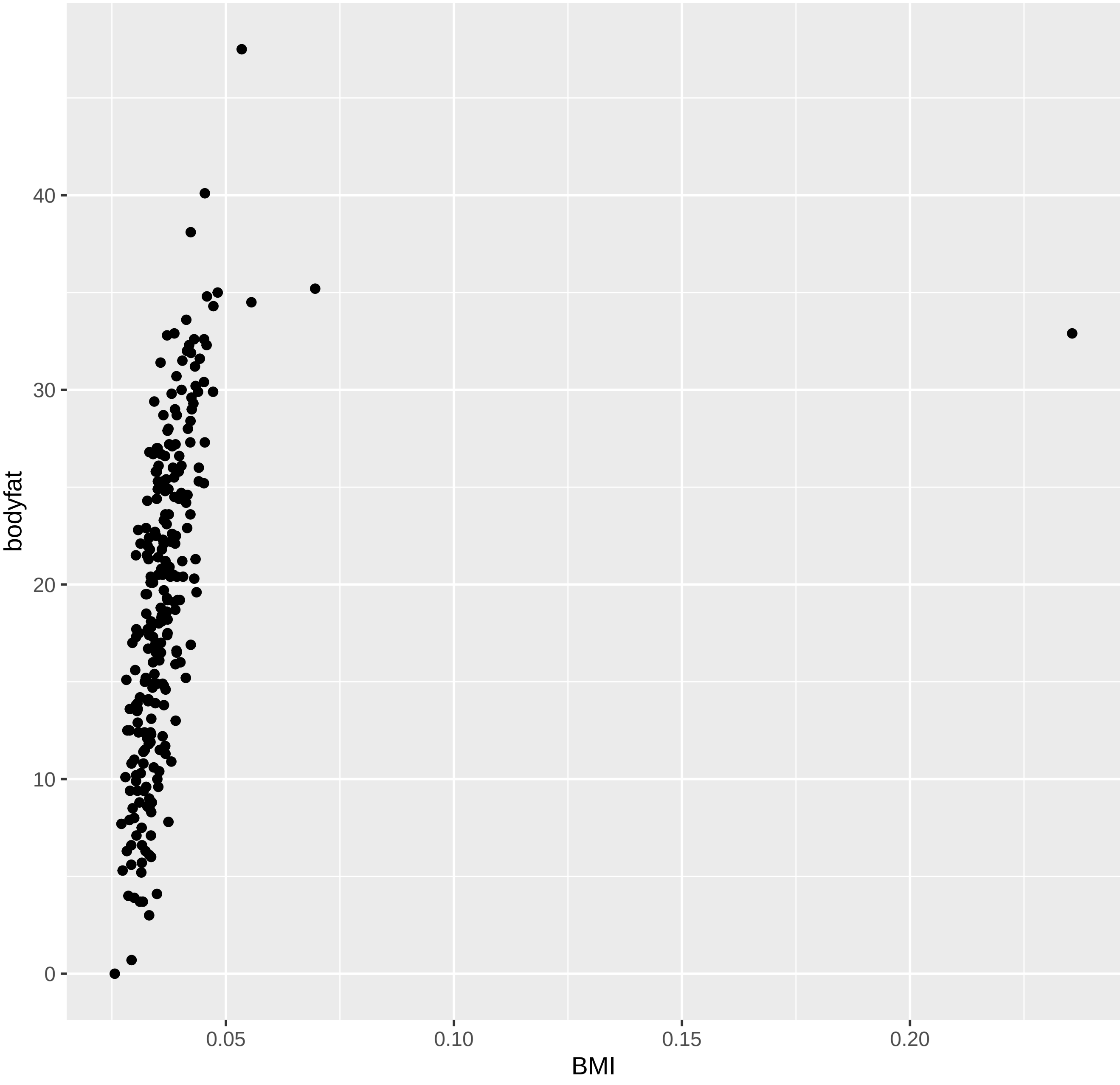

bodyfat vs Weight best line of it
bodyfat
40
20
0
200
300
Weight
colour
●
red

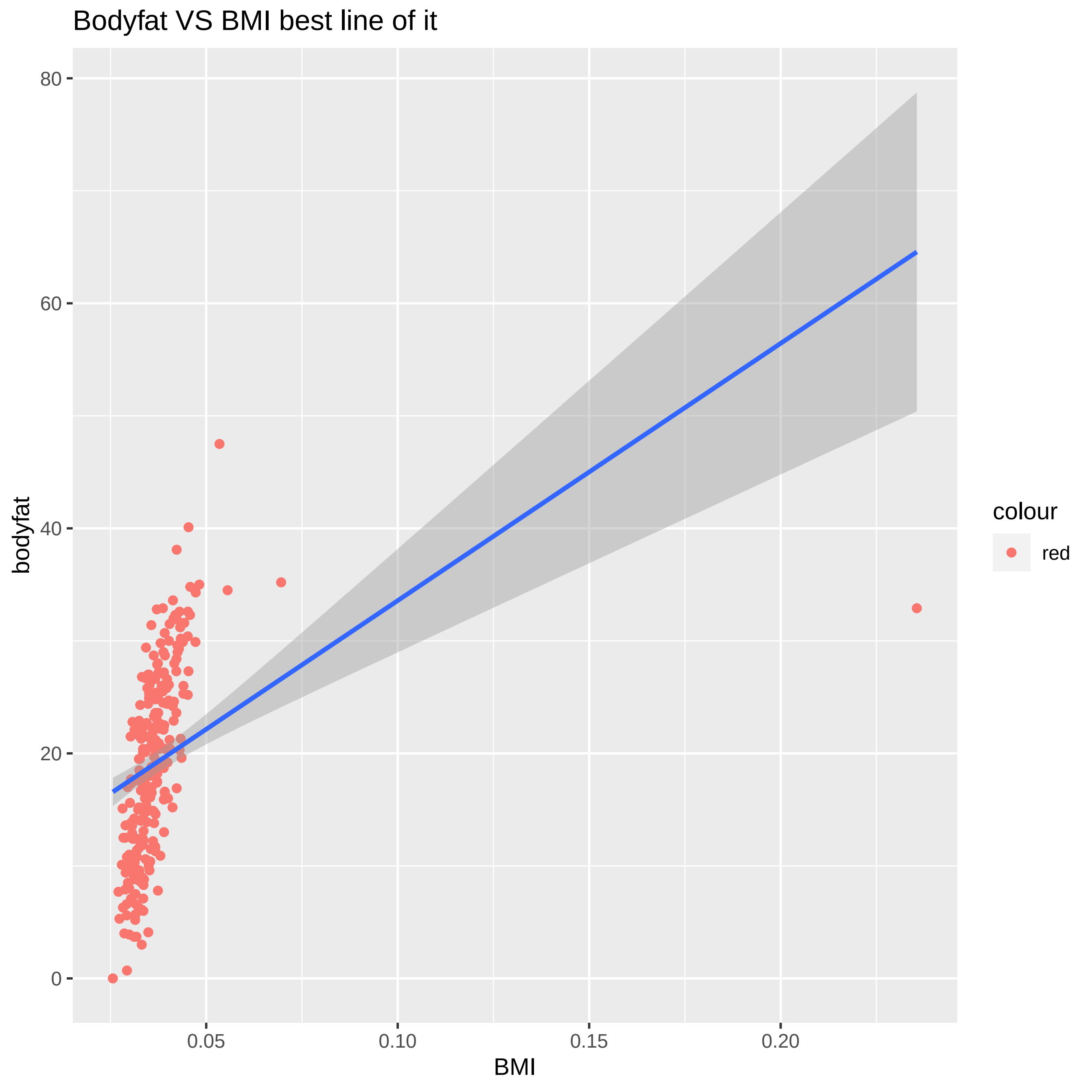

count
20
10
0
−20
−10
0
10
20
hello

e.2 Bodyfat vs Weight and Height



e.2 Bodyfat VS BMI




Bodyfat VS BMI best line of it
80
60
40
20
0
bodyfat
0.05
0.10
0.15
0.20
BMI
colour
red