State 462. Jiaqi Li

1. ① Prove $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$

for $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

we have $\varepsilon_i^2 = [y_i - (\beta_0 + \beta_1 x_i)]^2$

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

$$= \sum_{i=1}^{n} \left[y_i^2 - 2y_i(\beta_0 + \beta_1 x_i) + (\beta_0 + \beta_1 x_i)^2 \right]$$

$$= \sum_{i=1}^{n} \left[y_i^2 - 2y_i\beta_0 - 2y_i x_i \beta_1 + \beta_0^2 + 2\beta_0\beta_1 x_i + \beta_1^2 x_i^2 \right].$$

Since we want to minimize $\sum_{i=1}^{n} \varepsilon_i^2$, differentiate $\sum_{i=1}^{n} \varepsilon_i^2$ & set to zero:

$$\frac{\partial \sum_{i=1}^{n} \varepsilon_i^2}{\partial \beta_0} = \sum_{i=1}^{n} \left[-2y_i + 2\hat{\beta}_0 + 2\hat{\beta}_1 x_i \right] = 0.$$

$$-2\sum_{i=1}^{n} y_i + 2\sum_{i=1}^{n} \hat{\beta}_0 + 2\hat{\beta}_1 \sum_{i=1}^{n} x_i = 0.$$

$$\Rightarrow \sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \quad (*).$$

divide n on both sides $\Rightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$

② Prove $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$

From $(*)$, we have: $\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i$

$$= \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i). = \sum_{i=1}^{n} \hat{y}_i$$

$$\Rightarrow \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i. \quad (*)'$$

③ Prove $\sum_{i=1}^{n} e_i = 0$:

Since $e_i = y_i - \hat{y}_i$, $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{y}_i = 0$ by $(*)'$.

Thus, $\sum_{i=1}^{n} e_i = 0$.

2. $Y = \beta_0 + \varepsilon$

$\Rightarrow \varepsilon_i = Y_i - \beta_0$.

$\Rightarrow \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0)^2 = \sum_{i=1}^{n} (Y_i^2 - 2Y_i\beta_0 + \beta_0^2)$.

To minimize $\sum_{i=1}^{n} \varepsilon_i^2$, differentiate & set 0:

$$\frac{\partial \sum_{i=1}^{n} \varepsilon_i^2}{\partial \beta_0} = \sum_{i=1}^{n} (-2Y_i + 2\hat{\beta}_0) = 0.$$

$$n\hat{\beta}_0 = \sum_{i=1}^{n} Y_i$$

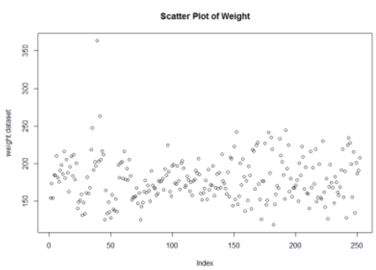
$$\hat{\beta}_0 = \overline{Y}$$

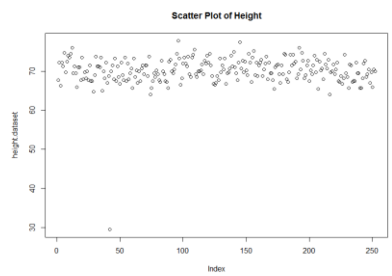
Since we can calculate $\overline{Y}$ by $\frac{1}{n}\sum_{i=1}^{n} Y_i$, $\hat{\beta}_0$ is a known quantity.

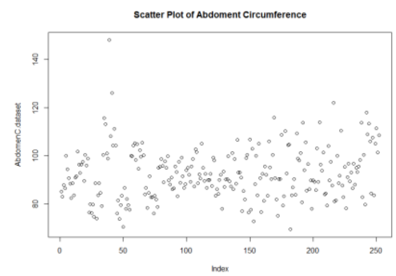
Application

A.

There is one value that is smaller than 0 but there is no value that is bigger than 100 in the recomputed dataset. So we will set the negative value to 0.
By comparing the data, there are some erroneous values in the variable "SiriBFPerc".
Thus, we employ the recomputed variable.
Check also variables "Weight", "Height", "AbdomenC" by plotting data:



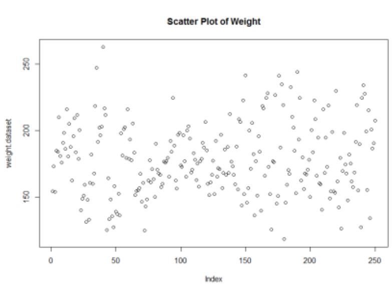


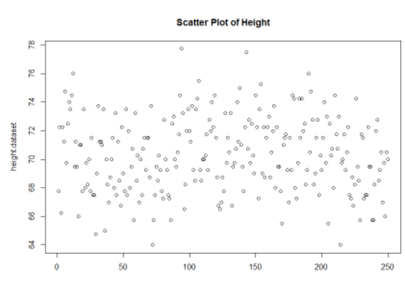


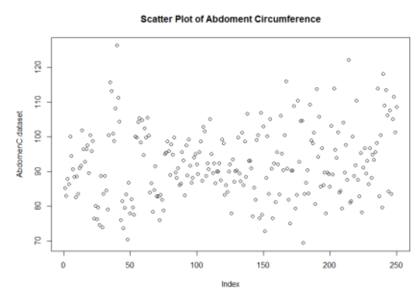
By observing the scatter plots, we can see some obvious mistakes.
Thus, we want to find these mistakes and remove them by checking

```
> bodyfat$Weight > 350
> bodyfat$AbdomenC > 140
> bodyfat$Height < 40
```

I find out that 2 observations contains those mistakes, so we remove those observations.
<u>Now, we should have a good dataset for further analysis</u>.







B.

For body fat percentage:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   12.40   19.20   18.99   25.18   47.50
```

Standard deviation = 8.357805 **IQR = 12.775**

For weight:

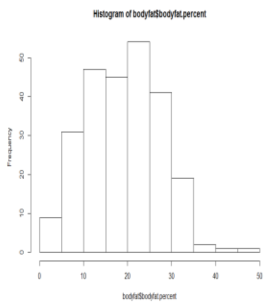
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  118.5   158.5   176.1   178.1   196.8   262.8
```

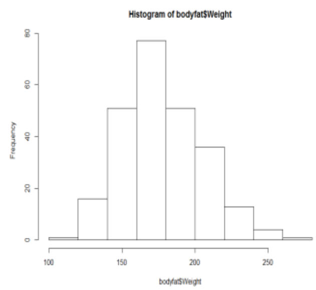
Standard deviation = 27.03549 **IQR = 38.25**

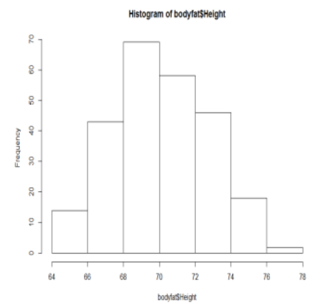
For height:

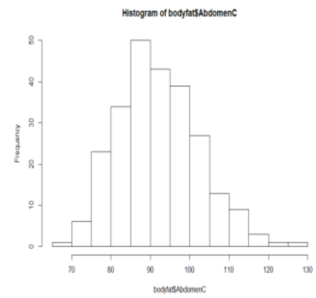
```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 64.00   68.25   70.00   70.30   72.25   77.75
```
Standard deviation = 2.616644   IQR = 4

For Abdomen Circumference:
```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 69.40   84.53   90.90   92.29   99.17  126.20
```
Standard deviation = 10.20744   IQR = 14.65

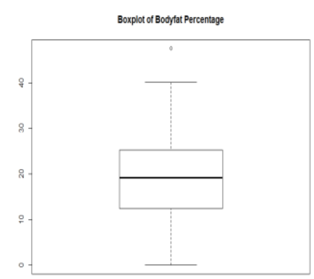


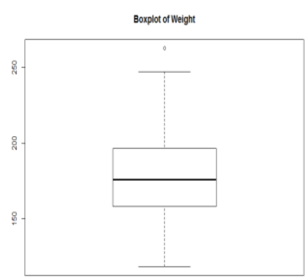


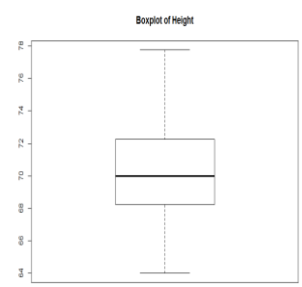


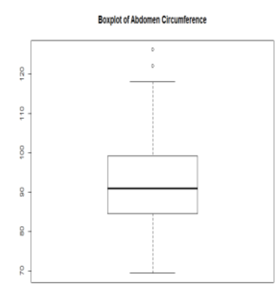


All these distributions above appear to be symmetric and bell-shaped.
The variable Weight has the largest variability while the variable Height has the smallest variability based of the standard deviation of each variables. (ie, $\sigma_{weight} > \sigma_{Abdomen\ Circumference} > \sigma_{bodyfat\ percentage} > \sigma_{height}$, same explanation for IQR)









Based on the boxplots, there are a few extreme values in distributions of Bodyfat Percentage, Weight, Abdomen Circumference.

<u>Hypothesis test (using normal distribution) on body fat percentage</u>:
H_0 : average body fat percentage ≤20%
H_1 : average body fat percentage > 20%
z-score = −1.91981   p-value = 0.972559
Thus, p-value = 0.972559 > α, accept H_0, which means the average body fat percentage does not exceed 20%.

<u>Hypothesis test (using normal distribution) on weight</u>:
H_0 : average weight ≤ 180 pounds
H_1 : average weight > 180 pounds
z-score = -1.121018   p-value = 0.8688599
Thus, p-value =0.8688599 > α, accept H_0, which means the average weight does not exceed 180 pounds.

C.

· **model employed:** $Y = \beta_0+\beta_1 X+\epsilon$

```
Correlation of body fat percentage and weight = 0.5981014
Residuals:
     Min       1Q   Median       3Q      Max
-27.1676  -4.6126   0.0375   4.9613  20.9494
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.94208    2.83363   -4.92 1.58e-06 ***
Weight        0.18490    0.01573   11.75  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.712 on 248 degrees of freedom
Multiple R-squared:  0.3577, Adjusted R-squared:  0.3551
F-statistic: 138.1 on 1 and 248 DF,  p-value: < 2.2e-16
```

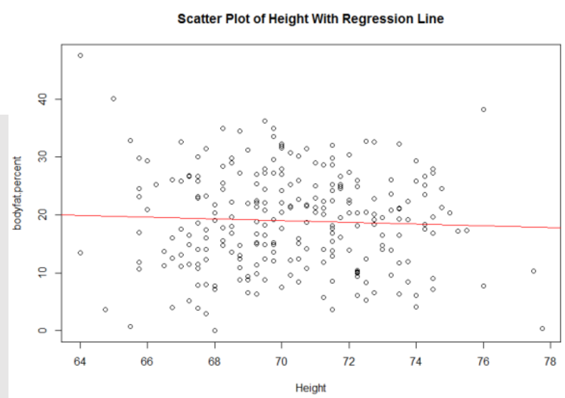
$\hat{\beta}_0 = -13.94208$ $\hat{\beta}_1 = 0.18490$ $\sigma^2 = 45.04566$ $R^2 = 0.3577$

$\hat{Y} = -13.94208 + 0.18490X$

Based on the slope of the regression line, Weight has positive relationship with body fat percentage. The slope of the regression line represents the rate of change in body fat percentage as Weight changes, its estimated value = 0.18490 can be interpreted by saying that an increase of 1 pound in Weight causes an increase of 0.18490 percent in body fat percentage.

· **model employed:** $Y = \beta_0+\beta_1 X+\epsilon$

Correlation of body fat percentage and height = -0.04854555




```
Residuals:
     Min       1Q   Median       3Q      Max
-19.3423  -6.5537   0.2821   6.2142  27.5375
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  29.8863    14.2522   2.097    0.037 *
Height       -0.1551     0.2026  -0.765    0.445
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 8.365 on 248 degrees of freedom
Multiple R-squared:  0.002357,       Adjusted R-squared:  -0.001666
F-statistic: 0.5858 on 1 and 248 DF,  p-value: 0.4448
```

$\hat{\beta}_0 = 29.8863$ $\hat{\beta}_1 = -0.1551$ $\sigma^2 = 69.96925$ $R^2 = 0.002357$

$\hat{Y} = 29.8863 + -0.1551X$

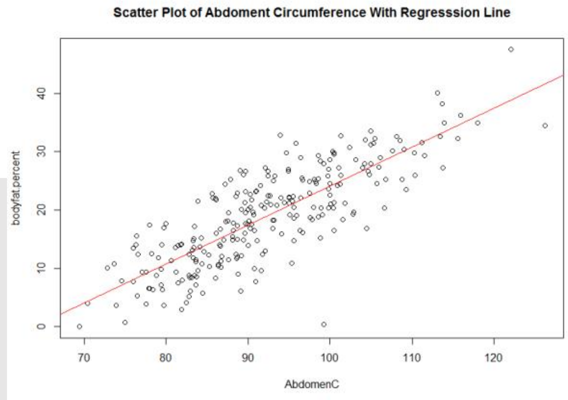
Based on the slope of the regression line, Height has negative relationship with body fat percentage. The slope of the regression line represents the rate of change in body fat percentage as Height changes, its estimated value = -0.1551 can be interpreted by saying that an increase of 1 cm in Height causes a decrease o f 0.1551 percent in body fat percentage.

· **model employed:** $Y = \beta_0+\beta_1 X+\epsilon$

Correlation of body fat percentage and Abdomen Circumference = 0.8110294

```
Residuals:
     Min       1Q   Median       3Q      Max
-23.1760  -3.5408   0.2143   3.1793  12.8435

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```




```
(Intercept) -42.29941    2.82409  -14.98   <2e-16 ***
AbdomenC      0.66407    0.03042   21.83   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.899 on 248 degrees of freedom
Multiple R-squared:  0.6578, Adjusted R-squared:  0.6564
F-statistic: 476.7 on 1 and 248 DF,  p-value: < 2.2e-16
```

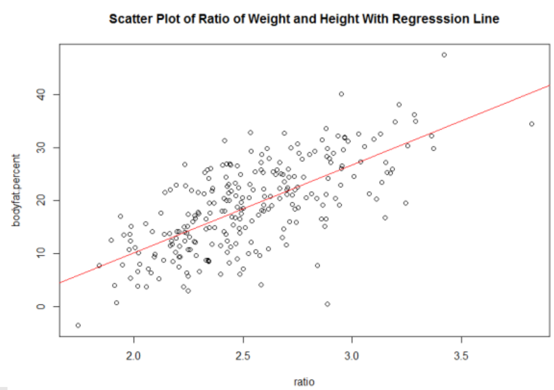
$\hat{\beta}_0$ = -42.29941 $\hat{\beta}_1$= 0.66407 σ^2=24.00225 R^2 = 0.6578
$\hat{Y}$ = -42.29941 + 0.66407X
Based on the slope of the regression line, Abdomen Circumference has positive relationship with body fat perce ntage.
The slope of the regression line represents the rate of change in body fat percentage as Abdomen Circumference changes, its estimated value = 0.66407 can be interpreted by saying that an increase of 1 cm in Abdomen Circumference causes an increase of 0.66407 percent in body fat percentage.

Since the regression between the body fat percentage and Abdomen Circumference has the largest R^2 = 0.6584, Abdomen Circumference appears to be the best predictor for body fat percentage.

D.

· Correlation of body fat percentage and ratio = 0.6852458

```
Residuals:
     Min       1Q   Median       3Q      Max
-24.4946  -4.0063   0.0771   4.1833  14.2680
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -22.746      2.843  -8.001 4.73e-14 ***
ratio         16.499      1.114  14.817  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 6.099 on 248 degrees of freedom
Multiple R-squared:  0.4696, Adjusted R-squared:  0.4674
F-statistic: 219.5 on 1 and 248 DF,  p-value: < 2.2e-16
```



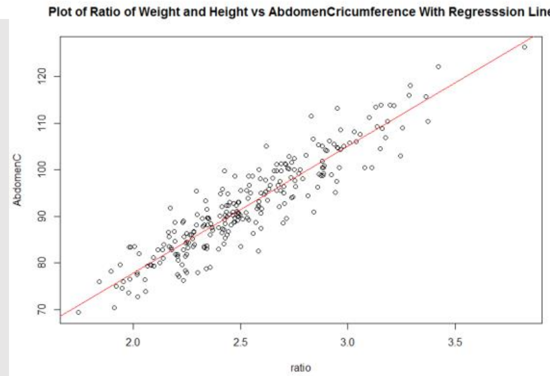
$\hat{\beta}_0$ = -22.746 $\hat{\beta}_1$= 16.499 σ^2=37.20206 R^2 = 0.4696
$\hat{Y}$ = -22.746 + 16.499X
Based on the slope of the regression line, ratio has positive relationship with body fat percentage.
The slope of the regression line represents the rate of change in body fat percentage as ratio changes, its estimated value = 16.499 can be interpreted by saying that an increase of 1 unit in ratio causes an increase of 16.499 percent in body fat percentage.

Since the value of R^2 of regression between ratio and body fat percentage is larger than that of regression between the body fat percentage and Height and that of regression between the body fat percentage and Weight, the ratio is a better predictor than weight and height.

E.

· **Correlation of Abdomen Circumference and ratio = 0.9236815**

```
Residuals:
     Min       1Q   Median       3Q      Max
-11.1934  -2.1244   0.0125   2.5932  11.0578
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.5862     1.8265   12.91   <2e-16 ***
ratio        27.1620     0.7155   37.96   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 3.919 on 248 degrees of freedom
Multiple R-squared:  0.8532, Adjusted R-squared:  0.8526
F-statistic:  1441 on 1 and 248 DF,  p-value: < 2.2e-16
```



$\hat{\beta}_0$ = 23.5862 $\hat{\beta}_1$= 27.1620 σ^2=15.35835 R^2 = 0.8532

$\hat{Y}$ = 23.5862 + 27.1620X

Based on the slope of the regression line, ratio has positive relationship with Abdomen Circumference.

The slope of the regression line represents the rate of change in Abdomen Circumference as ratio changes, its estimated value = 27.1620 can be interpreted by saying that an increase of 1 unit in ratio causes an increase of 27.1620 cm in Abdomen Circumference.

Since the regression between the ratio and Abdomen Circumference has a large R^2 = 0.8532, this regression line fits the data pretty well and ratio appears to be a very good predictor for Abdomen Circumference.

Thus, with strong correlation between Abdomen Circumference and ratio (correlation = 0.9236815, which is close to 1) and the well-fitted regression line we just addressed above, we can conclude that change in weight/height ratio and Abdomen Circumference will have similar effect to body fat percentage. Thus, weight/height ratio and Abdomen Circumference seem to “capture” the same underlying information.