Stat 462 HW4 Jiaqi Li

1. $\hat{\varepsilon} = Y - \hat{Y}$

$= Y - HY$

$= (I - H)Y$

$Var(\hat{\varepsilon}) = Var[(I-H)Y]$

$= (I-H) Var(\hat{\varepsilon}) (I-H)^T$

$= (I-H) \sigma^2 I (I-H)^T$

Since H is symmetric and idempotent, $H = H^T$ and $HH = H$ is true

Since I is symmetric, $(I-H)$ is also symmetric and idempotent, which means $(I-H) = (I-H)^T$ and $(I-H)(I-H) = (I-H)$

then, we have:

$Var(\hat{\varepsilon}) = (I-H)\sigma^2 I (I-H)$

$= \sigma^2 (I-H)(I-H)$

$= \sigma^2 (I-H)$

2. $X^T X \hat{\beta} = X^T y$

$$\Rightarrow \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 0 & -1 \\ 2 & 4 & 0 & -2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 0 & 0 \\ 1 & -1 & -2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 0 & -1 \\ 2 & 4 & 0 & -2 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \\ -1 \\ 0 \end{bmatrix}$$

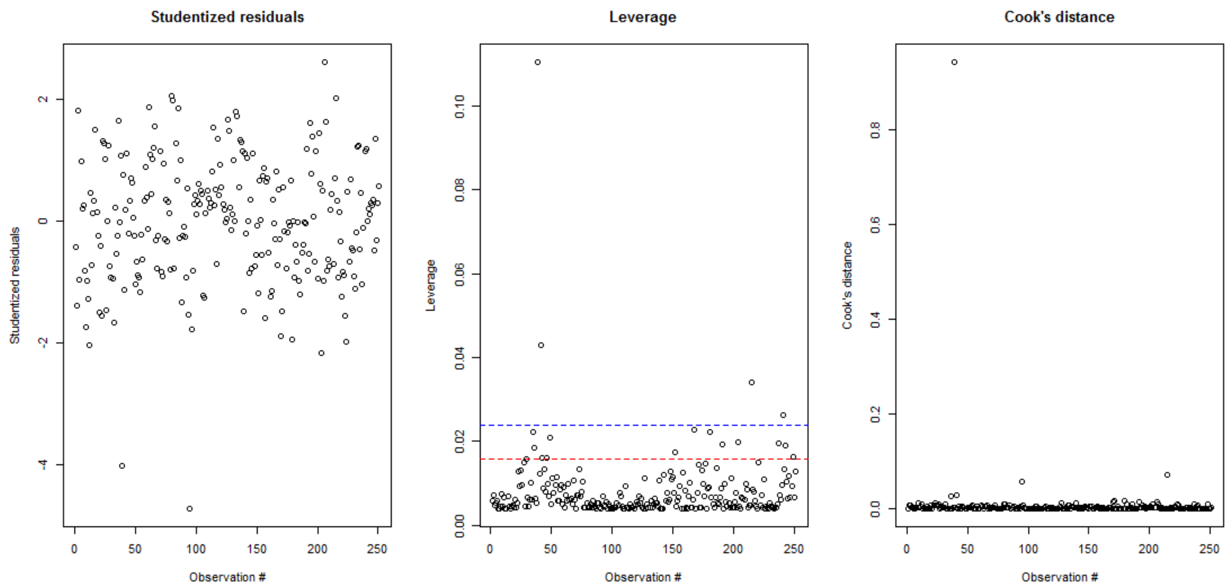
$$\Rightarrow \begin{bmatrix} 4 & 2 & 4 \\ 0 & 6 & 12 \\ 0 & 12 & 24 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} -2 \\ -2 \\ -4 \end{bmatrix} \Rightarrow \begin{cases} 4\beta_0 + 2\beta_1 + 4\beta_2 = -2 \\ 6\beta_1 + 12\beta_2 = -2 \\ 12\beta_1 + 24\beta_2 = -4 \end{cases}$$

$$\Rightarrow \begin{cases} \beta_0 = -\frac{1}{3} \\ \beta_1 = -\frac{1}{3} - 2\beta_2 \\ \beta_2 \text{ unbounded} \end{cases}$$

$\Rightarrow$ Since β_2 is unbounded here, we will have infinitely many solutions.

Application

A.



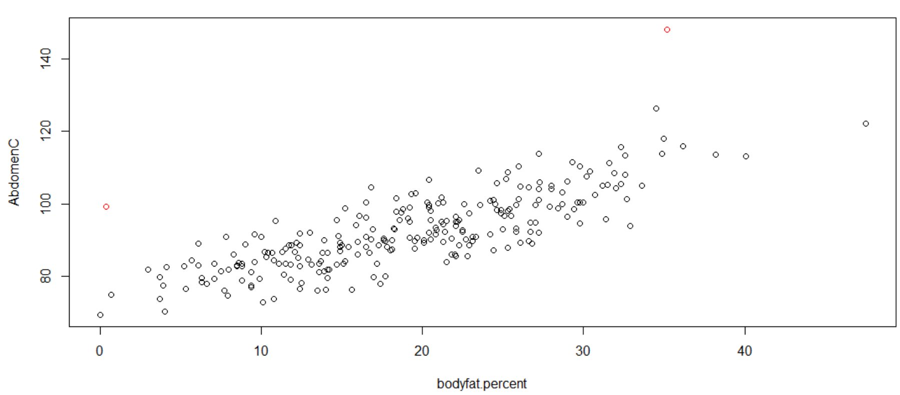

```
> residuals=lm_bodyfat_A$residuals
> sigma_hat=summary(lm_bodyfat_A)$sigma
> X1=model.matrix(bodyfat.percent~AbdomenC)
> H=X1%*%solve(t(X1)%*%X1)%*%t(X1)
> h=diag(H)
```

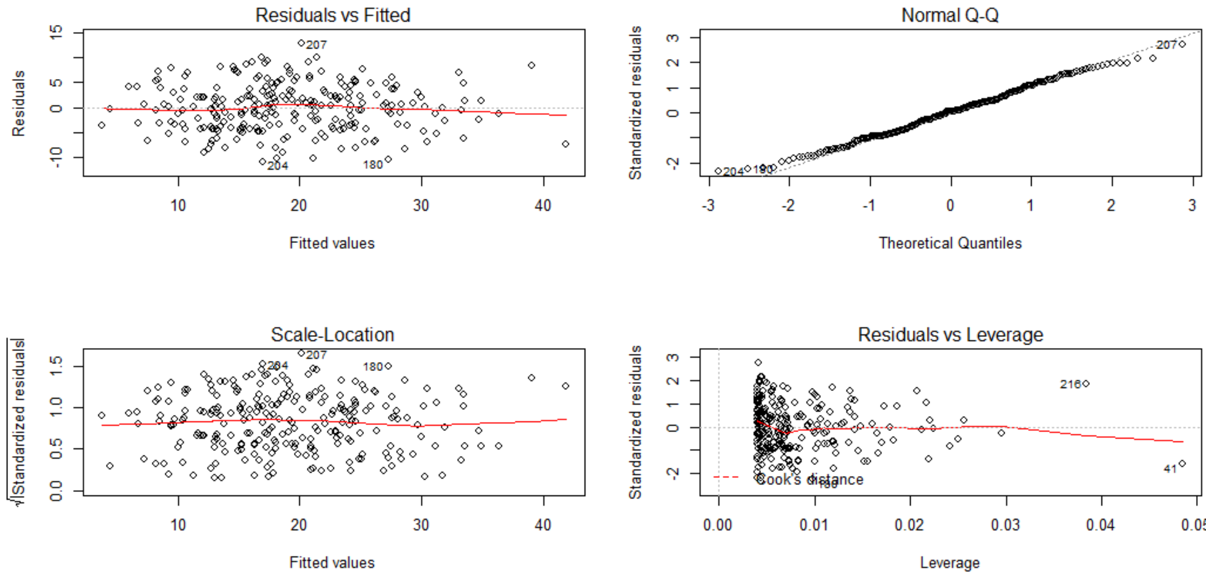
The leverages are computed as above.

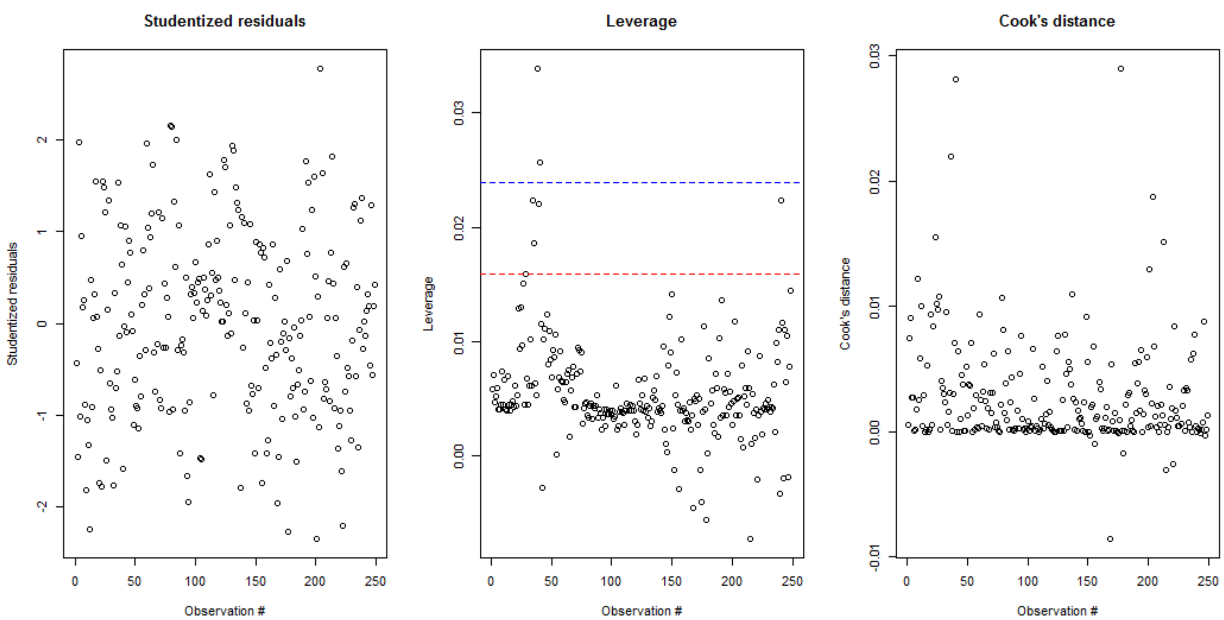
By observing the studentized residual graph, we can see that there are 2 points smaller -3, which may can be the influential points. Also, the Leverage graph shows that some points is above the blue line, which represents the threshold. From the cook's distance graph, we can see that there is an extreme point. So we want to find these extreme point:

> which(D1>0.8) 39 39	> which(h>0.1) 39 39	> which(abs(t)>3) 39 96 39 95

Then, we want to color these two points (in red) and remove them.







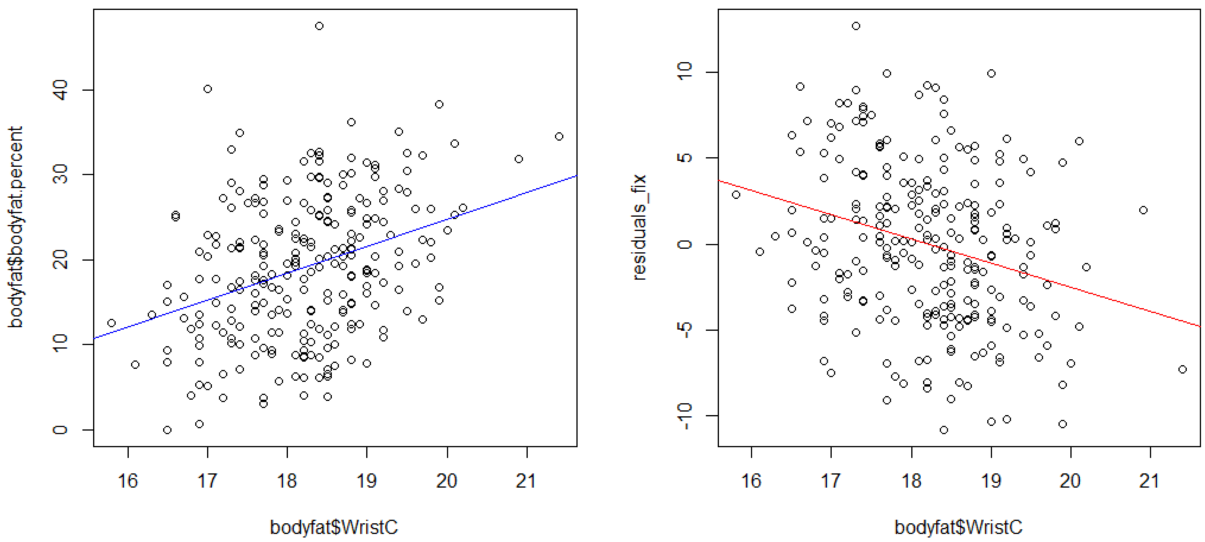

```
        Shapiro-Wilk normality test

data:  lm_bodyfat_A_fix$residuals
W = 0.99329, p-value = 0.3264
```

We can observe that data has constant variance and linearity; the Q-Q plot and the Shapiro-Wilk test (p-value=0.3264>α=0.05, we fail to reject null hypothesis and conclude that residuals are normally distributed) satisfy normality; no obvious outlier are shown by studentized residual graph, leverage graph, and cook's distance graph.

Thus, all assumptions are satisfied and this model is good for further study.

B.

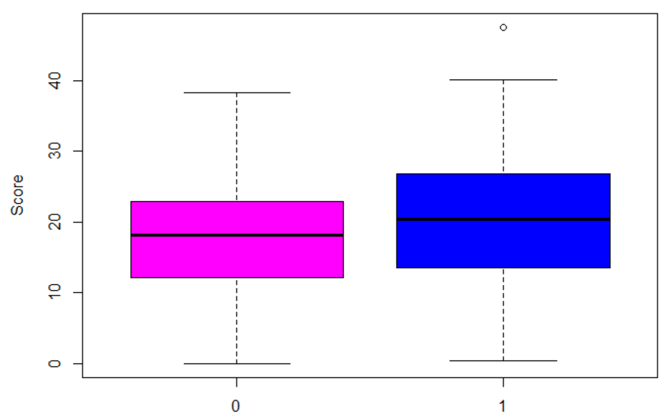


We can observe that the regression lines of the two graphs have opposite signs of slope.

According to the graph, we can say that there may be a strong correlation (collinearity) between Wrist Circumference and the body fat percentage.

That may because people with high body fat percentage have larger Wrist Circumference.

C.




```
> summary(lm_bodyfat_age)

Call:
lm(formula = bodyfat.percent ~ dummy)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-17.9591  -6.3063   0.0409   5.6937  27.0938 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   17.959      0.702  25.582   <2e-16 ***
dummyTRUE      2.447      1.047   2.338   0.0202 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.217 on 247 degrees of freedom
Multiple R-squared:  0.02165,	Adjusted R-squared:  0.01769 
F-statistic: 5.465 on 1 and 247 DF,  p-value: 0.0202
```

Model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $\epsilon \sim N(0,\sigma^2)$ with iid. y_i=bodyfat percentage, x_i=Over45

$$\hat{y}_i = \begin{cases} 17.959 + 2.447\mathrm{x_i}, & d = 1 \\ 17.959, & \mathrm{d} = 0 \end{cases}$$

$R^2 = 0.02165$

H_0: $\mu_{Over45} = \mu_{Under45}$ H_1: $\mu_{Over45} \neq \mu_{Under45}$

p-value = $0.0202 < \alpha = 0.05$

Thus, we reject the null hypothesis and conclude that there is significant difference in the bodyfat percentage between the two age groups.

D.

```
> summary(lm_bodyfat_combine)

Call:
lm(formula = bodyfat$bodyfat.percent ~ AbdomenC + dummy + AbdomenC:dummy)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.6664  -3.5929   0.0152   3.2720  13.1772 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -38.11237    3.54851 -10.740   <2e-16 ***
AbdomenC             0.61592    0.03873  15.901   <2e-16 ***
dummyTRUE          -10.10869    5.48312  -1.844   0.0664 .  
AbdomenC:dummyTRUE   0.11607    0.05887   1.971   0.0498 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.653 on 245 degrees of freedom
Multiple R-squared:  0.6889,	Adjusted R-squared:  0.6851 
F-statistic: 180.8 on 3 and 245 DF,  p-value: < 2.2e-16
```

Model: $y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 x_i d_i + \epsilon_i$, $\epsilon \sim N(0,\sigma^2)$ with iid. y_i=bodyfat percentage, x_i=Over45, d_i=dummy

$$\hat{y}_i = \begin{cases} -38.11237 + 0.61592\mathrm{x_i} - 10.10869 + 0.11607\mathrm{x_i}, & d = 1 \\ -38.11237 + 0.61592\mathrm{x_i}, & \mathrm{d} = 0 \end{cases}$$

```
> anova(lm_bodyfat_A_fix,lm_bodyfat_combine)
Analysis of Variance Table

Model 1: bodyfat$bodyfat.percent ~ bodyfat$AbdomenC
Model 2: bodyfat$bodyfat.percent ~ AbdomenC + dummy + AbdomenC:dummy
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    247 5412.3
2    245 5303.6  2    108.66 2.5098 0.08337 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

H_0: the mean response of the reduced model is the same as that of the model with the dummy variable.

H_1: the mean response of the reduced model is NOT the same as that of the model with the dummy variable.

Test-statistic = $F_{2,245}$ = 2.5098

p-value=0.08337 > α =0.05

Thus, we fail to reject the null hypothesis and conclude that the reduced model is the same as the model with the dummy variable.

Thus, the model with the dummy variable is not better than the reduced model and the predictor Over45 is not significant.