Homework 3

Juwon Lee, Economics and Statistics, UCLA

2023-01-13

tinytex::install_tinytex()

1.

(a).

Based on the output for model (3.7) a business analyst concluded the following:

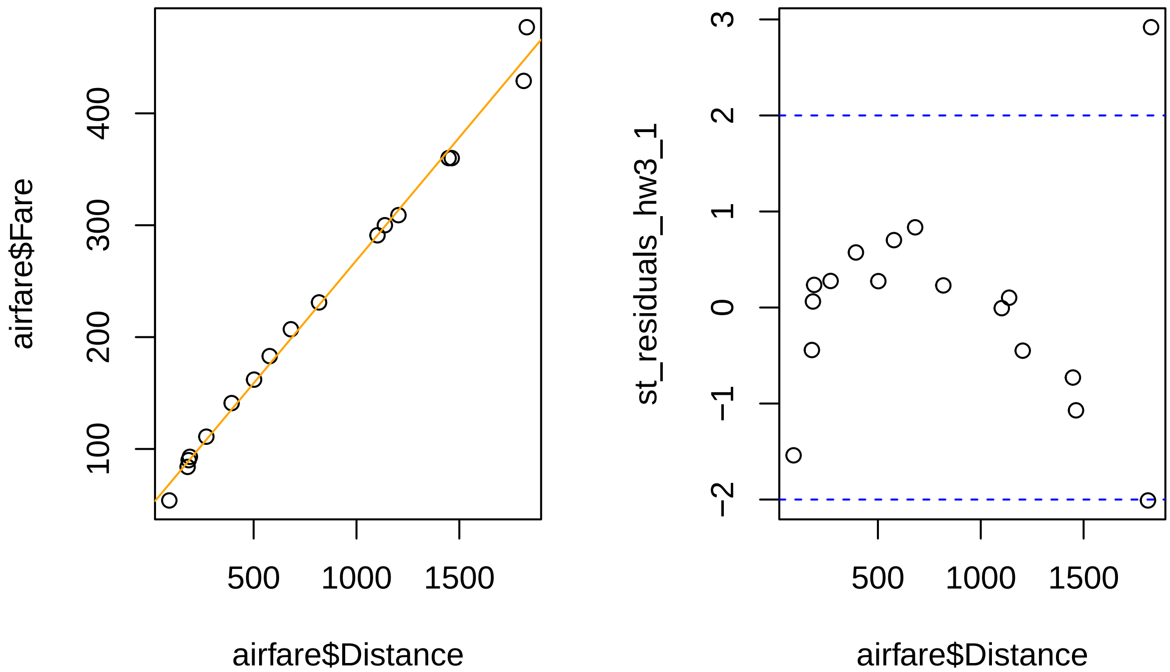
The regression coefficient of the predictor variable, Distance is highly statistically significant and the model explains 99.4% of the variability in the Y-variable, Fare. Thus model (1) is a highly effective model for both understanding the effects of Distance on Fare and for predicting future values of Fare given the value of the predictor variable, Distance.

There are three methods to provide a detailed critique, $\begin{cases} h_{ii} > \frac{4}{n} \to \ \frac{4}{17} \approx 0.235 \\ |\gamma_i| > 2 \\ D_i > \frac{4}{n-2} \to \ \frac{4}{15} \approx 0.267 \end{cases}$,

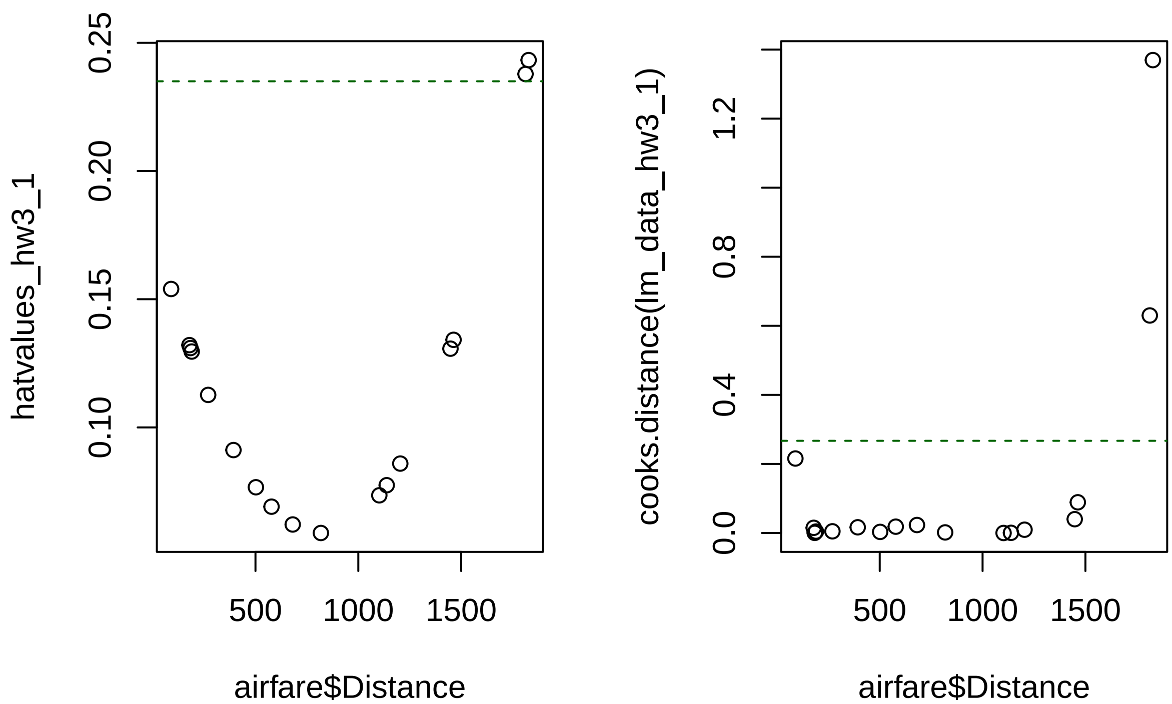
```r
airfare <- read.table("airfares.txt", header=T)

lm_data_hw3_1 <- lm(airfare$Fare~airfare$Distance, data=airfare)

s_hw3_1 <- (sum((lm_data_hw3_1$residuals - mean(lm_data_hw3_1$residuals))^2) / (length(airfare$Fare)-2))

hatvalues_hw3_1 <- hatvalues(lm_data_hw3_1)

st_residuals_hw3_1 <- lm_data_hw3_1$residuals / (s_hw3_1 * (1-hatvalues_hw3_1)^(1/2))

cooks.distance(lm_data_hw3_1)
```

```
##            1            2            3            4            5            6
## 8.883565e-02 3.997799e-02 2.310385e-02 4.856507e-03 4.128465e-03 1.648618e-02
##            7            8            9           10           11           12
## 1.784460e-06 1.826705e-02 9.494655e-03 4.401410e-04 2.934379e-04 3.125113e-03
##           13           14           15           16           17
## 1.369600e+00 1.492552e-02 1.654116e-03 2.156824e-01 6.299398e-01
```

```r
par(mfcol=c(1,2))
plot(airfare$Distance, airfare$Fare)
abline(lm_data_hw3_1$coefficients[1], lm_data_hw3_1$coefficients[2], col='orange')

plot(airfare$Distance, st_residuals_hw3_1)
abline(2, 0, col='blue', lty='dashed')
abline(-2, 0, col='blue', lty='dashed')
```




```r
par(mfcol=c(1,2))
plot(airfare$Distance, hatvalues_hw3_1)
abline(0.235, 0, col='darkgreen', lty='dashed')

plot(airfare$Distance, cooks.distance(lm_data_hw3_1))
abline(0.267, 0, col='darkgreen', lty='dashed')
```



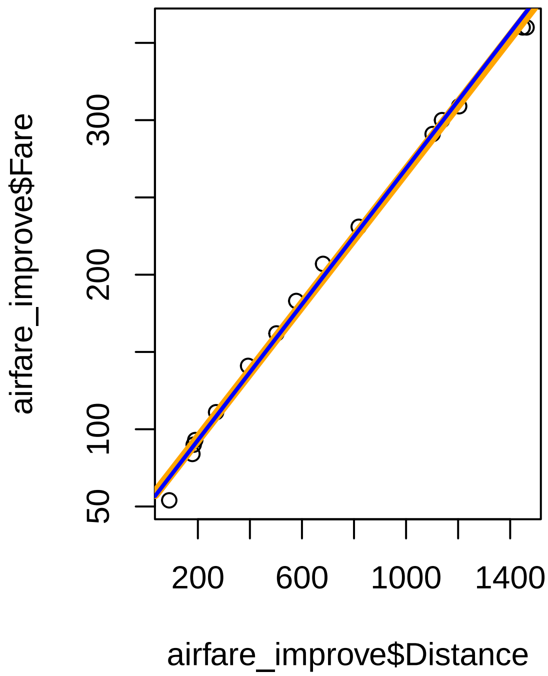
(b)

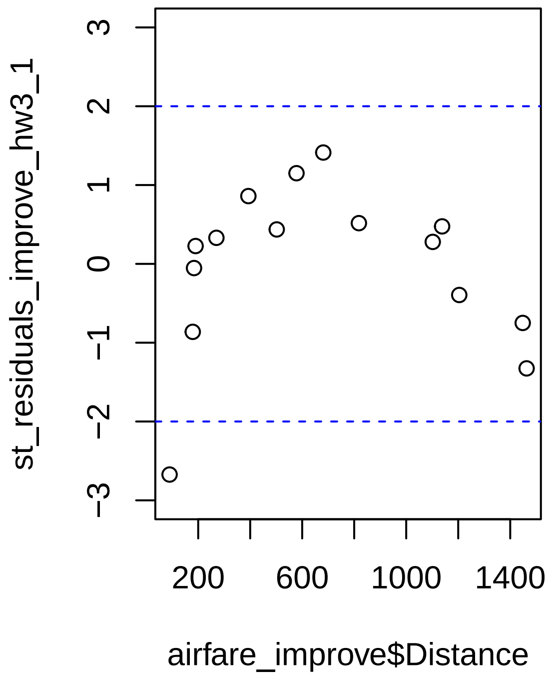
Thus, two values who have more than 1500 distances are bad leverage points.
Also, they have big Cook's distance, too.

```
airfare_improve <- airfare[c(-13,-17),]

lm_data_improve_hw3_1 <- lm(airfare_improve$Fare~airfare_improve$Distance, data=airfare_improve)

s_improve_hw3_1 <- (sum((lm_data_improve_hw3_1$residuals - mean(lm_data_improve_hw3_1$residuals))^2) /

hatvalues_improve_hw3_1 <- hatvalues(lm_data_improve_hw3_1)

st_residuals_improve_hw3_1 <- lm_data_improve_hw3_1$residuals / (s_improve_hw3_1 * (1-hatvalues_improve_

cooks.distance(lm_data_improve_hw3_1)
```

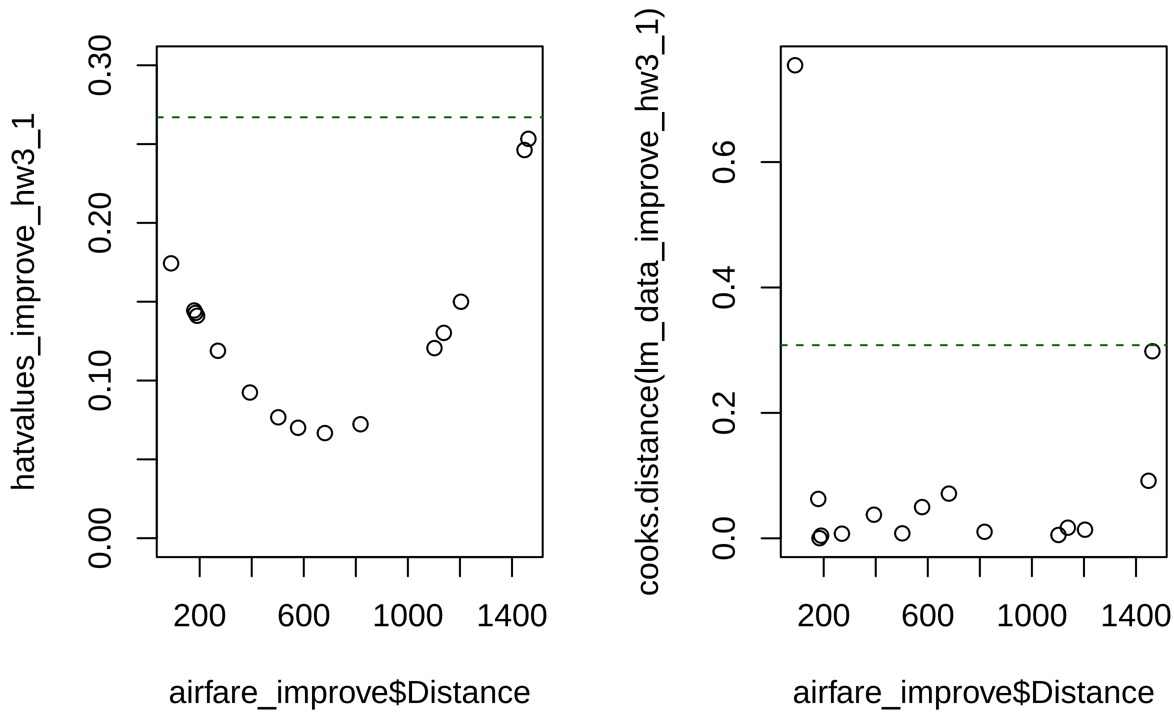
```
##            1            2            3            4            5            6
## 0.2982606823 0.0917448061 0.0712568213 0.0073769587 0.0042049174 0.0376435716
##            7            8            9           10           11           12
## 0.0053263543 0.0498181939 0.0137501029 0.0169499453 0.0002344767 0.0079227965
##           14           15           16
## 0.0627871691 0.0104044128 0.7543280904
```

```
par(mfcol=c(1,2))
plot(airfare_improve$Distance, airfare_improve$Fare)
abline(lm_data_improve_hw3_1$coefficients[1], lm_data_improve_hw3_1$coefficients[2], col='orange', lwd=
abline(lm_data_hw3_1$coefficients[1], lm_data_hw3_1$coefficients[2], col='blue', lwd=2)

plot(airfare_improve$Distance, st_residuals_improve_hw3_1, ylim=c(-3,3))
abline(2, 0, col='blue', lty='dashed')
abline(-2, 0, col='blue', lty='dashed')
```





```
par(mfcol=c(1,2))
plot(airfare_improve$Distance, hatvalues_improve_hw3_1, ylim=c(0,0.3))
abline(0.267, 0, col='darkgreen', lty='dashed')

plot(airfare_improve$Distance, cooks.distance(lm_data_improve_hw3_1))
abline(0.308, 0, col='darkgreen', lty='dashed')
```



They have new ones such that $|\gamma_i| > 2$, but we had better not eliminate it because of the originality.

2.

An analyst for the auto industry has asked for your help in modeling data on the prices of new cars. Interest centers on modeling suggested retail price as a function of the cost to the dealer for 234 new cars. The data set, which is available on the book website in the file cars04.csv, is a subset of the data from http://www.amstat.org/publications/jse/datasets/04cars.txt

The first model to fit to the data was
Suggested Retail Price $= \beta_0 + \beta_1 *$ Dealer Cost $+ e$.

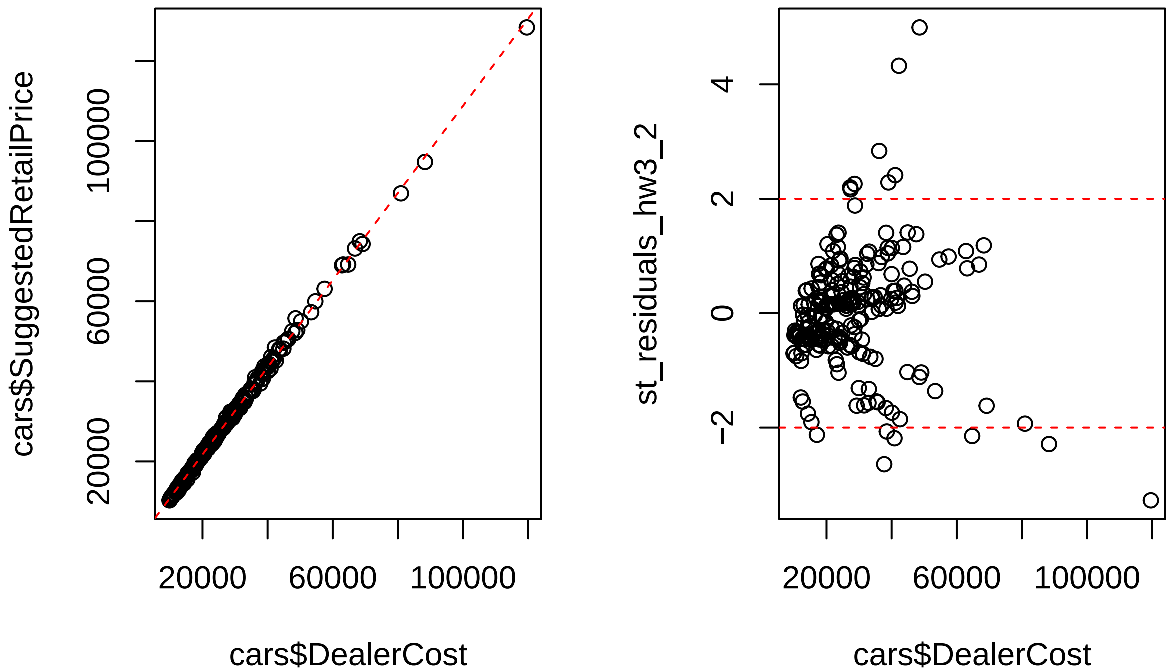
(a)

Based on the output for model, the analyst concluded the following:

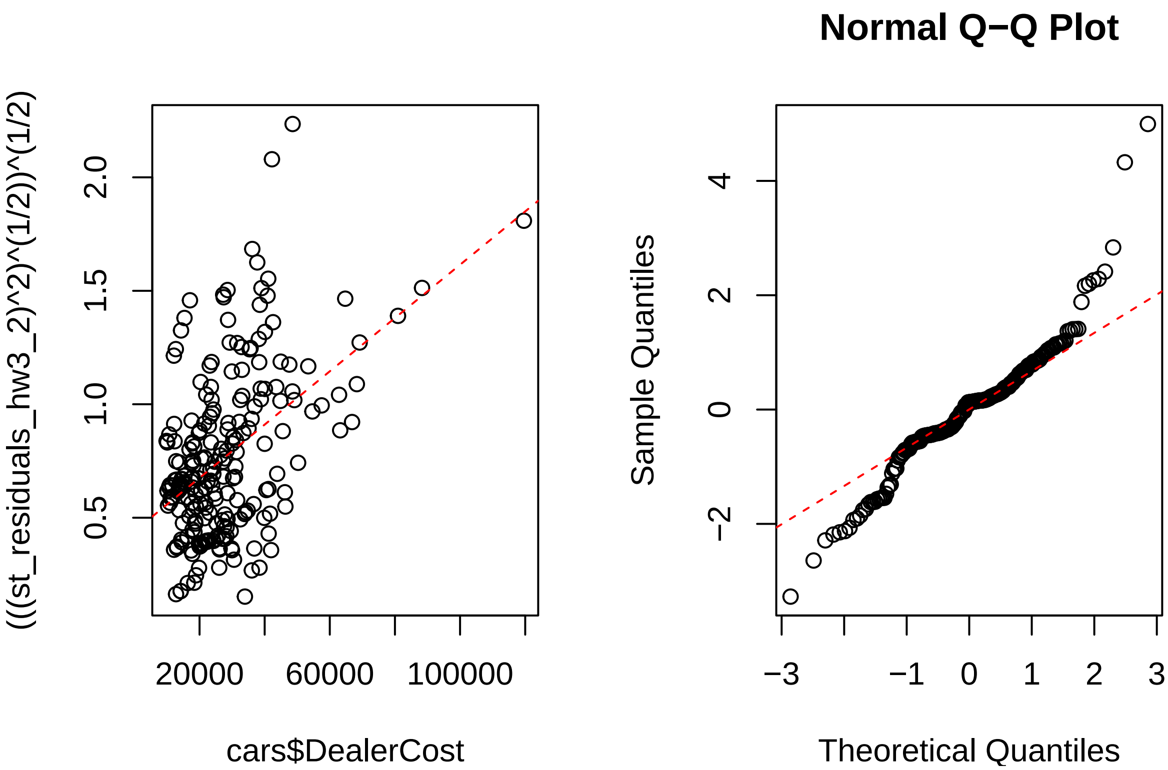
Since the model explains just more than 99.8% of the variabilty in Suggested Retail Price and the coefficient of Dealer Cost has a t-value greater than 412, model (1) is a highly effective model for producting prediction intervals for Suggested Retail Price.

Provide a detailed critique of this conclusion.

```
cars <- read.csv("cars04.csv", header=T)

lm_data_hw3_2 <- lm(cars$SuggestedRetailPrice~cars$DealerCost, data=cars)

s_hw3_2 <- (sum((lm_data_hw3_2$residuals - mean(lm_data_hw3_2$residuals))^2) / (length(cars$DealerCost)

hatvalues_hw3_2 <- hatvalues(lm_data_hw3_2)

st_residuals_hw3_2 <- lm_data_hw3_2$residuals / (s_hw3_2 * (1-hatvalues_hw3_2)^(1/2))

lm_data_residual_hw3_2 <- lm((((st_residuals_hw3_2)^2)^(1/2))^(1/2)~cars$DealerCost, data=cars)

par(mfrow=c(1,2))
plot(cars$DealerCost, cars$SuggestedRetailPrice)
abline(lm_data_hw3_2$coefficients[1], lm_data_hw3_2$coefficients[2], col='red', lty='dashed')

plot(cars$DealerCost, st_residuals_hw3_2)
abline(2,0, col='red', lty='dashed')
abline(-2,0,col='red', lty='dashed')
```

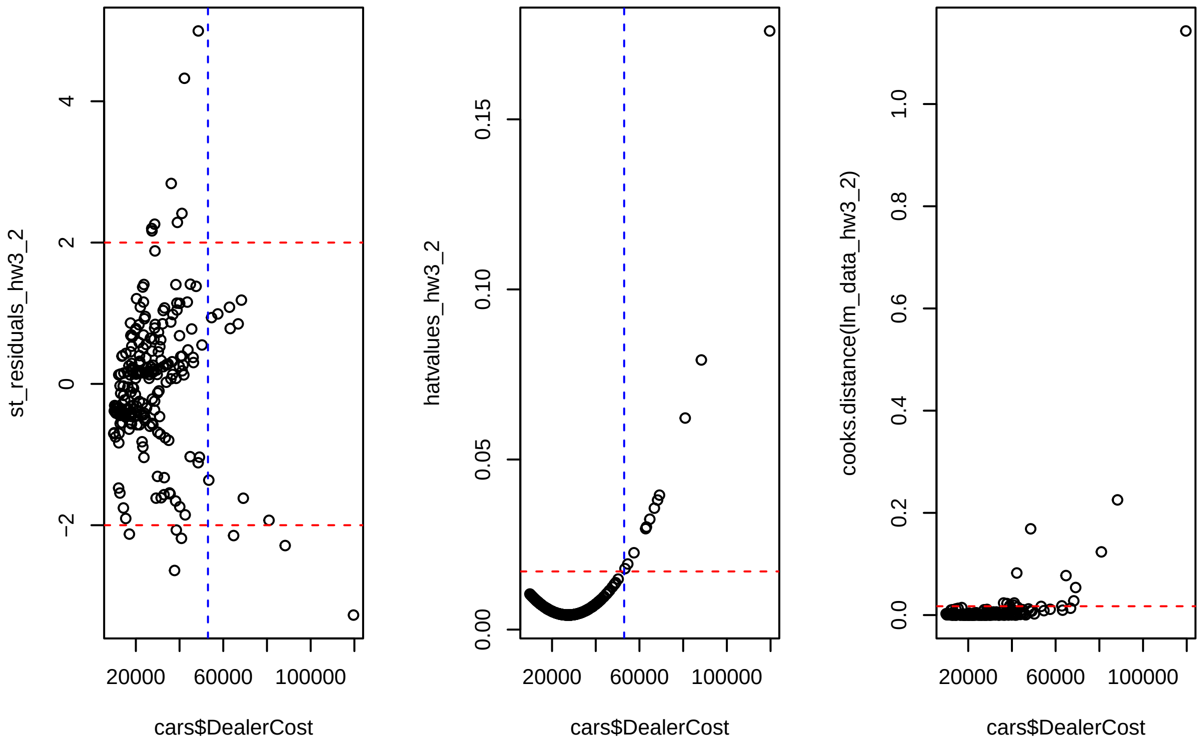



```
par(mfrow=c(1,2))
plot(cars$DealerCost, (((st_residuals_hw3_2)^2)^(1/2))^(1/2))
abline(lm_data_residual_hw3_2$coefficients[1], lm_data_residual_hw3_2$coefficients[2], col='red', lty='d

qqnorm(st_residuals_hw3_2)
qqline(st_residuals_hw3_2, col='red', lty='dashed')
```



There are three methods to provide a detailed critique, $\begin{cases} h_{ii} > \frac{4}{n} \to \ \frac{4}{234} \approx 0.017 \\ |\gamma_i| > 2 \\ D_i > \frac{4}{n-2} \to \ \frac{4}{232} \approx 0.0172 \end{cases}$,

```
par(mfrow=c(1,3))
plot(cars$DealerCost, st_residuals_hw3_2)
abline(2, 0, col='red', lty='dashed')
abline(-2, 0, col='red', lty='dashed')
abline(v=53000, col='blue', lty='dashed')

plot(cars$DealerCost, hatvalues_hw3_2)
abline(4/234,0, col='red', lty='dashed')
abline(v=53000, col='blue', lty='dashed')

plot(cars$DealerCost, cooks.distance(lm_data_hw3_2))
abline(4/232,0,col='red', lty='dashed')
```




```
badleverage <- ((st_residuals_hw3_2)^2)^(1/2) > 2 & hatvalues_hw3_2 > 4/234
badleverage[badleverage==TRUE]
```

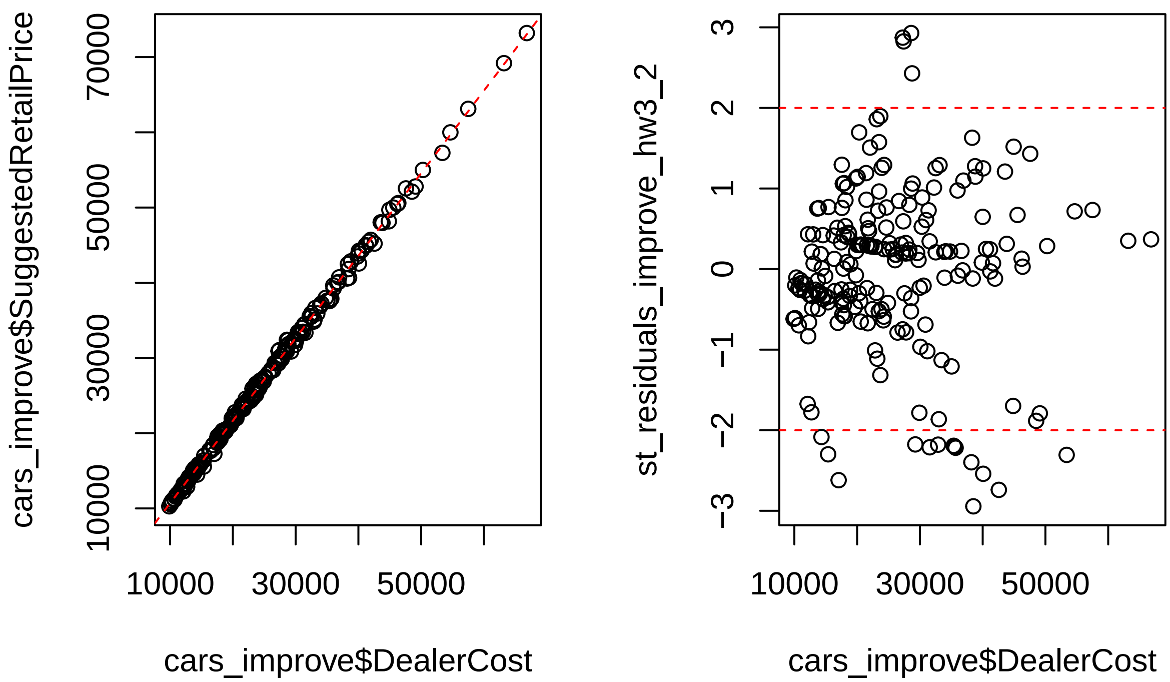
```
##  194  222  223
## TRUE TRUE TRUE
```

```
cooks.distance(lm_data_hw3_2)[cooks.distance(lm_data_hw3_2) > 4/232]
```

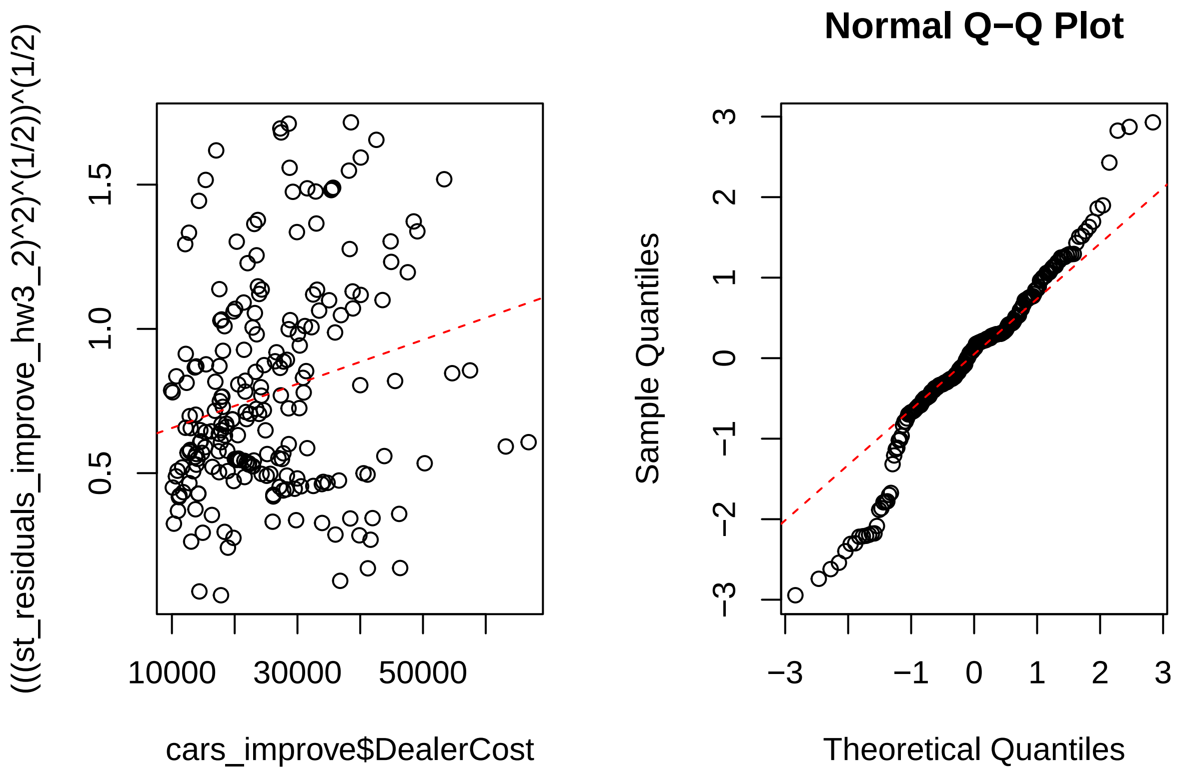
```
##        178        188        189        194        210        212        213
## 0.02256804 0.01841797 0.02367993 0.07728761 0.01800359 0.02781819 0.02363324
##        214        215        222        223        228        229        231
## 0.08232065 0.16876037 0.22534700 1.14307623 0.05388900 0.12363746 0.01915348
```

Thus, three components are bad leverage points.
And we can detect big Cook's distance, too.

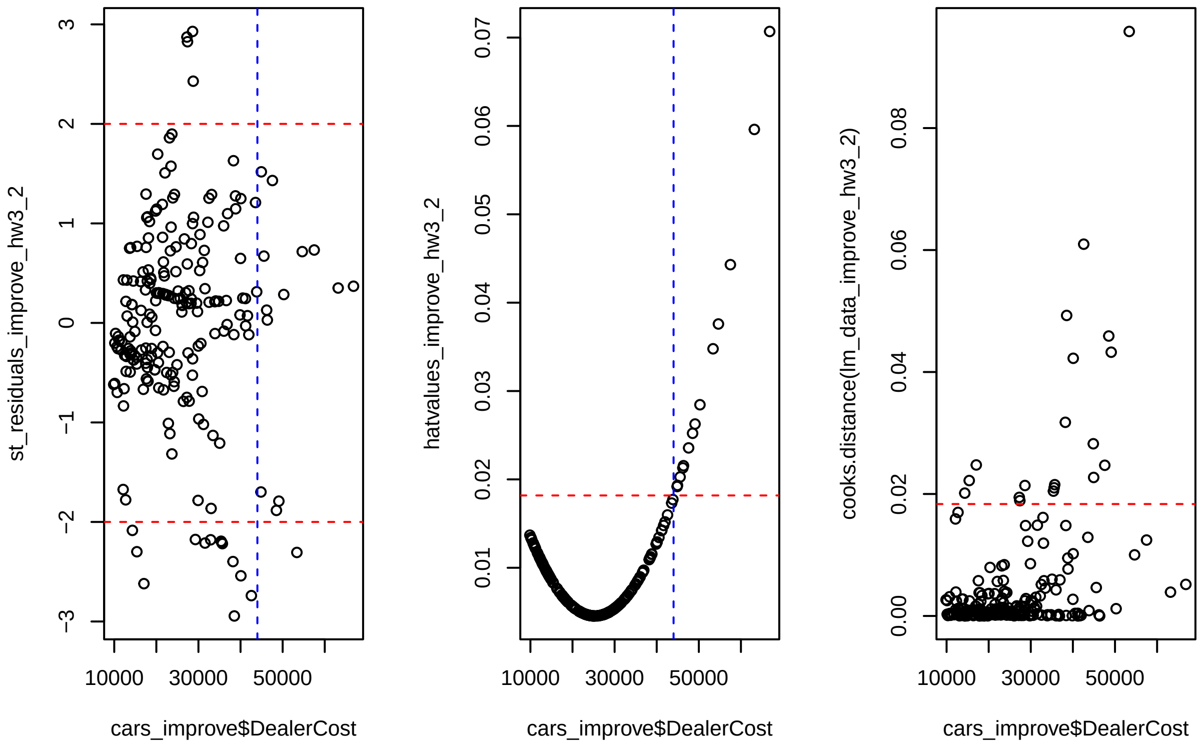
```
cars_improve <- cars[c(-178, -188, -189, -194, -210, -212, -213, -214, -215, -222, -223, -228, -229, -2

lm_data_improve_hw3_2 <- lm(cars_improve$SuggestedRetailPrice~cars_improve$DealerCost, data=cars)

s_improve_hw3_2 <- (sum((lm_data_improve_hw3_2$residuals - mean(lm_data_improve_hw3_2$residuals))^2) /

hatvalues_improve_hw3_2 <- hatvalues(lm_data_improve_hw3_2)

st_residuals_improve_hw3_2 <- lm_data_improve_hw3_2$residuals / (s_improve_hw3_2 * (1-hatvalues_improve

lm_data_residual_improve_hw3_2<-lm((((st_residuals_improve_hw3_2)^2)^(1/2))^(1/2)~cars_improve$DealerCo

par(mfrow=c(1,2))
plot(cars_improve$DealerCost, cars_improve$SuggestedRetailPrice)
abline(lm_data_improve_hw3_2$coefficients[1], lm_data_improve_hw3_2$coefficients[2], col='red', lty='da

plot(cars_improve$DealerCost, st_residuals_improve_hw3_2)
abline(2,0, col='red', lty='dashed')
abline(-2,0,col='red', lty='dashed')
```




```
par(mfrow=c(1,2))
plot(cars_improve$DealerCost, (((st_residuals_improve_hw3_2)^2)^(1/2))^(1/2))
abline(lm_data_residual_improve_hw3_2$coefficients[1], lm_data_residual_improve_hw3_2$coefficients[2],

qqnorm(st_residuals_improve_hw3_2)
qqline(st_residuals_improve_hw3_2, col='red', lty='dashed')
```




```r
par(mfrow=c(1,3))
plot(cars_improve$DealerCost, st_residuals_improve_hw3_2)
abline(2, 0, col='red', lty='dashed')
abline(-2, 0, col='red', lty='dashed')
abline(v=44000, col='blue', lty='dashed')

plot(cars_improve$DealerCost, hatvalues_improve_hw3_2)
abline(4/220,0, col='red', lty='dashed')
abline(v=44000, col='blue', lty='dashed')

plot(cars_improve$DealerCost, cooks.distance(lm_data_improve_hw3_2))
abline(4/218,0,col='red', lty='dashed')
```



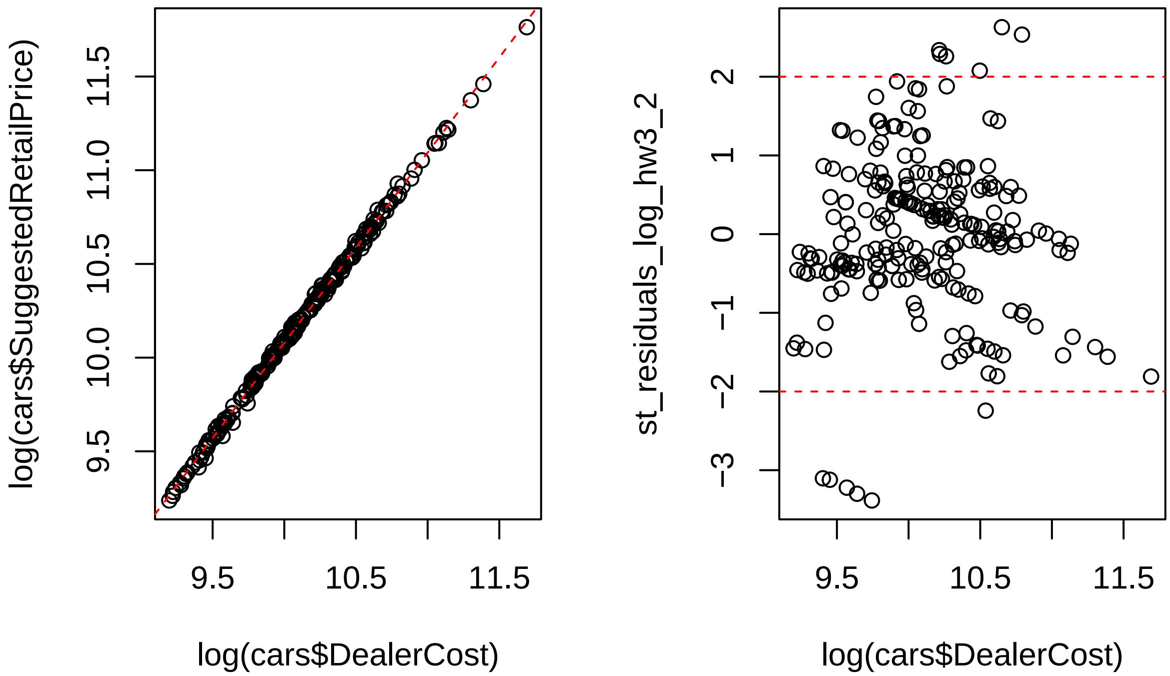
(b)

Carefully describe all the shortcomings evident in model (3.10). For each shortcoming, describe the steps needed to overcome the shortcoming.
(1) The sqaure root of standardized residual has steep slope. $\rightarrow$ we can use log-scale.
(2) It has a heavy-tail in QQ-plot.

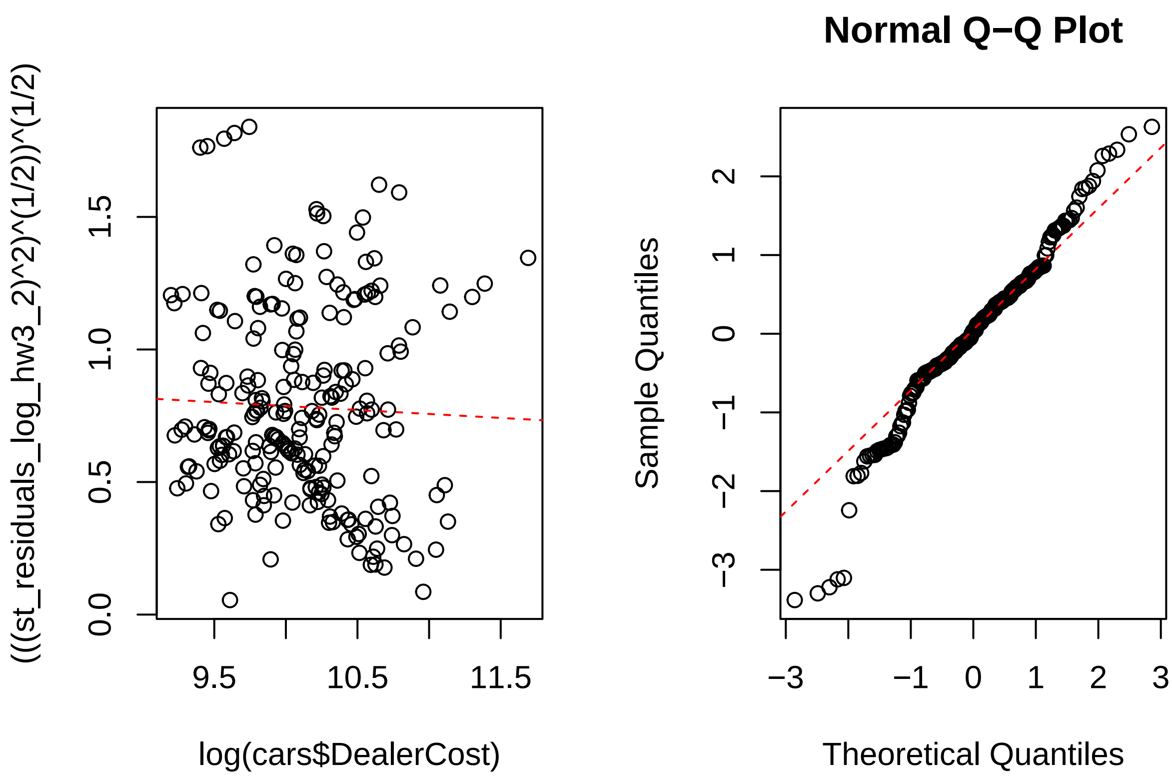
(c)

The second model fitted to the data was
$\log(\text{Suggested Retail Price}) = \beta_0 + \beta_1 \log(\text{Dealer Cost}) + e$.

```
lm_data_log_hw3_2 <- lm(log(cars$SuggestedRetailPrice)~log(cars$DealerCost), data=cars)

s_log_hw3_2 <- (sum((lm_data_log_hw3_2$residuals - mean(lm_data_log_hw3_2$residuals))^2) / (length(cars$

hatvalues_log_hw3_2 <- hatvalues(lm_data_log_hw3_2)

st_residuals_log_hw3_2 <- lm_data_log_hw3_2$residuals / (s_log_hw3_2 * (1-hatvalues_log_hw3_2)^(1/2))

lm_data_residual_log_hw3_2 <- lm((((st_residuals_log_hw3_2)^2)^(1/2))^(1/2)~log(cars$DealerCost), data=c

par(mfrow=c(1,2))
plot(log(cars$DealerCost), log(cars$SuggestedRetailPrice))
abline(lm_data_log_hw3_2$coefficients[1], lm_data_log_hw3_2$coefficients[2], col='red', lty='dashed')

plot(log(cars$DealerCost), st_residuals_log_hw3_2)
abline(2,0, col='red', lty='dashed')
abline(-2,0,col='red', lty='dashed')
```



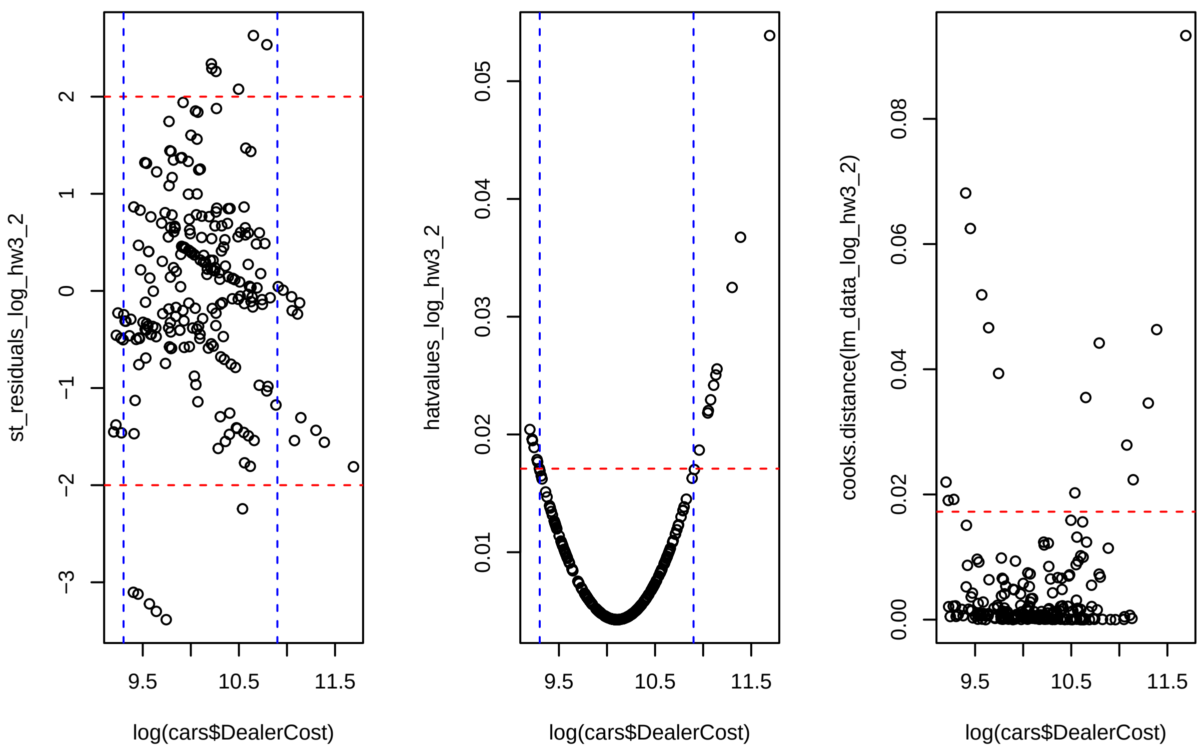

```
par(mfrow=c(1,2))
plot(log(cars$DealerCost), (((st_residuals_log_hw3_2)^2)^(1/2))^(1/2))
abline(lm_data_residual_log_hw3_2$coefficients[1], lm_data_residual_log_hw3_2$coefficients[2], col='red

qqnorm(st_residuals_log_hw3_2)
qqline(st_residuals_log_hw3_2, col='red', lty='dashed')
```



Thus, the log-scale model is more fitted than above one. This is because

(1) More γ_i are in (-2,2).
(2) Square root of standardized residual has flatter regression.
(3) Normality is better.

```
par(mfrow=c(1,3))
plot(log(cars$DealerCost), st_residuals_log_hw3_2)
abline(2, 0, col='red', lty='dashed')
abline(-2, 0, col='red', lty='dashed')
abline(v=9.3, col='blue', lty='dashed')
abline(v=10.9, col='blue', lty='dashed')

plot(log(cars$DealerCost), hatvalues_log_hw3_2)
abline(4/234,0, col='red', lty='dashed')
abline(v=9.3, col='blue', lty='dashed')
abline(v=10.9, col='blue', lty='dashed')

plot(log(cars$DealerCost), cooks.distance(lm_data_log_hw3_2))
abline(4/232,0,col='red', lty='dashed')
```

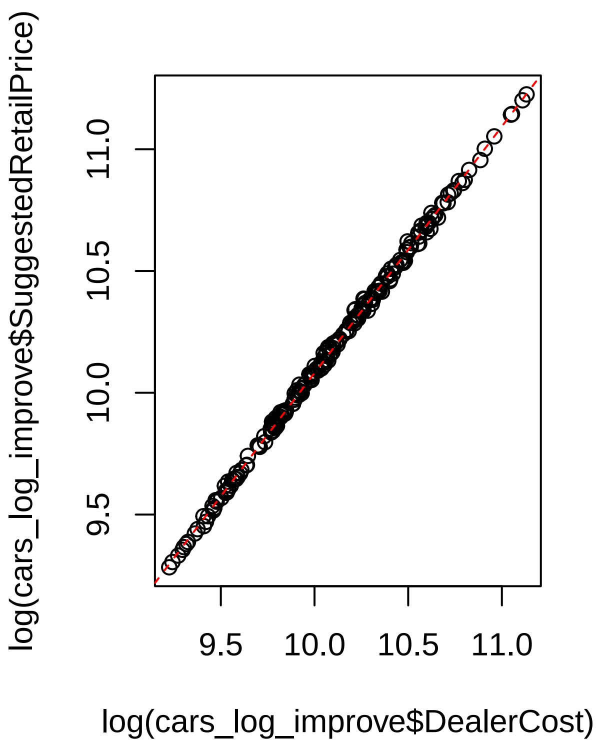


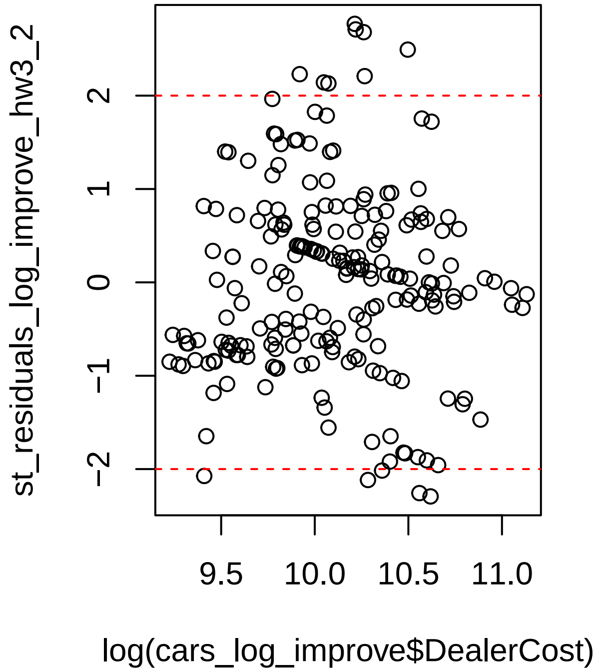
Thus, there are no bad leverage points,
and if we eliminate the values having big Cook's distances,

```
cooks.distance(lm_data_log_hw3_2)[cooks.distance(lm_data_log_hw3_2) > 4/232]
```

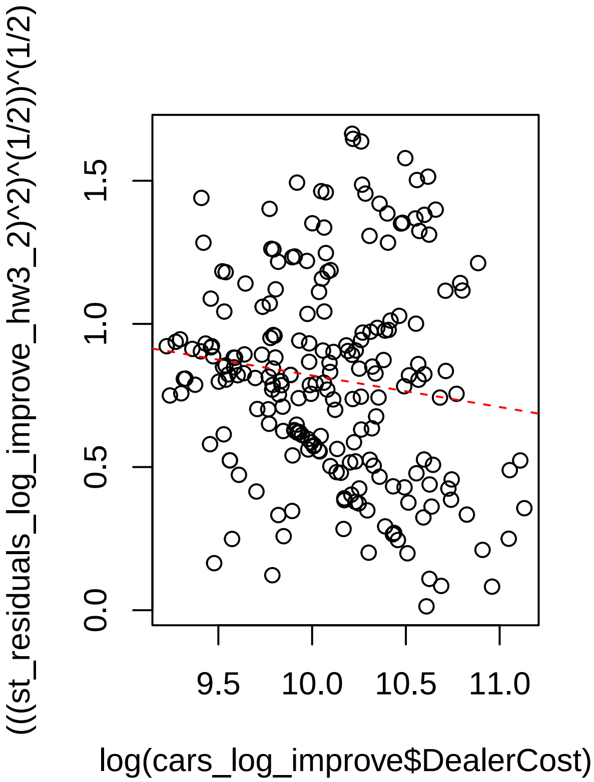
```
##         15         22         23         37         38         39         40
## 0.01903889 0.02196987 0.01921043 0.06248367 0.05188559 0.06814664 0.04663131
##         83        178        194        214        215        222        223
## 0.03933094 0.02024618 0.02788756 0.03548507 0.04418252 0.04633358 0.09330748
##        228        229
## 0.02234703 0.03459348
```

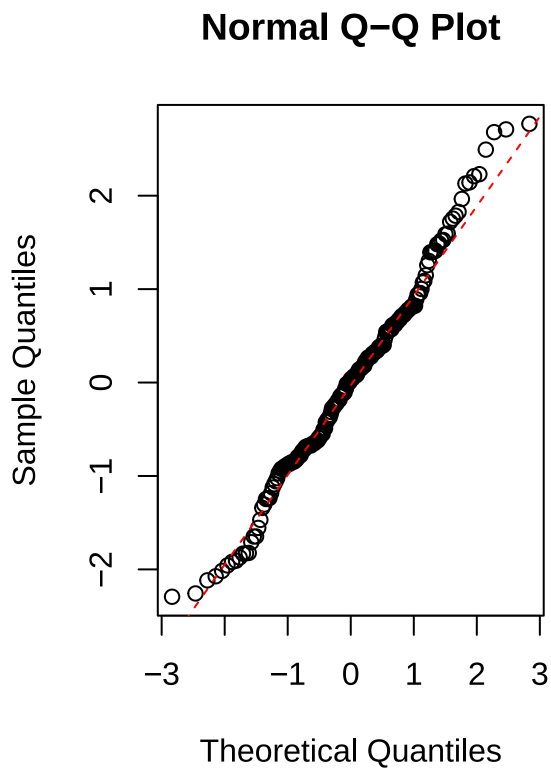
```
cars_log_improve <- cars[c(-15,-22,-23,-37,-38,-39,-40,-83,-178,-194,-214,-215,-222,-223,-228,-229),]

lm_data_log_improve_hw3_2 <- lm(log(cars_log_improve$SuggestedRetailPrice)~log(cars_log_improve$DealerC

s_log_improve_hw3_2 <- (sum((lm_data_log_improve_hw3_2$residuals - mean(lm_data_log_improve_hw3_2$resid

hatvalues_log_improve_hw3_2 <- hatvalues(lm_data_log_improve_hw3_2)

st_residuals_log_improve_hw3_2 <- lm_data_log_improve_hw3_2$residuals / (s_log_improve_hw3_2 * (1-hatva

lm_data_residual_log_improve_hw3_2 <- lm((((st_residuals_log_improve_hw3_2)^2)^(1/2))^(1/2)~log(cars_lo

par(mfrow=c(1,2))
plot(log(cars_log_improve$DealerCost), log(cars_log_improve$SuggestedRetailPrice))
abline(lm_data_log_improve_hw3_2$coefficients[1], lm_data_log_improve_hw3_2$coefficients[2], col='red',

plot(log(cars_log_improve$DealerCost), st_residuals_log_improve_hw3_2)
abline(2,0, col='red', lty='dashed')
abline(-2,0,col='red', lty='dashed')
```



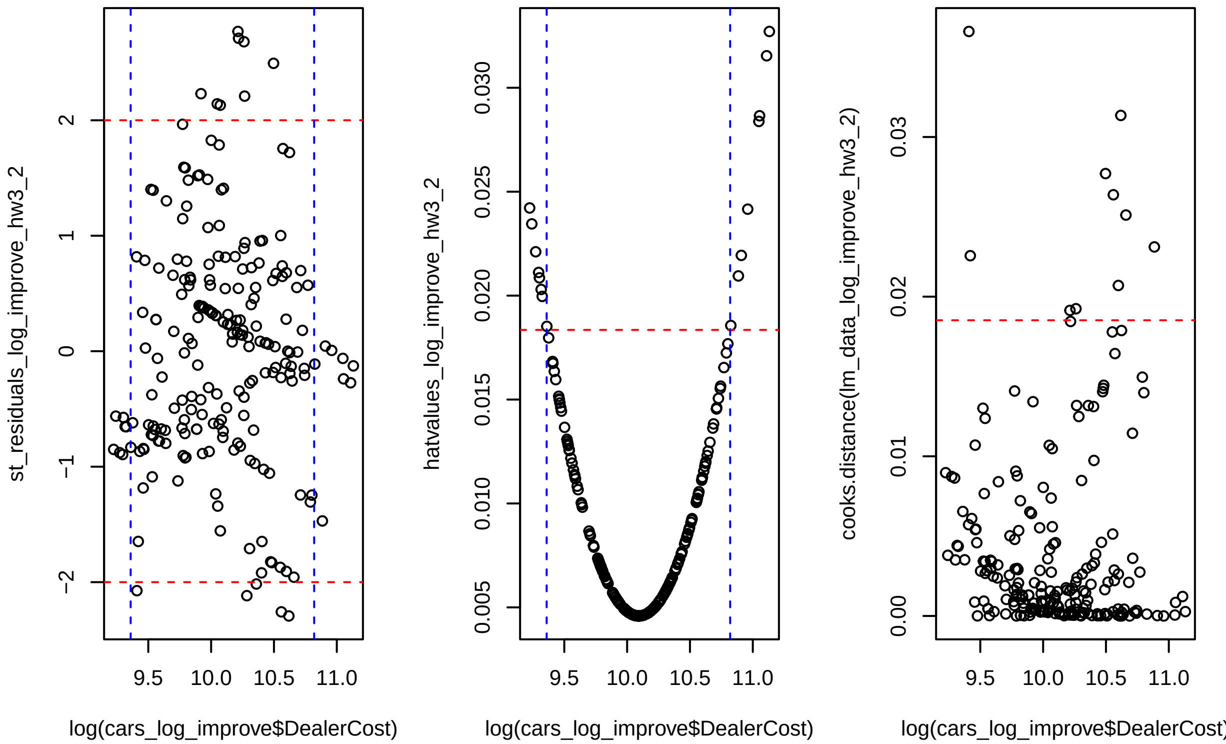



```
par(mfrow=c(1,2))
plot(log(cars_log_improve$DealerCost), (((st_residuals_log_improve_hw3_2)^2)^(1/2))^(1/2))
abline(lm_data_residual_log_improve_hw3_2$coefficients[1], lm_data_residual_log_improve_hw3_2$coefficie

qqnorm(st_residuals_log_improve_hw3_2)
qqline(st_residuals_log_improve_hw3_2, col='red', lty='dashed')
```






```
par(mfrow=c(1,3))
plot(log(cars_log_improve$DealerCost), st_residuals_log_improve_hw3_2)
abline(2, 0, col='red', lty='dashed')
abline(-2, 0, col='red', lty='dashed')
abline(v=9.36, col='blue', lty='dashed')
abline(v=10.82, col='blue', lty='dashed')

plot(log(cars_log_improve$DealerCost), hatvalues_log_improve_hw3_2)
abline(4/218,0, col='red', lty='dashed')
abline(v=9.36, col='blue', lty='dashed')
abline(v=10.82, col='blue', lty='dashed')

plot(log(cars_log_improve$DealerCost), cooks.distance(lm_data_log_improve_hw3_2))
abline(4/216,0,col='red', lty='dashed')
```



(d)

log(Dealer Cost) = 1.01484, which is the amount of change of Suggested Retail Price when Dealer Cost fluctuates.

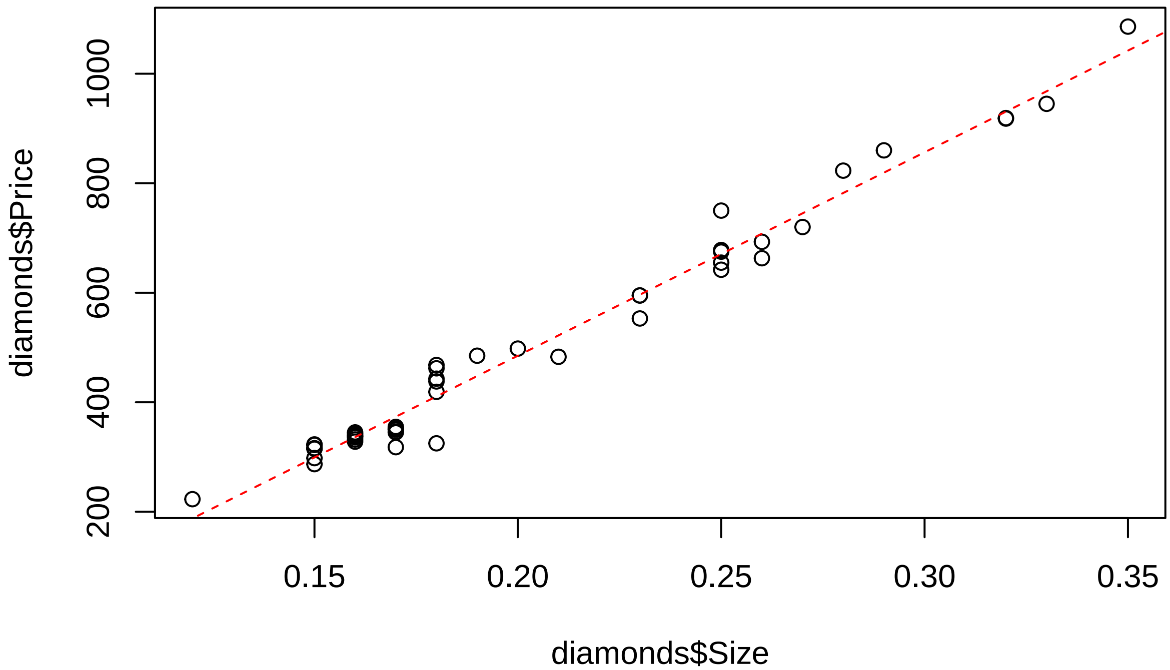
(e)

3.

Chu (1996) discusses the development of a regression model to predict the price of diamond rings from the size of their diamond stones (in terms of their weight in carats). Data on both variables were obtained from a full page advertisement placed in the *Straits Times* newspaper by a Singapore-based retailer of diamond jewelry. Only rings made with 20 carat gold and mounted with a single diamond stone were included in the data set. There were 48 such rings of varying designs. (Information on the designs was available but not used in the modeling.)

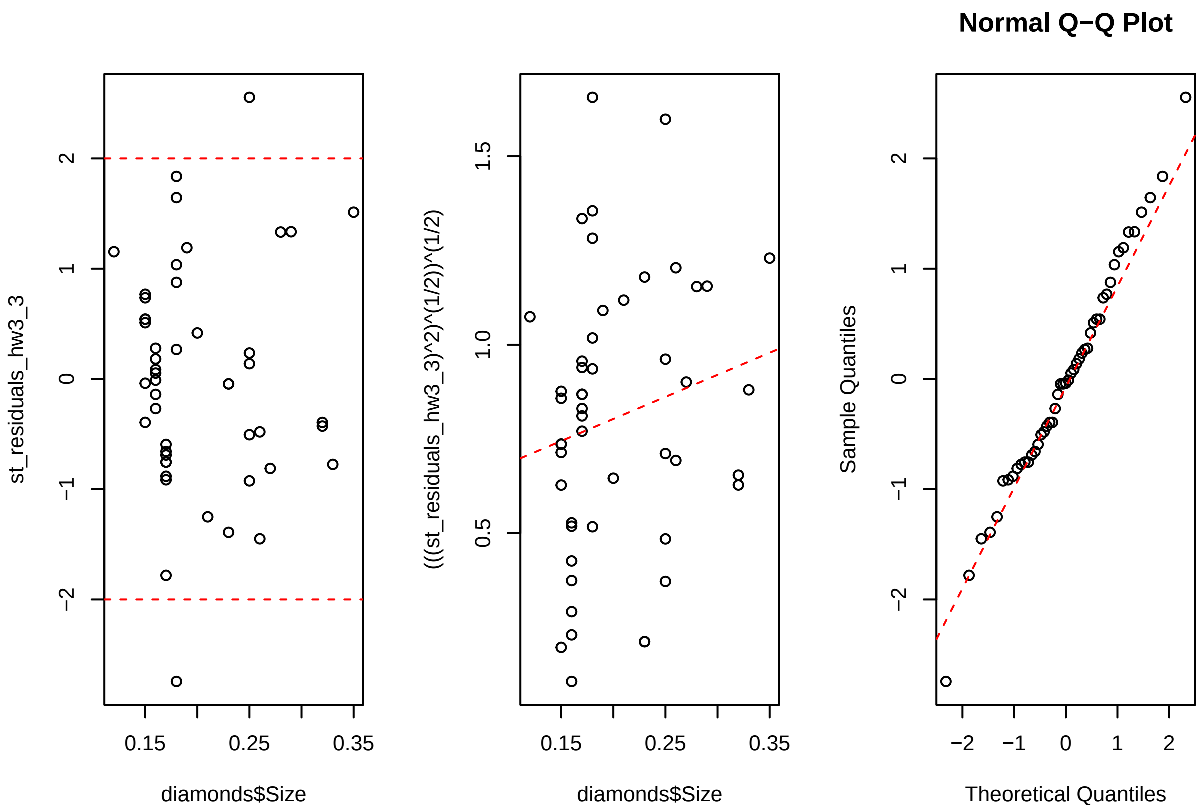
Part 1 - (a)

```
diamonds <- read.table("/Users/user/Desktop/Yonsei/Junior/3-2/Introduction to Data Analysis and Regressi

lm_data_hw3_3 <- lm(diamonds$Price~diamonds$Size, data=diamonds)

plot(diamonds$Size, diamonds$Price)
abline(lm_data_hw3_3$coefficients[1], lm_data_hw3_3$coefficients[2], col='red', lty='dashed')
```



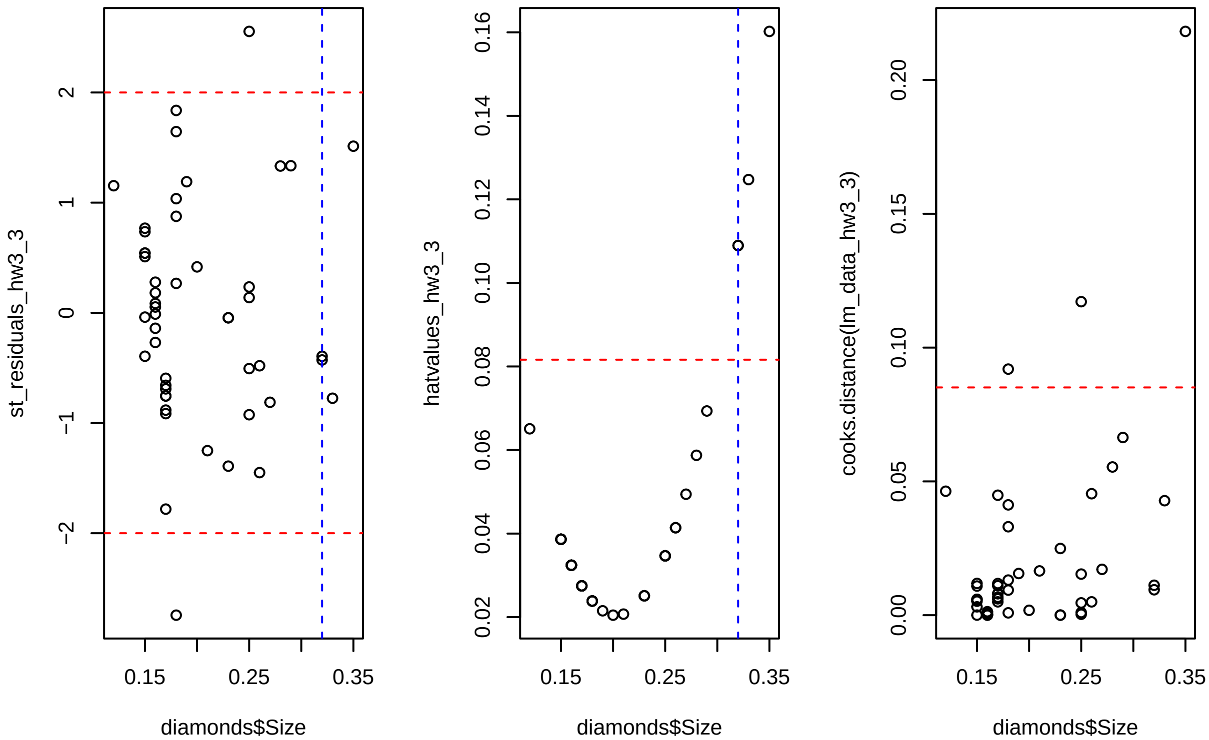

```
###

s_hw3_3 <- (sum((lm_data_hw3_3$residuals - mean(lm_data_hw3_3$residuals))^2) / (length(diamonds$Price)-2

hatvalues_hw3_3 <- hatvalues(lm_data_hw3_3)

st_residuals_hw3_3 <- lm_data_hw3_3$residuals / (s_hw3_3 * (1-hatvalues_hw3_3)^(1/2))

lm_data_residual_hw3_3 <- lm((((st_residuals_hw3_3)^2)^(1/2))^(1/2)~diamonds$Size, data=diamonds)

###

par(mfrow=c(1,3))
plot(diamonds$Size, st_residuals_hw3_3)
abline(2,0,col='red', lty='dashed')
```

```
abline(-2,0,col='red', lty='dashed')

plot(diamonds$Size, (((st_residuals_hw3_3)^2)^(1/2))^(1/2))
abline(lm_data_residual_hw3_3$coefficients[1], lm_data_residual_hw3_3$coefficients[2], col='red', lty='

qqnorm(st_residuals_hw3_3)
qqline(st_residuals_hw3_3, col='red', lty='dashed')
```



And when we check our power of justification,

```
par(mfrow=c(1,3))
plot(diamonds$Size, st_residuals_hw3_3)
abline(2, 0, col='red', lty='dashed')
abline(-2, 0, col='red', lty='dashed')
abline(v=0.32, col='blue', lty='dashed')

plot(diamonds$Size, hatvalues_hw3_3)
abline(4/length(diamonds$Size),0, col='red', lty='dashed')
abline(v=0.32, col='blue', lty='dashed')

plot(diamonds$Size, cooks.distance(lm_data_hw3_3))
abline(4/(length(diamonds$Size)-2),0,col='red', lty='dashed')
```

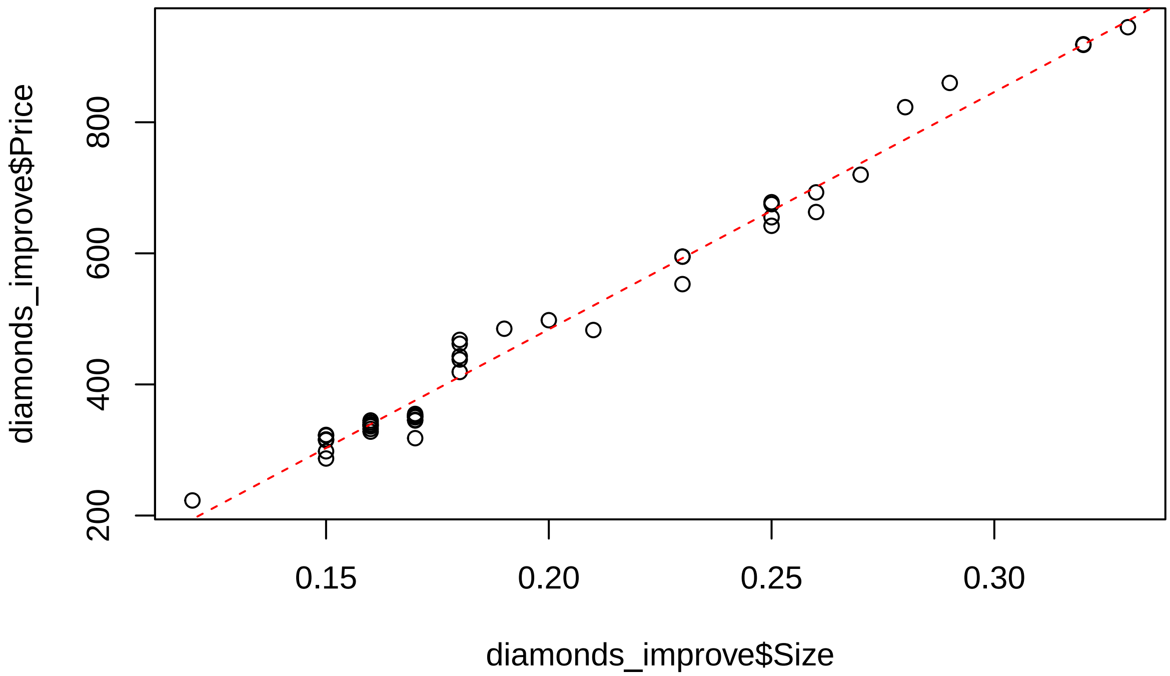


Thus, they don't have any bad leverage points.
If we eliminate values having 'big' cook's distance,

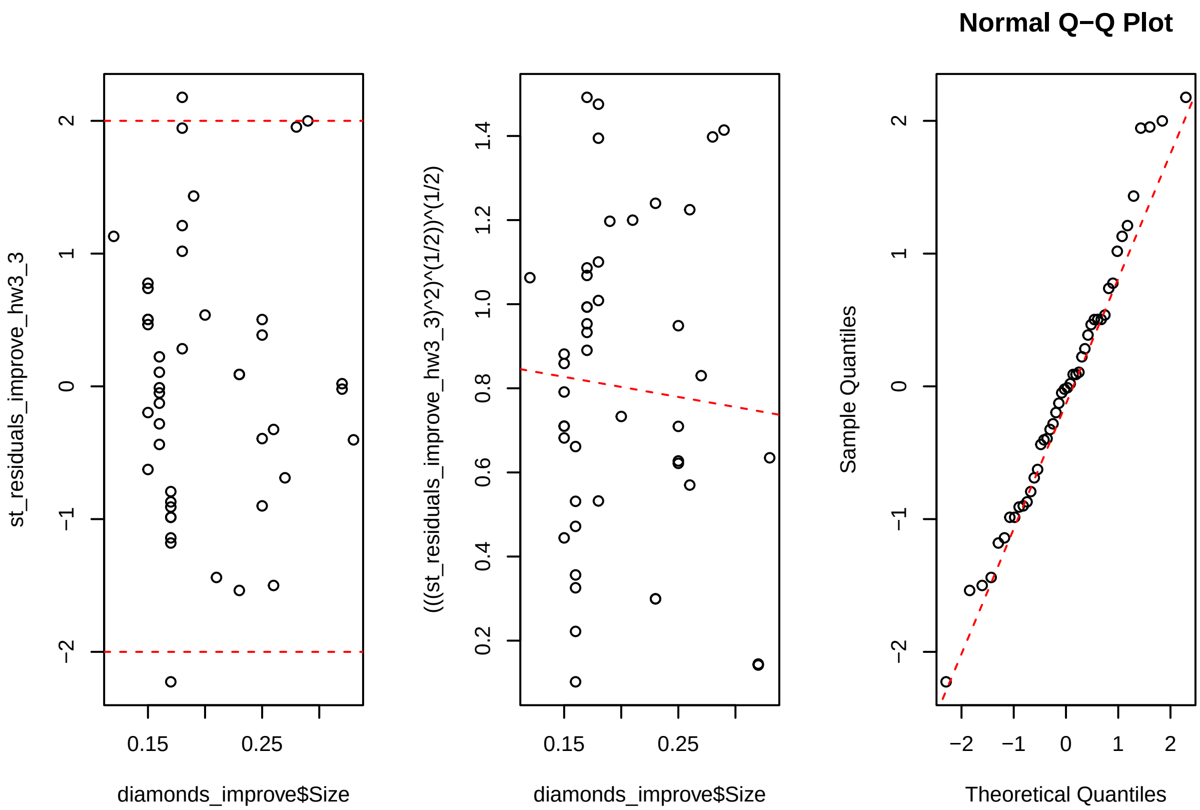
```
cooks.distance(lm_data_hw3_3)[cooks.distance(lm_data_hw3_3) > 4/(length(diamonds$Price)-2)]
```

```
##          4          19          42
## 0.09196098 0.11715838 0.21815953
```

```
diamonds_improve <- diamonds[c(-4,-19,-42),]

lm_data_improve_hw3_3 <- lm(diamonds_improve$Price~diamonds_improve$Size, data=diamonds_improve)

plot(diamonds_improve$Size, diamonds_improve$Price)
abline(lm_data_improve_hw3_3$coefficients[1], lm_data_improve_hw3_3$coefficients[2], col='red', lty='da
```

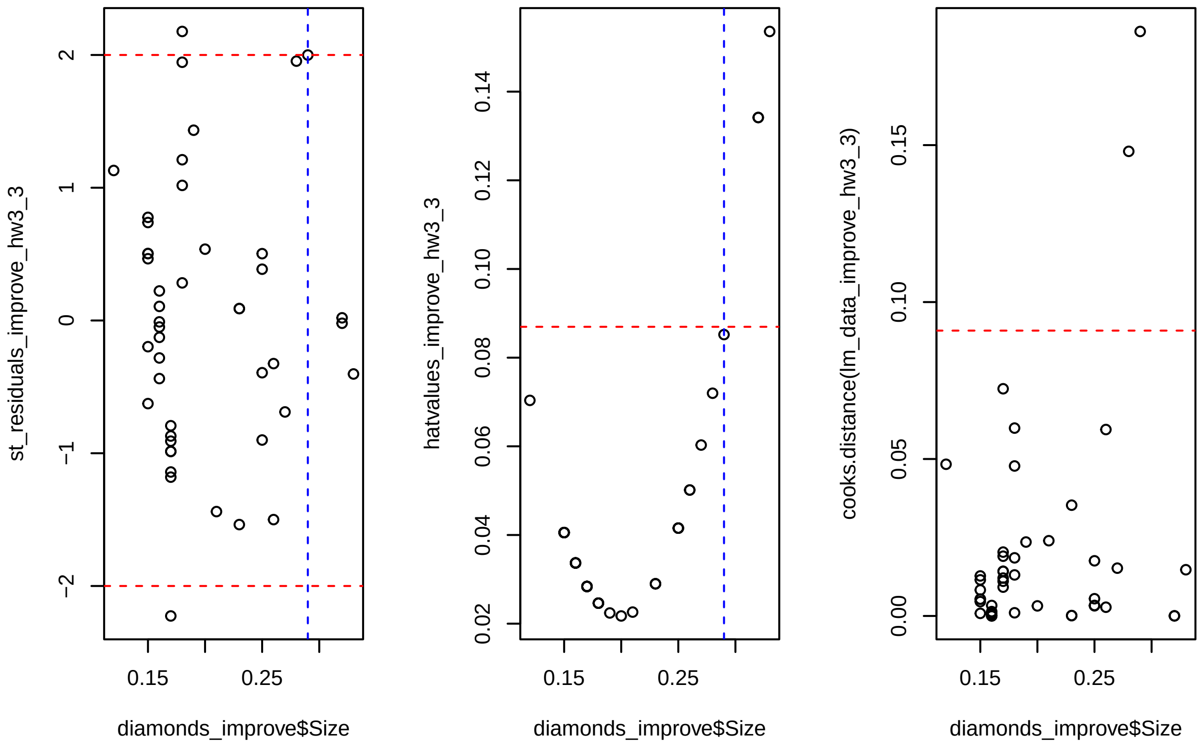



```
###

s_improve_hw3_3 <- (sum((lm_data_improve_hw3_3$residuals - mean(lm_data_improve_hw3_3$residuals))^2) /

hatvalues_improve_hw3_3 <- hatvalues(lm_data_improve_hw3_3)

st_residuals_improve_hw3_3 <- lm_data_improve_hw3_3$residuals / (s_improve_hw3_3 * (1-hatvalues_improve_

lm_data_residual_improve_hw3_3 <- lm((((st_residuals_improve_hw3_3)^2)^(1/2))^(1/2)~diamonds_improve$Si

###

par(mfrow=c(1,3))
plot(diamonds_improve$Size, st_residuals_improve_hw3_3)
abline(2,0,col='red', lty='dashed')
abline(-2,0,col='red', lty='dashed')

plot(diamonds_improve$Size, (((st_residuals_improve_hw3_3)^2)^(1/2))^(1/2))
abline(lm_data_residual_improve_hw3_3$coefficients[1], lm_data_residual_improve_hw3_3$coefficients[2],

qqnorm(st_residuals_improve_hw3_3)
qqline(st_residuals_improve_hw3_3, col='red', lty='dashed')
```



Then we can get this outcome. If we check the power of justification,

```
par(mfrow=c(1,3))
plot(diamonds_improve$Size, st_residuals_improve_hw3_3)
abline(2, 0, col='red', lty='dashed')
abline(-2, 0, col='red', lty='dashed')
abline(v=0.29, col='blue', lty='dashed')

plot(diamonds_improve$Size, hatvalues_improve_hw3_3)
abline(4/length(diamonds_improve$Size),0, col='red', lty='dashed')
abline(v=0.29, col='blue', lty='dashed')

plot(diamonds_improve$Size, cooks.distance(lm_data_improve_hw3_3))
abline(4/(length(diamonds_improve$Size)-2),0,col='red', lty='dashed')
```



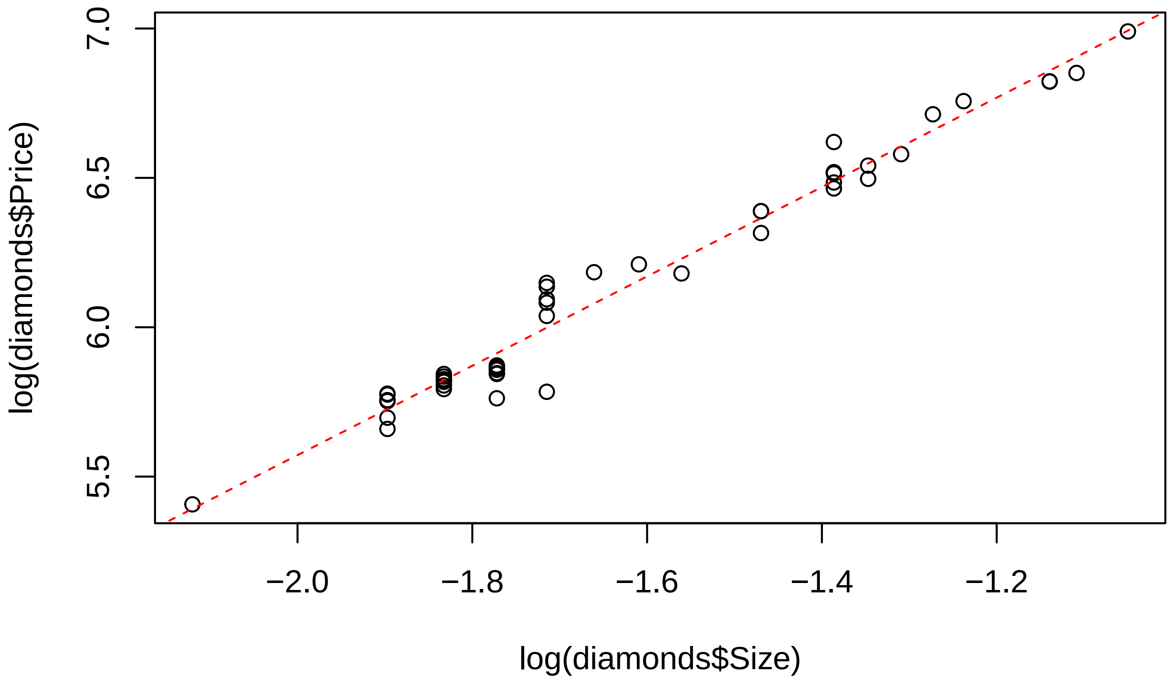
Part 1 - (b)

The number of data are small.

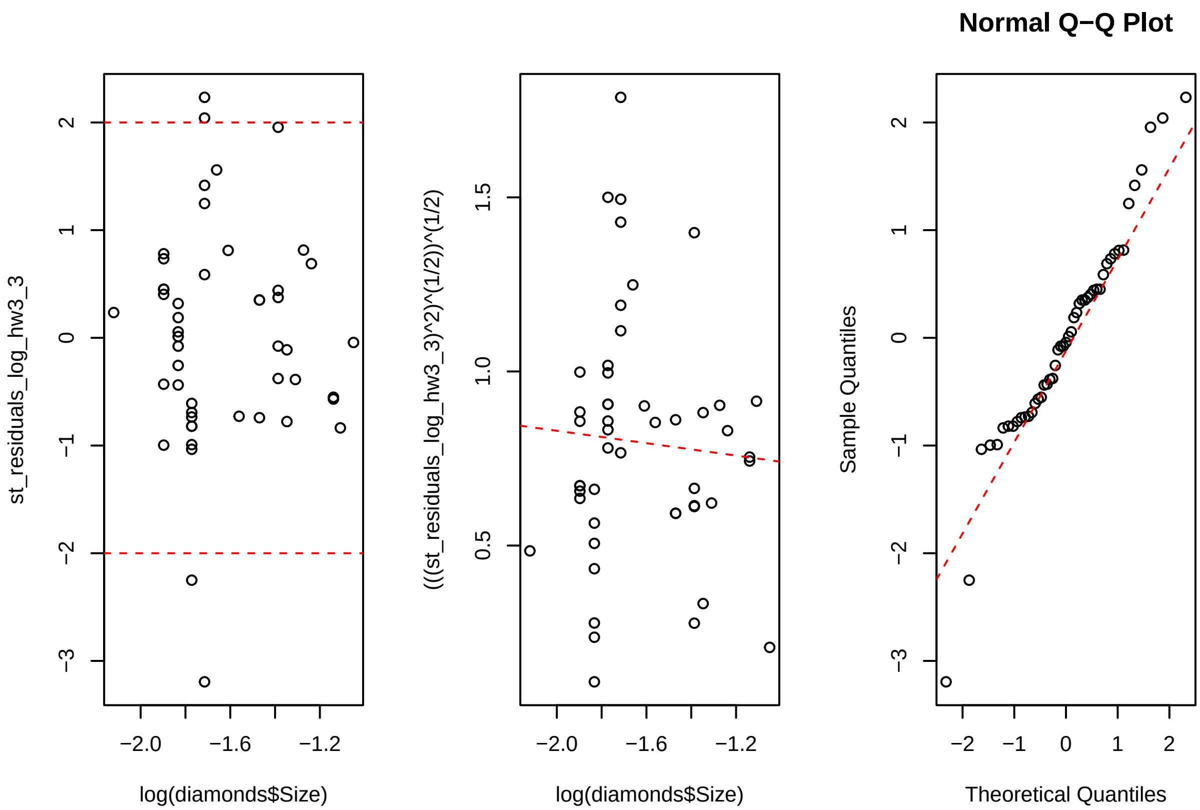
Part 2 - (a)

We can use log-scale SLR model.

```
lm_data_log_hw3_3 <- lm(log(diamonds$Price)~log(diamonds$Size), data=diamonds)

plot(log(diamonds$Size), log(diamonds$Price))
abline(lm_data_log_hw3_3$coefficients[1], lm_data_log_hw3_3$coefficients[2], col='red', lty='dashed')
```

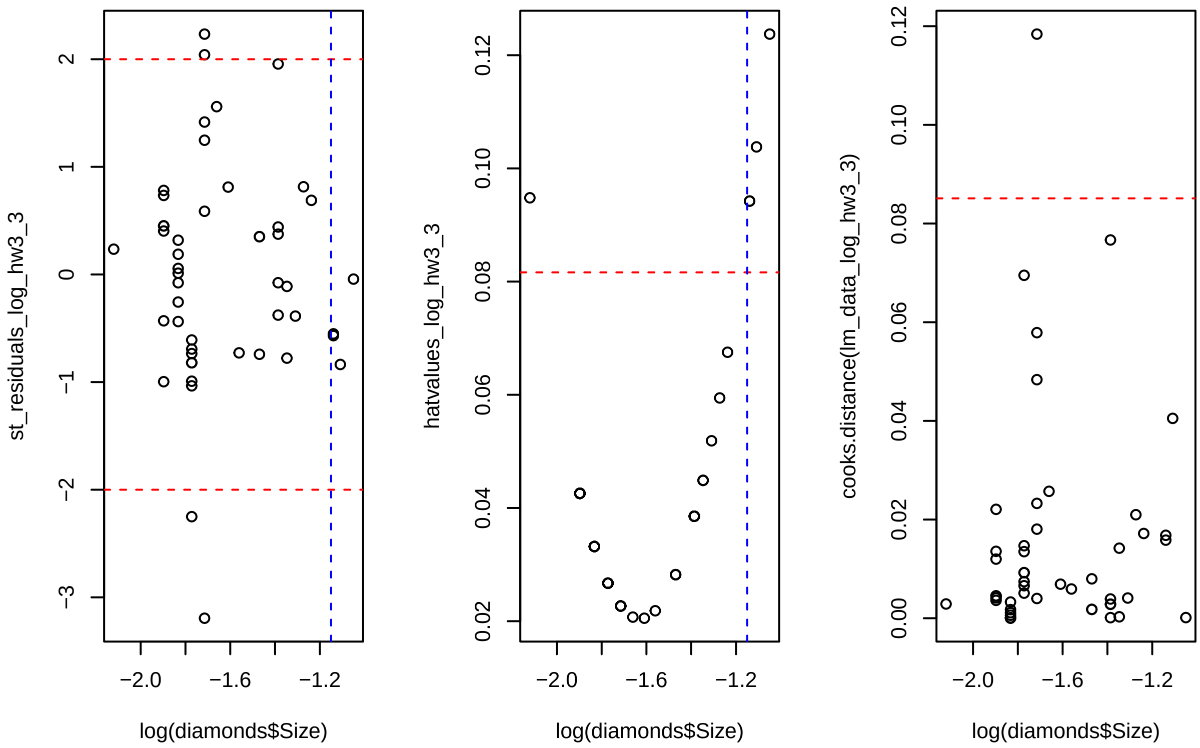



```
###

s_log_hw3_3 <- (sum((lm_data_log_hw3_3$residuals - mean(lm_data_log_hw3_3$residuals))^2) / (length(diam

hatvalues_log_hw3_3 <- hatvalues(lm_data_log_hw3_3)

st_residuals_log_hw3_3 <- lm_data_log_hw3_3$residuals / (s_log_hw3_3 * (1-hatvalues_log_hw3_3)^(1/2))

lm_data_residual_log_hw3_3 <- lm((((st_residuals_log_hw3_3)^2)^(1/2))^(1/2)~log(diamonds$Size), data=di

###

par(mfrow=c(1,3))
plot(log(diamonds$Size), st_residuals_log_hw3_3)
abline(2,0,col='red', lty='dashed')
abline(-2,0,col='red', lty='dashed')

plot(log(diamonds$Size), (((st_residuals_log_hw3_3)^2)^(1/2))^(1/2))
abline(lm_data_residual_log_hw3_3$coefficients[1], lm_data_residual_log_hw3_3$coefficients[2], col='red

qqnorm(st_residuals_log_hw3_3)
qqline(st_residuals_log_hw3_3, col='red', lty='dashed')
```



When we check the power of justification,

```
par(mfrow=c(1,3))
plot(log(diamonds$Size), st_residuals_log_hw3_3)
abline(2, 0, col='red', lty='dashed')
abline(-2, 0, col='red', lty='dashed')
abline(v=-1.15, col='blue', lty='dashed')

plot(log(diamonds$Size), hatvalues_log_hw3_3)
abline(4/length(diamonds$Size),0, col='red', lty='dashed')
abline(v=-1.15, col='blue', lty='dashed')

plot(log(diamonds$Size), cooks.distance(lm_data_log_hw3_3))
abline(4/(length(diamonds$Size)-2),0,col='red', lty='dashed')
```

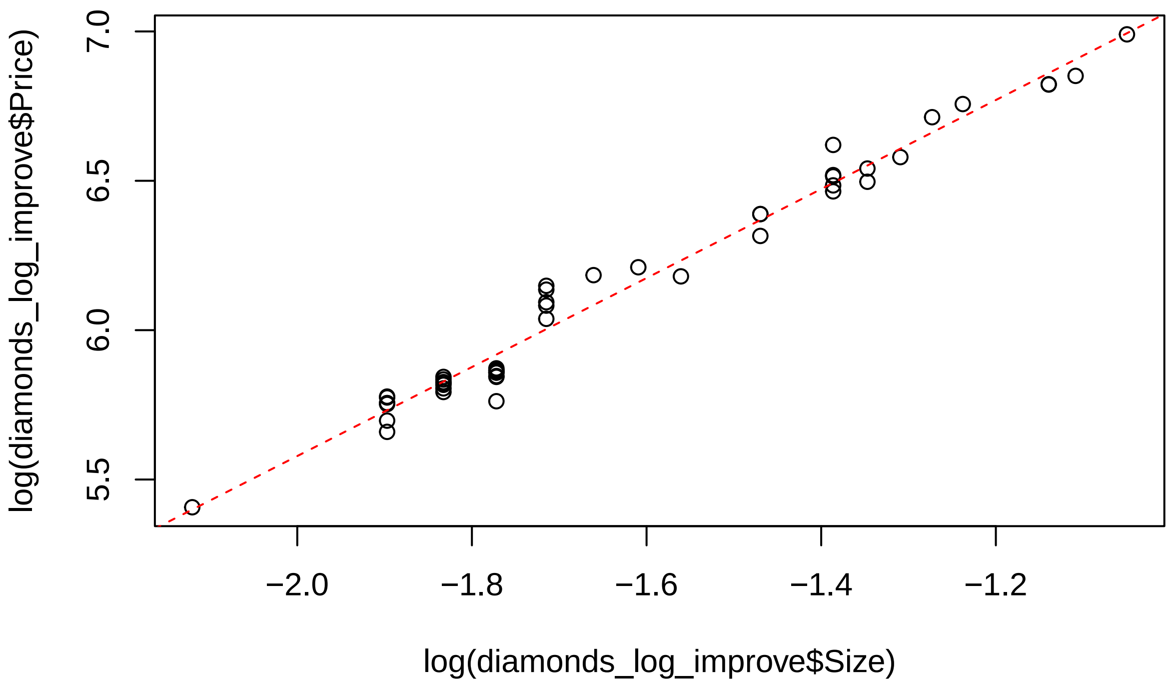


Thus, there are no bad leverage points.
If we eliminate the data having big cook's distance,

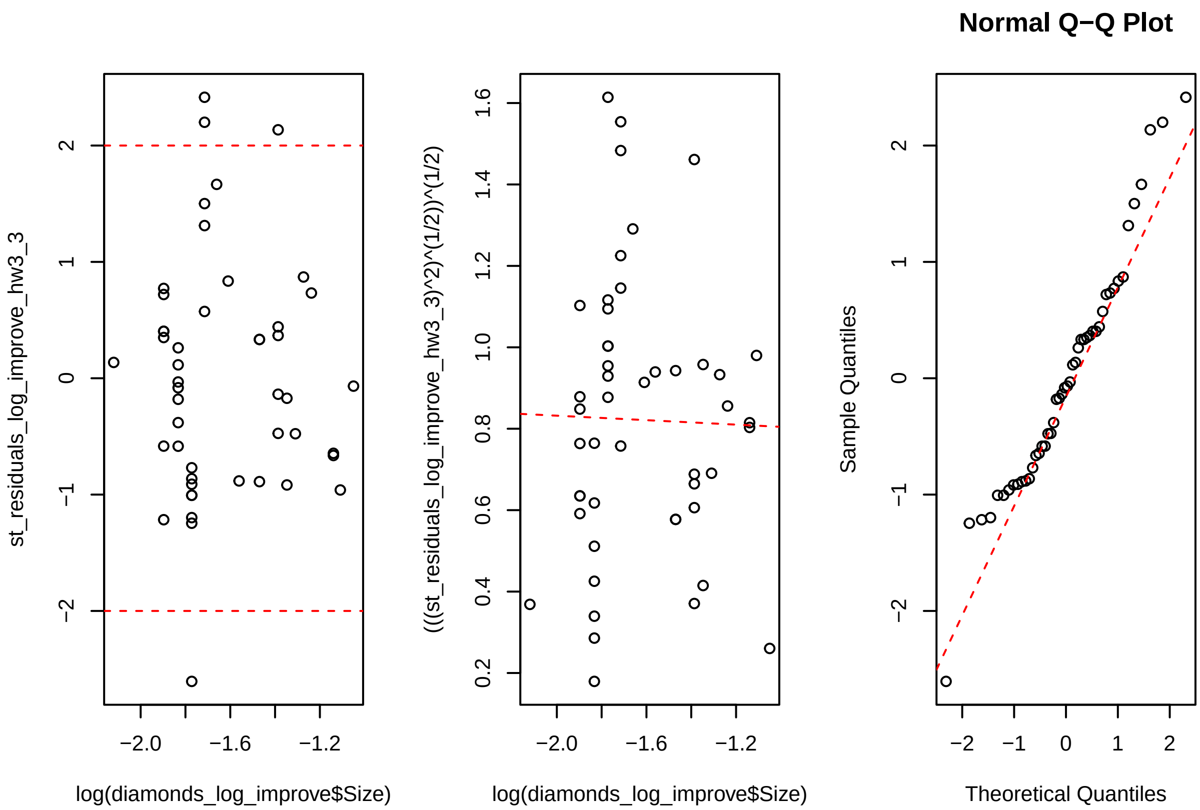
```
cooks.distance(lm_data_log_hw3_3)[cooks.distance(lm_data_log_hw3_3) > 4/(length(diamonds$Size)-2)]
```

```
##         4
## 0.1183951
```

```
diamonds_log_improve <- diamonds[c(-4),]

lm_data_log_improve_hw3_3 <- lm(log(diamonds_log_improve$Price)~log(diamonds_log_improve$Size), data=di

plot(log(diamonds_log_improve$Size), log(diamonds_log_improve$Price))
abline(lm_data_log_improve_hw3_3$coefficients[1], lm_data_log_improve_hw3_3$coefficients[2], col='red',
```




```
###

s_log_improve_hw3_3 <- (sum((lm_data_log_improve_hw3_3$residuals - mean(lm_data_log_improve_hw3_3$residu

hatvalues_log_improve_hw3_3 <- hatvalues(lm_data_log_improve_hw3_3)

st_residuals_log_improve_hw3_3 <- lm_data_log_improve_hw3_3$residuals / (s_log_improve_hw3_3 * (1-hatval

lm_data_residual_log_improve_hw3_3 <- lm((((st_residuals_log_improve_hw3_3)^2)^(1/2))^(1/2)~log(diamonds

###

par(mfrow=c(1,3))
plot(log(diamonds_log_improve$Size), st_residuals_log_improve_hw3_3)
abline(2,0,col='red', lty='dashed')
abline(-2,0,col='red', lty='dashed')

plot(log(diamonds_log_improve$Size), (((st_residuals_log_improve_hw3_3)^2)^(1/2))^(1/2))
abline(lm_data_residual_log_improve_hw3_3$coefficients[1], lm_data_residual_log_improve_hw3_3$coefficien

qqnorm(st_residuals_log_improve_hw3_3)
qqline(st_residuals_log_improve_hw3_3, col='red', lty='dashed')
```



Part 2 - (b)

The number of data are small.

Part 3

Part B has a better model, because the regression of sum of squared of standardized residual is flatter.