

# 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, 
$$\begin{cases} h_{ii} > \frac{4}{n} \rightarrow \frac{4}{17} \approx 0.235 \\ |\gamma_i| > 2 \\ D_i > \frac{4}{n-2} \rightarrow \frac{4}{15} \approx 0.267 \end{cases},$$

```
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)

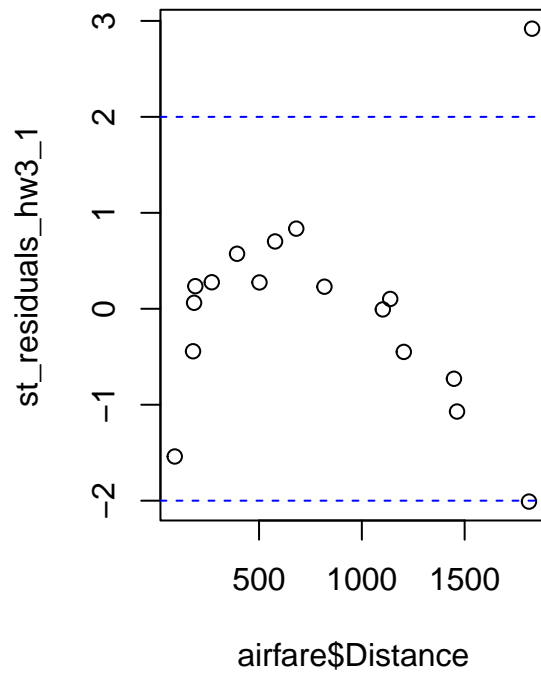
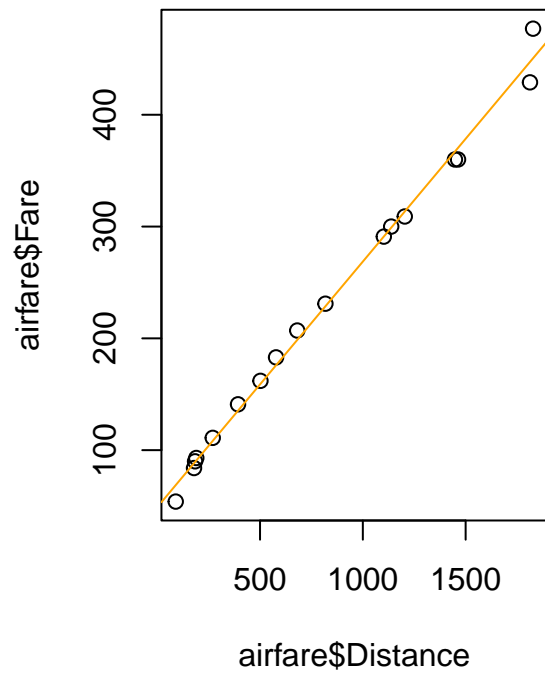
st_residuals_hw3_1 <- lm_data_hw3_1$residuals / (s_hw3_1 * (1-hatvalues_hw3_1)^(1/2))

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

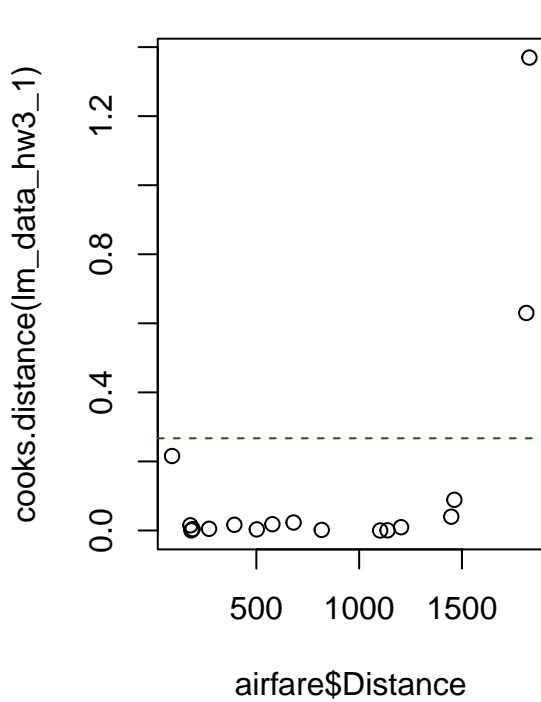
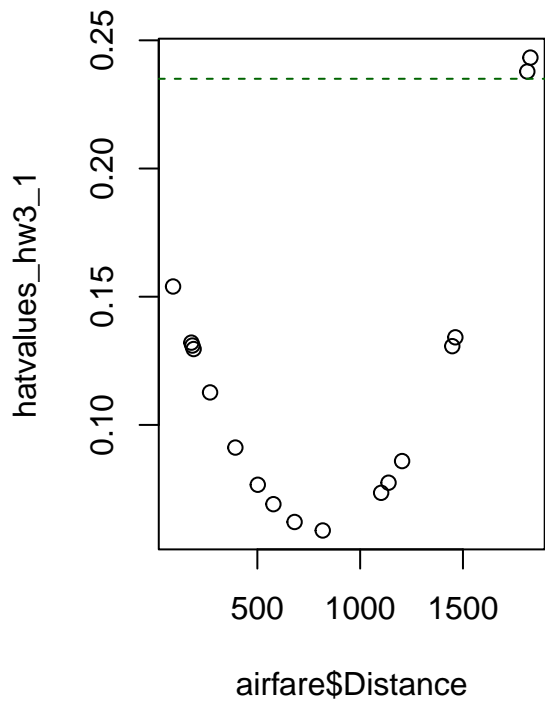
par(mfcol=c(1,2))
plot(airfare$Distance, airfare$Fare)
abline(lm_data_hw3_1$coefficients[1], lm_data_hw3_1$coefficients[2], col='orange')

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



```
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)

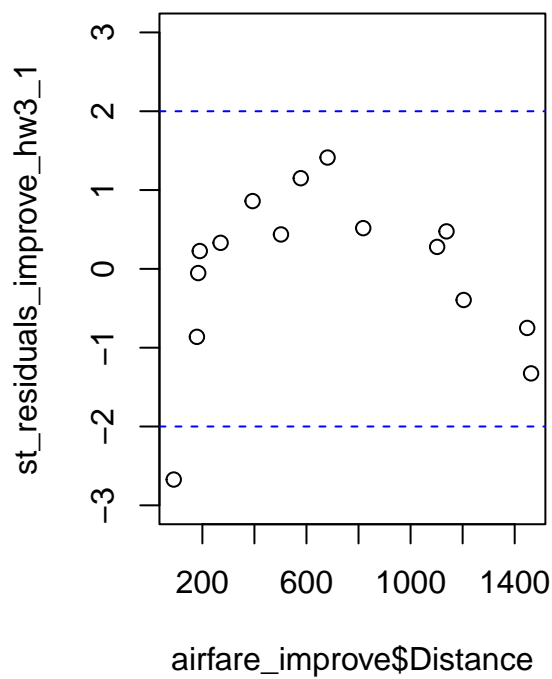
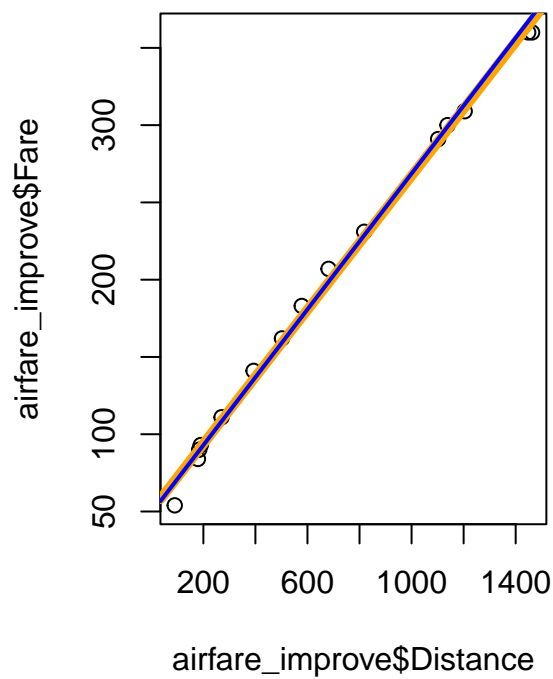
st_residuals_improve_hw3_1 <- lm_data_improve_hw3_1$residuals / (s_improve_hw3_1 * (1-hatvalues_improve

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=2)
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')
```

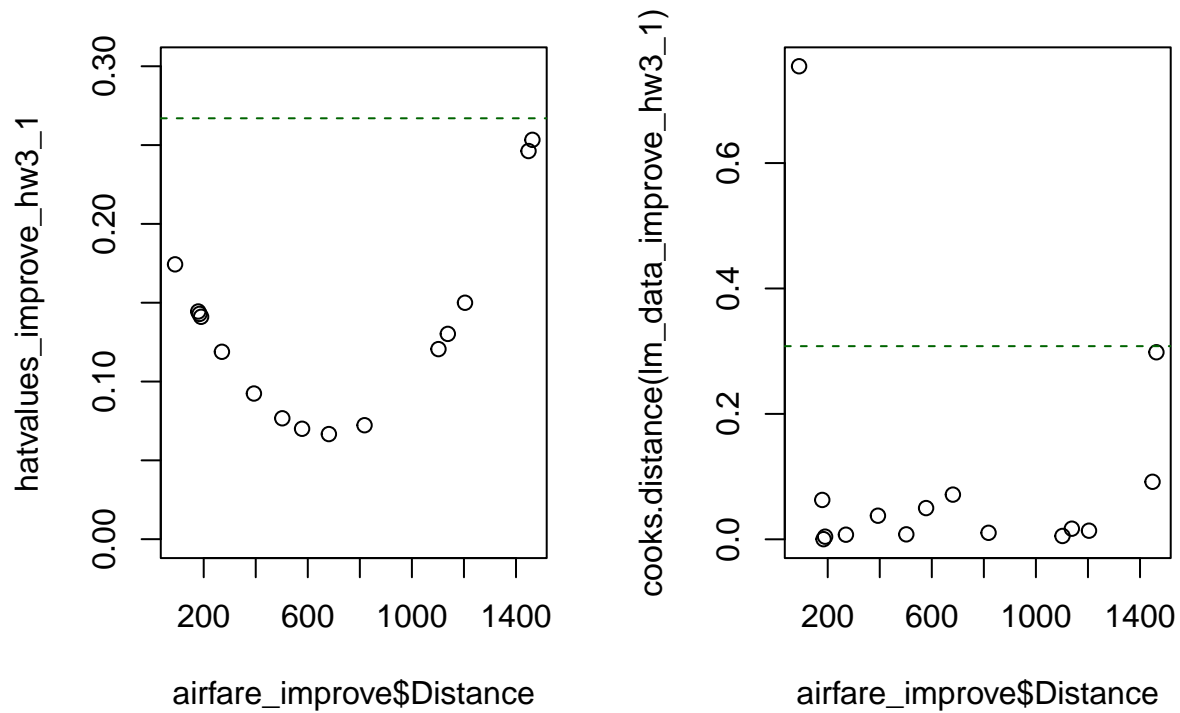


```

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 =  $\beta_0 + \beta_1 * \text{Dealer Cost} + e$ .

(a)

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

*Since the model explains just more than 99.8% of the variability 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 producing 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) - 2))

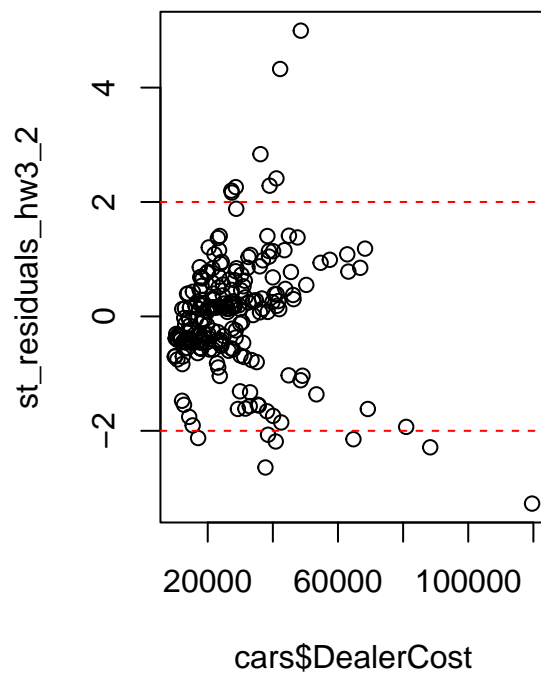
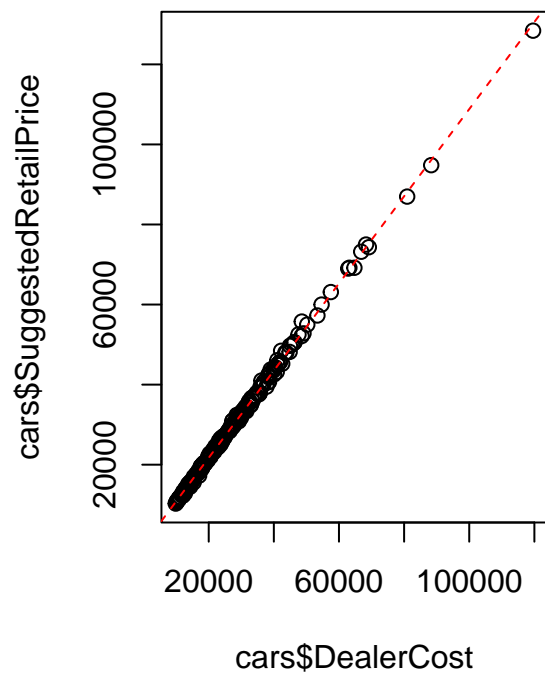
hatvalues_hw3_2 <- hatvalues(lm_data_hw3_2)

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')

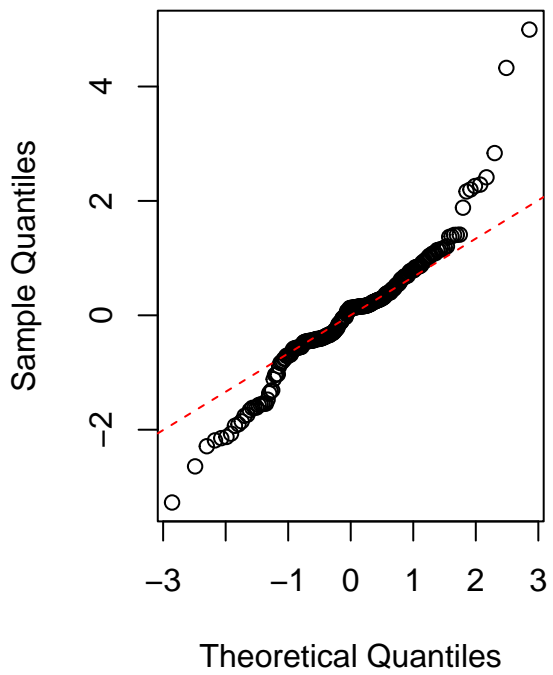
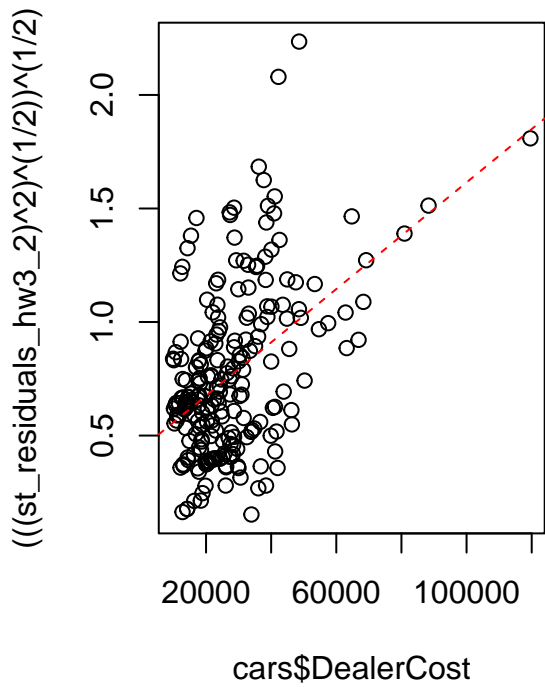
plot(cars$DealerCost, st_residuals_hw3_2)
abline(2,0, col='red', lty='dashed')
abline(-2,0,col='red', lty='dashed')
```



```
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='dashed')

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

### Normal Q-Q Plot

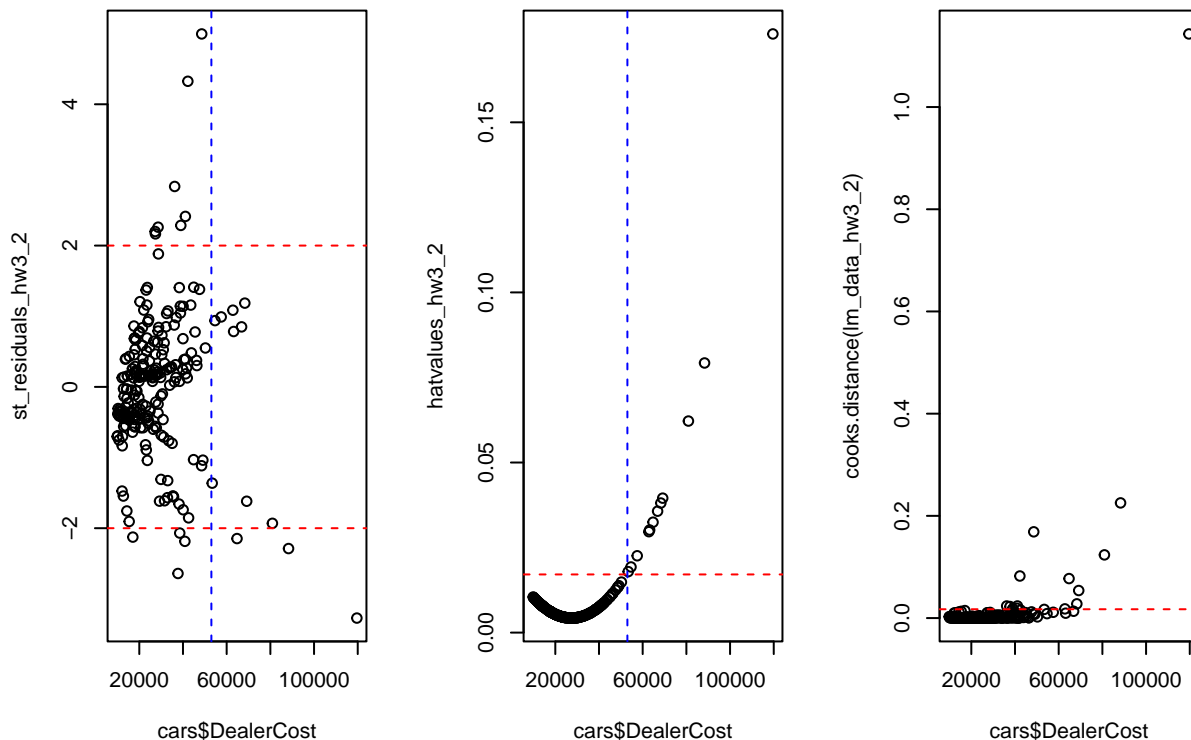


There are three methods to provide a detailed critique,  $\begin{cases} h_{ii} > \frac{4}{n} \rightarrow \frac{4}{234} \approx 0.017 \\ |\gamma_i| > 2 \\ D_i > \frac{4}{n-2} \rightarrow \frac{4}{232} \approx 0.0172 \end{cases}$ ,

```
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, -230, -231, -232, -233, -234, -235, -236, -237, -238, -239, -240, -241, -242, -243, -244, -245, -246, -247, -248, -249, -250, -251, -252, -253, -254, -255, -256, -257, -258, -259, -260, -261, -262, -263, -264, -265, -266, -267, -268, -269, -270, -271, -272, -273, -274, -275, -276, -277, -278, -279, -280, -281, -282, -283, -284, -285, -286, -287, -288, -289, -290, -291, -292, -293, -294, -295, -296, -297, -298, -299, -300, -301, -302, -303, -304, -305, -306, -307, -308, -309, -310, -311, -312, -313, -314, -315, -316, -317, -318, -319, -320, -321, -322, -323, -324, -325, -326, -327, -328, -329, -330, -331, -332, -333, -334, -335, -336, -337, -338, -339, -340, -341, -342, -343, -344, -345, -346, -347, -348, -349, -350, -351, -352, -353, -354, -355, -356, -357, -358, -359, -360, -361, -362, -363, -364, -365, -366, -367, -368, -369, -370, -371, -372, -373, -374, -375, -376, -377, -378, -379, -380, -381, -382, -383, -384, -385, -386, -387, -388, -389, -390, -391, -392, -393, -394, -395, -396, -397, -398, -399, -400, -401, -402, -403, -404, -405, -406, -407, -408, -409, -410, -411, -412, -413, -414, -415, -416, -417, -418, -419, -420, -421, -422, -423, -424, -425, -426, -427, -428, -429, -430, -431, -432, -433, -434, -435, -436, -437, -438, -439, -440, -441, -442, -443, -444, -445, -446, -447, -448, -449, -450, -451, -452, -453, -454, -455, -456, -457, -458, -459, -460, -461, -462, -463, -464, -465, -466, -467, -468, -469, -470, -471, -472, -473, -474, -475, -476, -477, -478, -479, -480, -481, -482, -483, -484, -485, -486, -487, -488, -489, -490, -491, -492, -493, -494, -495, -496, -497, -498, -499, -500, -501, -502, -503, -504, -505, -506, -507, -508, -509, -510, -511, -512, -513, -514, -515, -516, -517, -518, -519, -520, -521, -522, -523, -524, -525, -526, -527, -528, -529, -530, -531, -532, -533, -534, -535, -536, -537, -538, -539, -540, -541, -542, -543, -544, -545, -546, -547, -548, -549, -550, -551, -552, -553, -554, -555, -556, -557, -558, -559, -560, -561, -562, -563, -564, -565, -566, -567, -568, -569, -570, -571, -572, -573, -574, -575, -576, -577, -578, -579, -580, -581, -582, -583, -584, -585, -586, -587, -588, -589, -590, -591, -592, -593, -594, -595, -596, -597, -598, -599, -600, -601, -602, -603, -604, -605, -606, -607, -608, -609, -610, -611, -612, -613, -614, -615, -616, -617, -618, -619, -620, -621, -622, -623, -624, -625, -626, -627, -628, -629, -630, -631, -632, -633, -634, -635, -636, -637, -638, -639, -640, -641, -642, -643, -644, -645, -646, -647, -648, -649, -650, -651, -652, -653, -654, -655, -656, -657, -658, -659, -660, -661, -662, -663, -664, -665, -666, -667, -668, -669, -670, -671, -672, -673, -674, -675, -676, -677, -678, -679, -680, -681, -682, -683, -684, -685, -686, -687, -688, -689, -690, -691, -692, -693, -694, -695, -696, -697, -698, -699, -700, -701, -702, -703, -704, -705, -706, -707, -708, -709, -710, -711, -712, -713, -714, -715, -716, -717, -718, -719, -720, -721, -722, -723, -724, -725, -726, -727, -728, -729, -730, -731, -732, -733, -734, -735, -736, -737, -738, -739, -740, -741, -742, -743, -744, -745, -746, -747, -748, -749, -750, -751, -752, -753, -754, -755, -756, -757, -758, -759, -760, -761, -762, -763, -764, -765, -766, -767, -768, -769, -770, -771, -772, -773, -774, -775, -776, -777, -778, -779, -780, -781, -782, -783, -784, -785, -786, -787, -788, -789, -790, -791, -792, -793, -794, -795, -796, -797, -798, -799, -800, -801, -802, -803, -804, -805, -806, -807, -808, -809, -810, -811, -812, -813, -814, -815, -816, -817, -818, -819, -820, -821, -822, -823, -824, -825, -826, -827, -828, -829, -830, -831, -832, -833, -834, -835, -836, -837, -838, -839, -840, -841, -842, -843, -844, -845, -846, -847, -848, -849, -850, -851, -852, -853, -854, -855, -856, -857, -858, -859, -860, -861, -862, -863, -864, -865, -866, -867, -868, -869, -870, -871, -872, -873, -874, -875, -876, -877, -878, -879, -880, -881, -882, -883, -884, -885, -886, -887, -888, -889, -890, -891, -892, -893, -894, -895, -896, -897, -898, -899, -900, -901, -902, -903, -904, -905, -906, -907, -908, -909, -910, -911, -912, -913, -914, -915, -916, -917, -918, -919, -920, -921, -922, -923, -924, -925, -926, -927, -928, -929, -930, -931, -932, -933, -934, -935, -936, -937, -938, -939, -940, -941, -942, -943, -944, -945, -946, -947, -948, -949, -950, -951, -952, -953, -954, -955, -956, -957, -958, -959, -960, -961, -962, -963, -964, -965, -966, -967, -968, -969, -970, -971, -972, -973, -974, -975, -976, -977, -978, -979, -980, -981, -982, -983, -984, -985, -986, -987, -988, -989, -990, -991, -992, -993, -994, -995, -996, -997, -998, -999, 1000], 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)

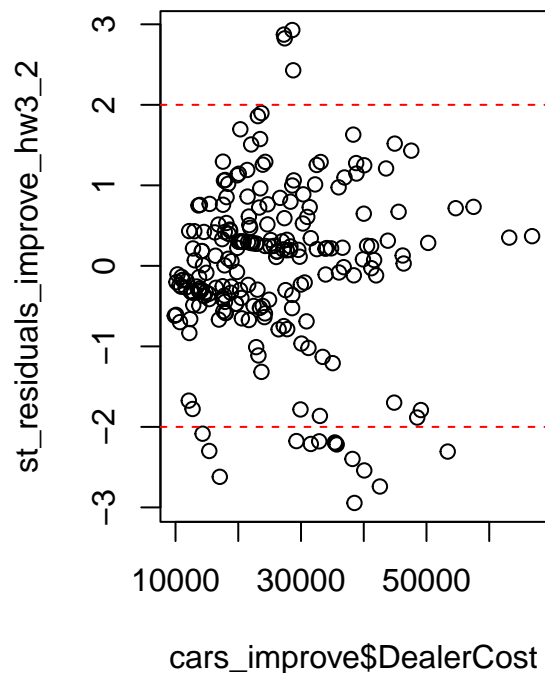
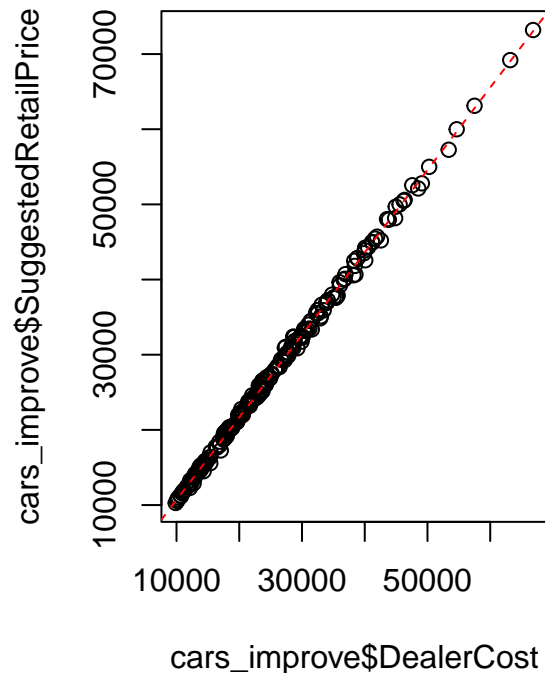
st_residuals_improve_hw3_2 <- lm_data_improve_hw3_2$residuals / (s_improve_hw3_2 * (1-hatvalues_improve_hw3_2))

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

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='dashed')

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

```



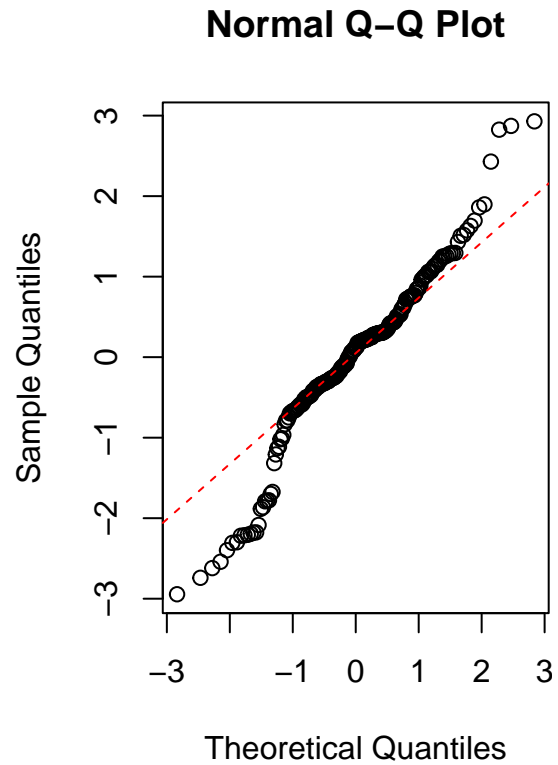
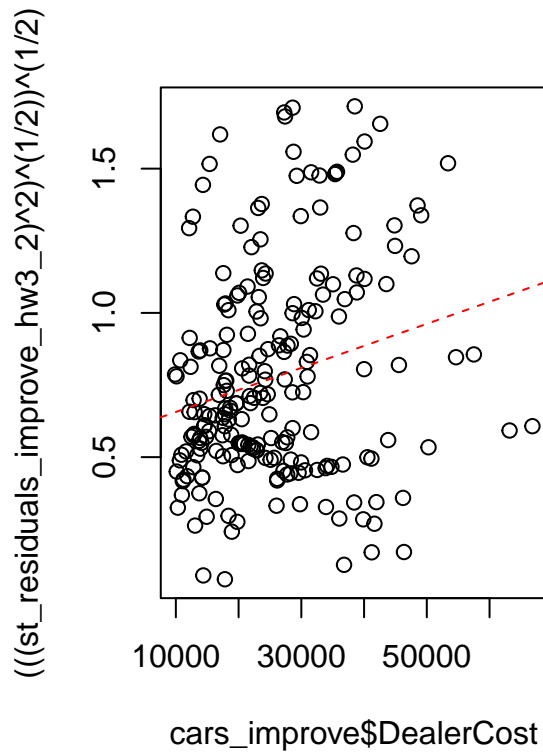
```

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], col='red', lty='dashed')

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

```

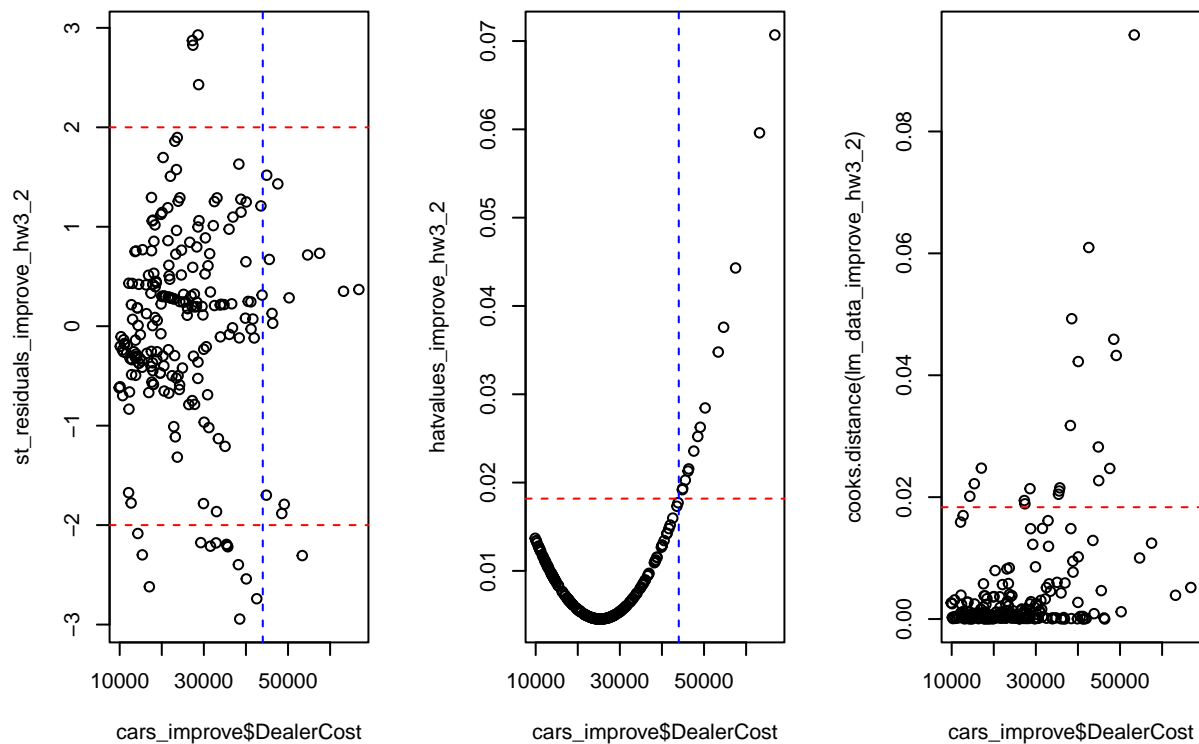




```
par(mfrow=c(1,3))
plot(cars_improve$DealerCost, st_residuals_improve_hw3_2)
abline(2, 0, col='red', lty='dashed')
abline(-2, 0, col='red', lty='dashed')
abline(v=44000, col='blue', lty='dashed')

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 square root of standardized residual has steep slope. → 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$SuggestedRetailPrice) - 2))^(1/2)

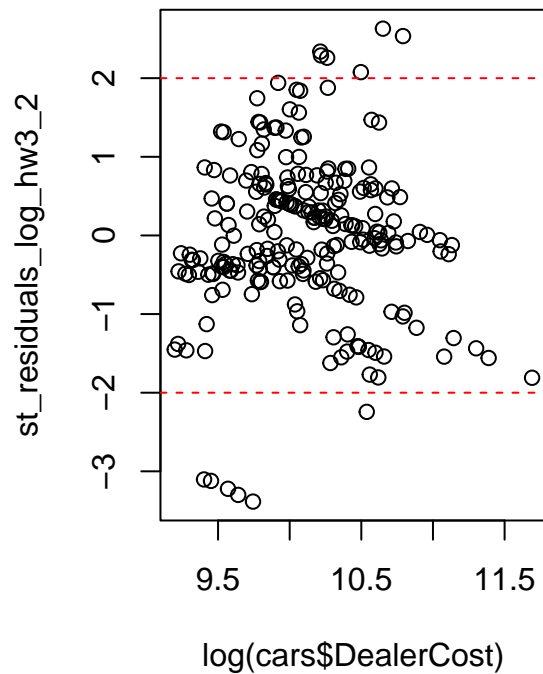
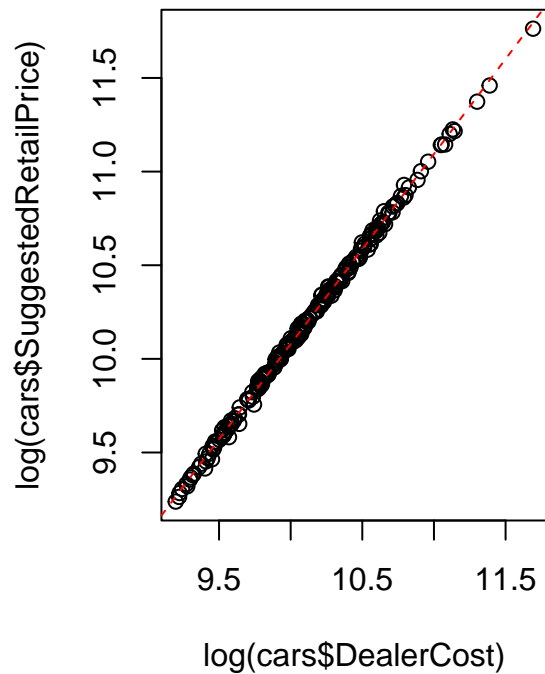
hatvalues_log_hw3_2 <- hatvalues(lm_data_log_hw3_2)

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=cars)

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')

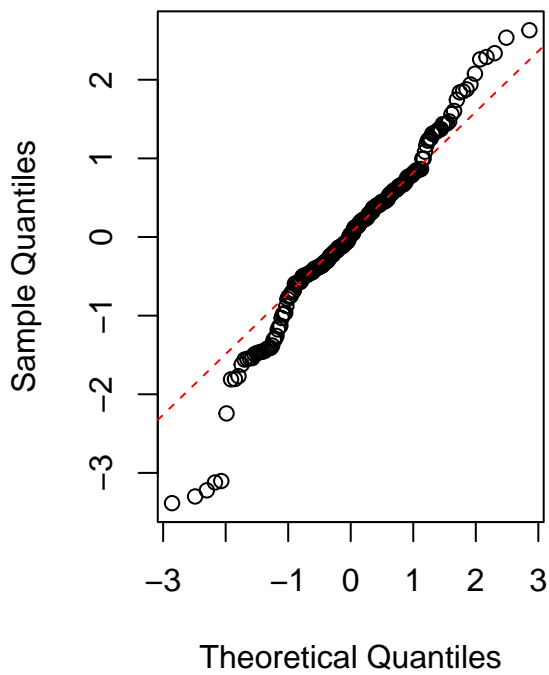
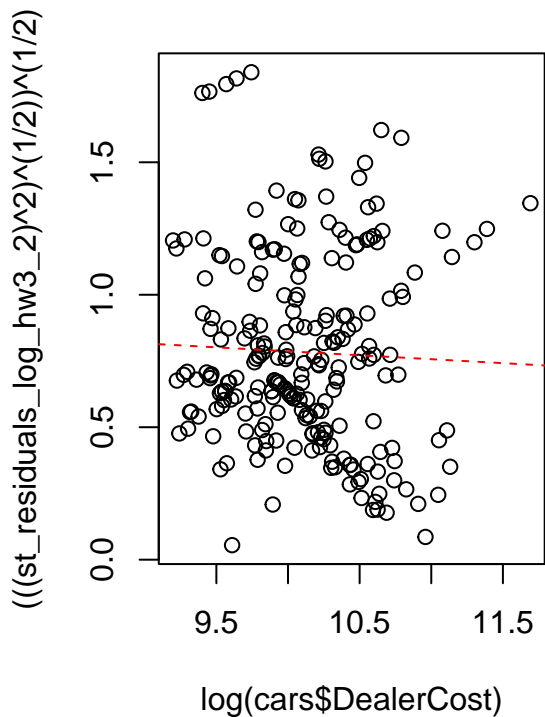
plot(log(cars$DealerCost), st_residuals_log_hw3_2)
abline(2,0, col='red', lty='dashed')
abline(-2,0,col='red', lty='dashed')
```



```
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')
```

### Normal Q-Q Plot



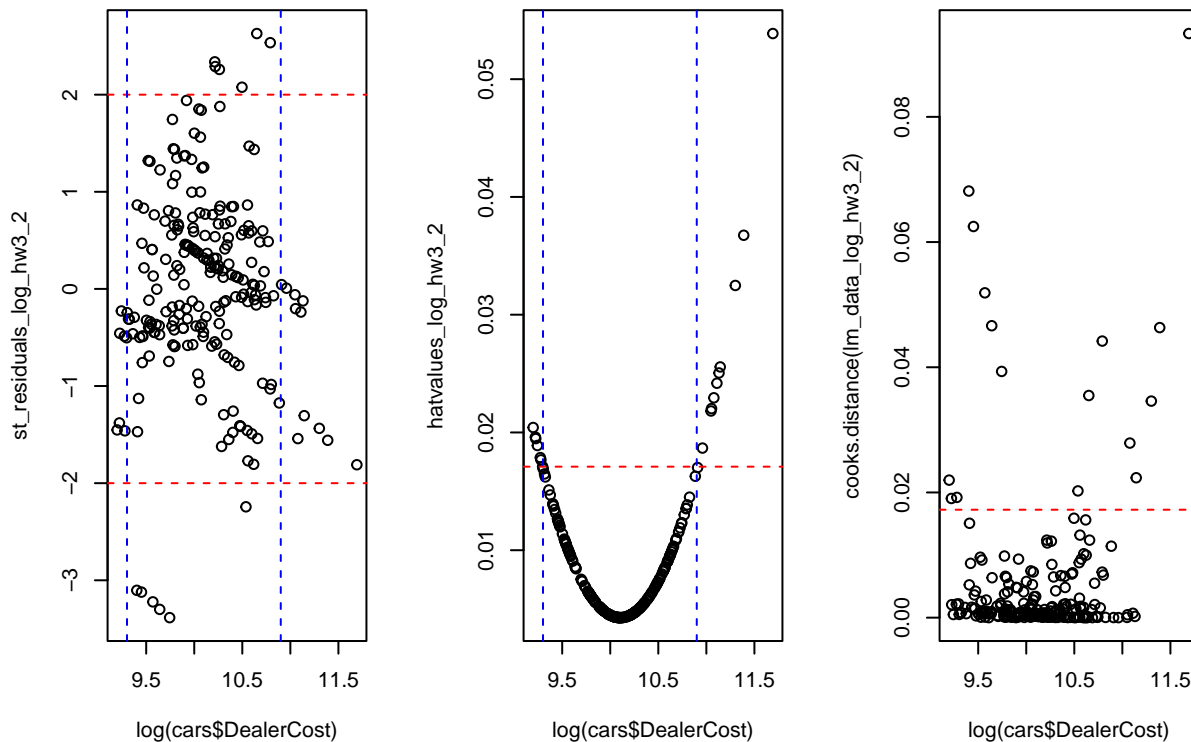
Thus, the log-scale model is more fitted than above one. This is because

- (1) More  $\gamma_i$  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$SuggestedRetailPrice)~log(cars_log_improve$DealerCost))

s_log_improve_hw3_2 <- (sum((lm_data_log_improve_hw3_2$residuals - mean(lm_data_log_improve_hw3_2$residuals))^2))^0.5

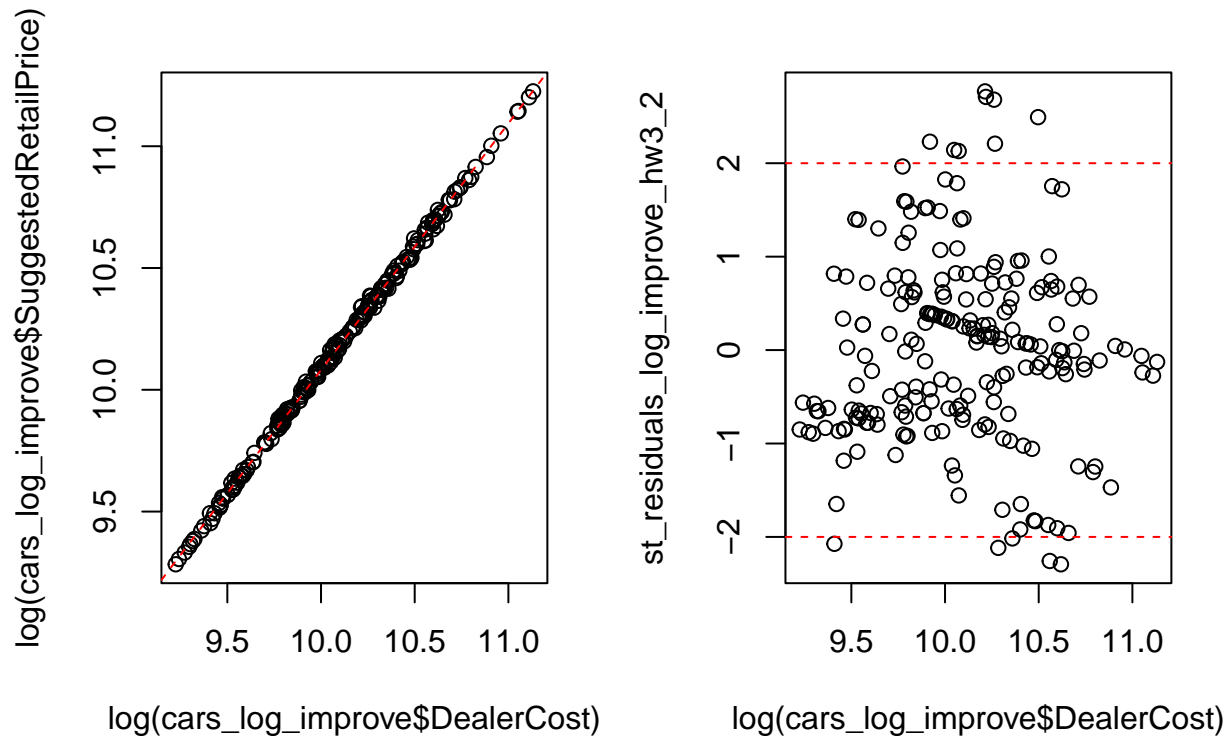
hatvalues_log_improve_hw3_2 <- hatvalues(lm_data_log_improve_hw3_2)

st_residuals_log_improve_hw3_2 <- lm_data_log_improve_hw3_2$residuals / (s_log_improve_hw3_2 * (1-hatvalues_log_improve_hw3_2))

lm_data_residual_log_improve_hw3_2 <- lm((((st_residuals_log_improve_hw3_2)^2)^(1/2))^(1/2)~log(cars_log_improve$DealerCost))

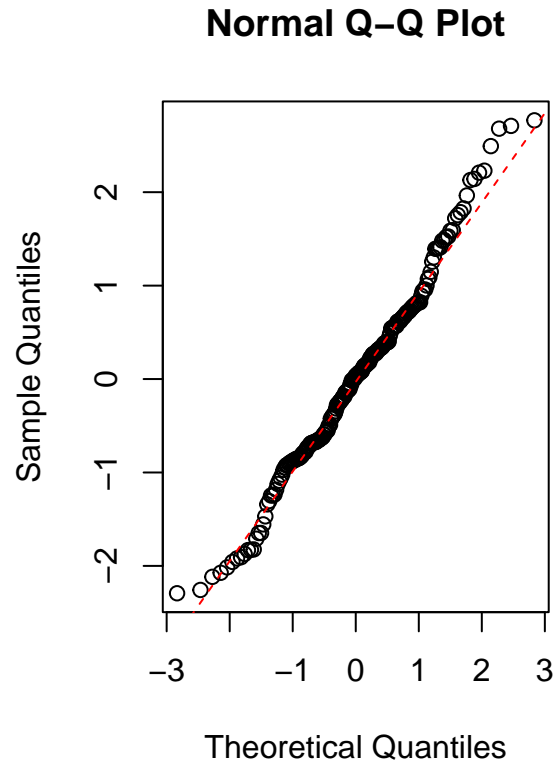
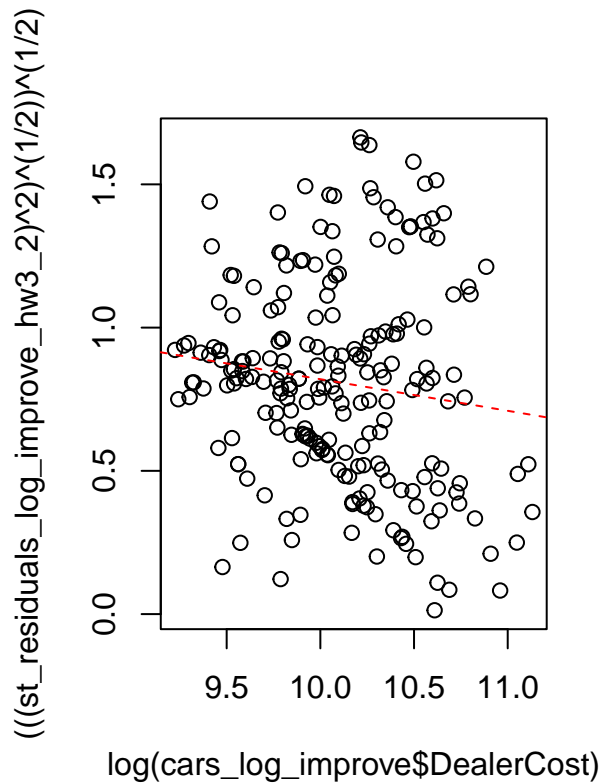
par(mfrow=c(1,2))
plot(log(cars_log_improve$DealerCost), log(cars_log_improve$SuggestedRetailPrice))
abline(lm_data_log_improve_hw3_2$coefficients[1], lm_data_log_improve_hw3_2$coefficients[2], col='red', lty='dashed')

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$coefficients[2], col='red', lty='dashed')

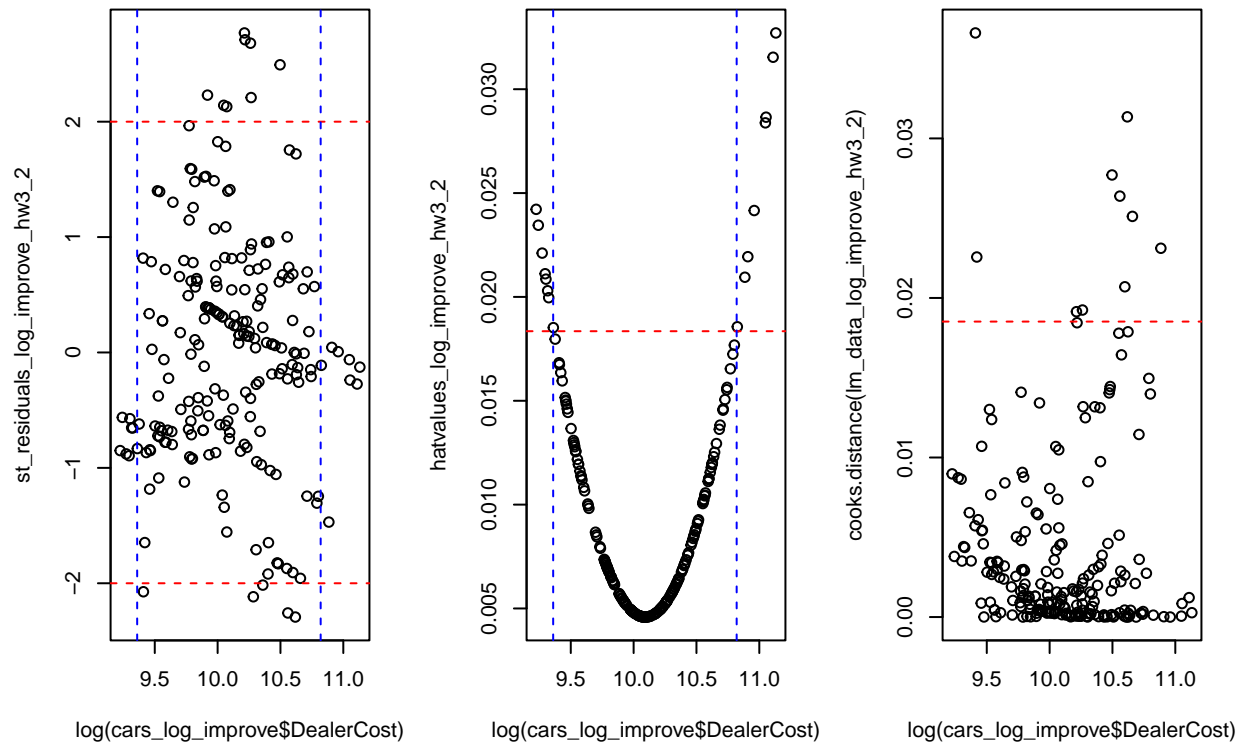
qqnorm(st_residuals_log_improve_hw3_2)
qqline(st_residuals_log_improve_hw3_2, col='red', lty='dashed')
```



```
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)

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

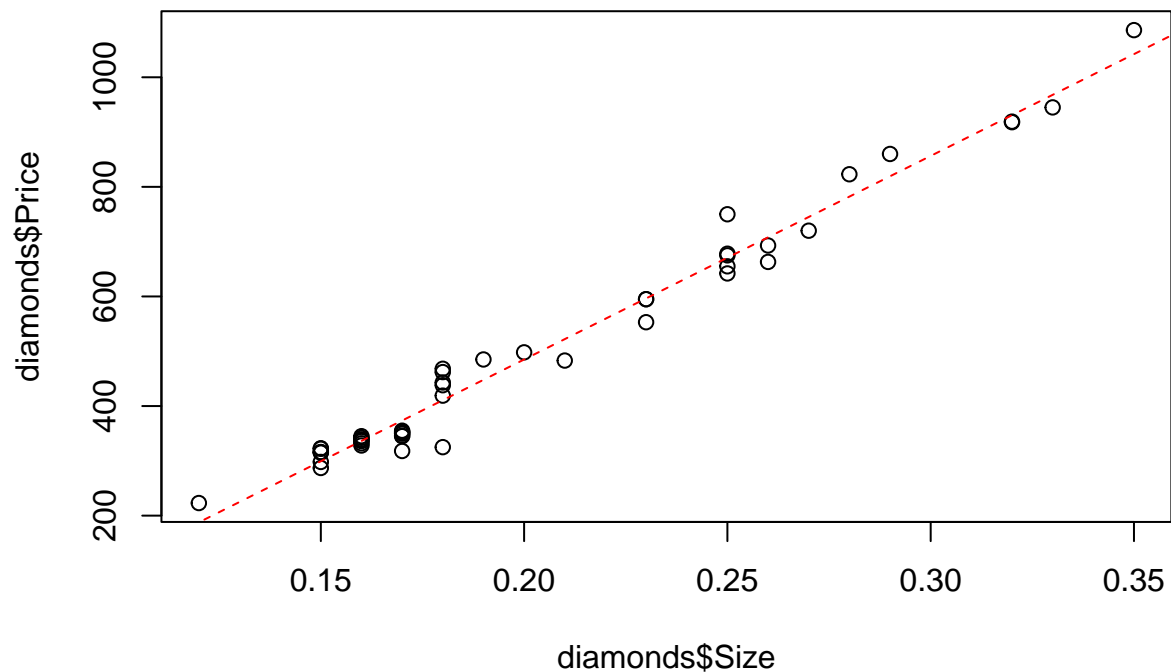
(e)

### 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 Regression  
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')
```



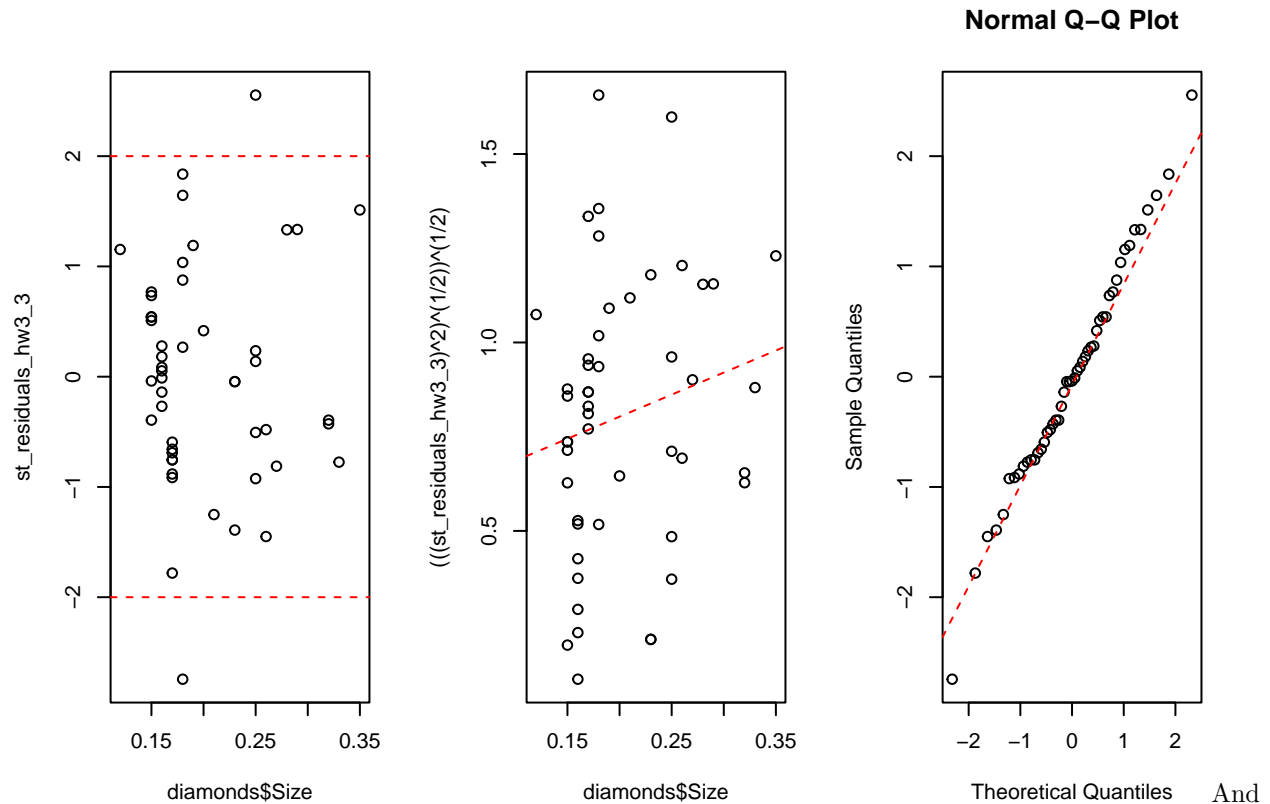
```
###  
s_hw3_3 <- (sum((lm_data_hw3_3$residuals - mean(lm_data_hw3_3$residuals))^2) / (length(diamonds$Price)-  
hatvalues_hw3_3 <- hatvalues(lm_data_hw3_3)  
  
st_residuals_hw3_3 <- lm_data_hw3_3$residuals / (s_hw3_3 * (1-hatvalues_hw3_3)^(1/2))  
  
lm_data_resid_hw3_3 <- lm((((st_residuals_hw3_3)^2)^(1/2))^(1/2)~diamonds$Size, 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='dashed')

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

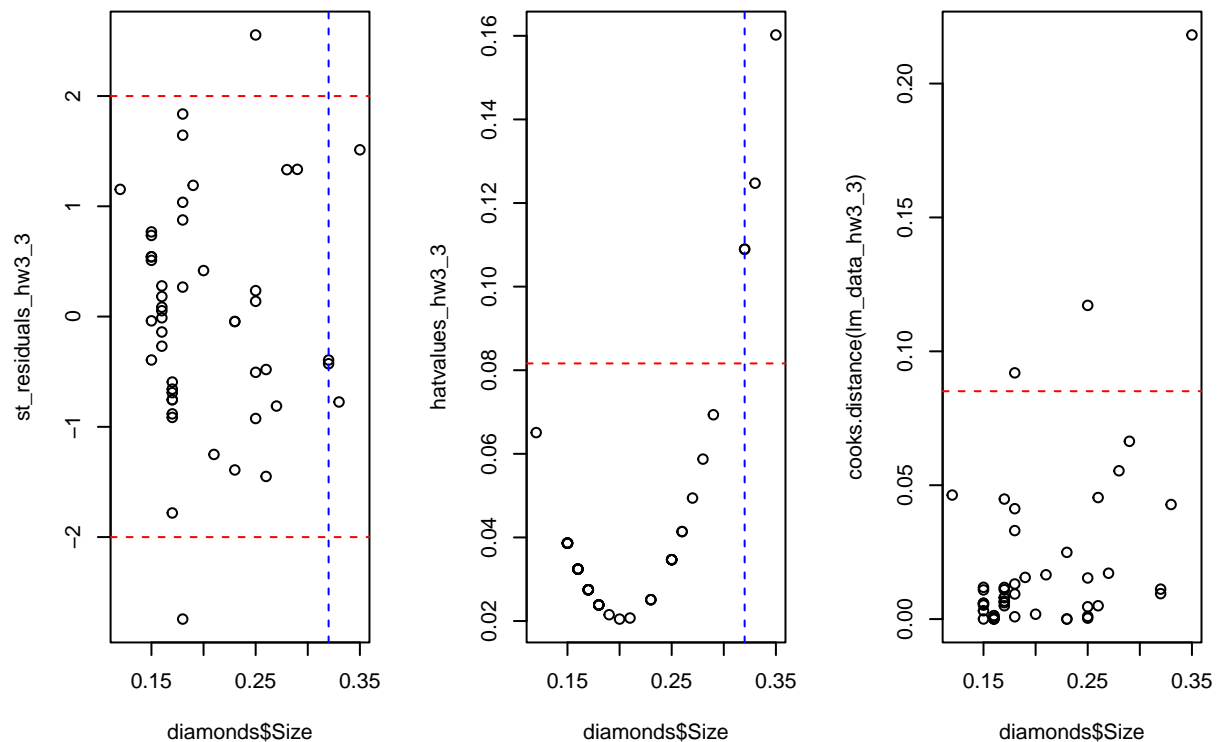


when we check our power of justification,

```
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
```

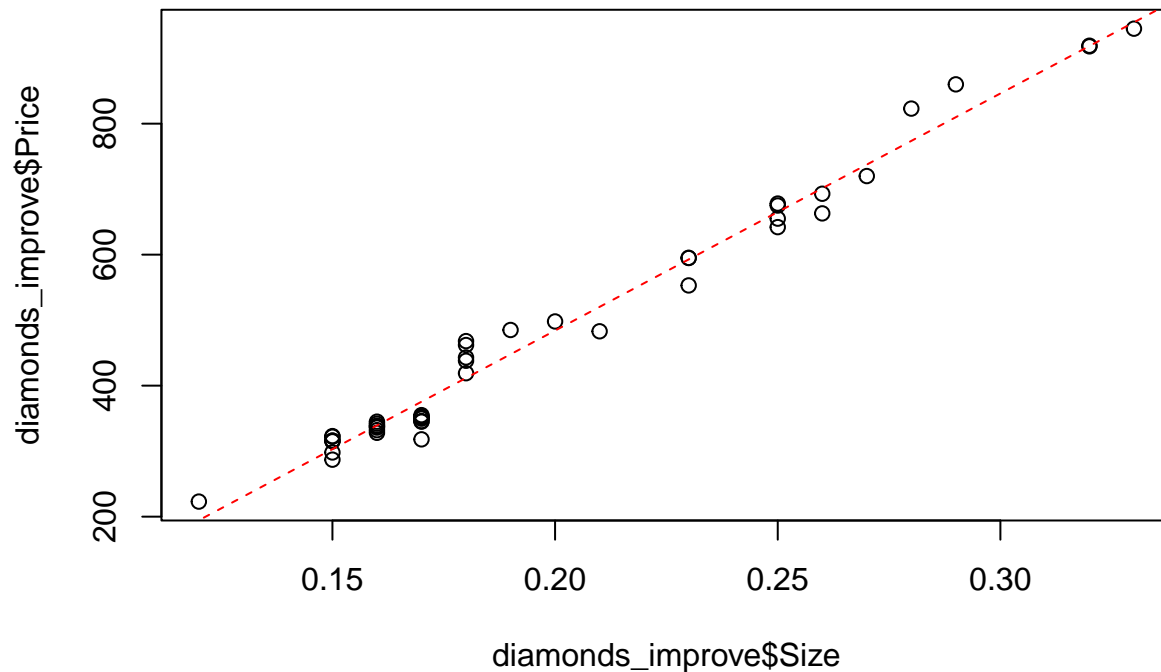
```
## 0.09196098 0.11715838 0.21815953
```

```
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
```

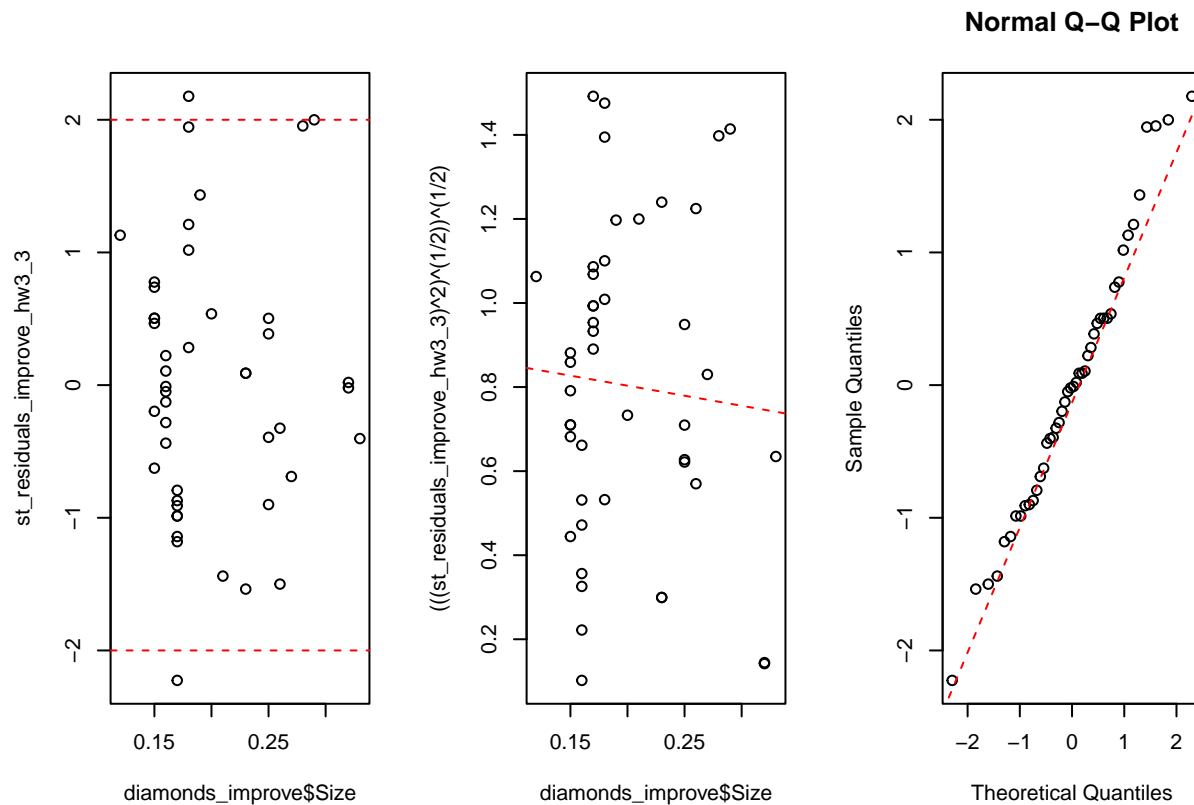


```
###
```

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

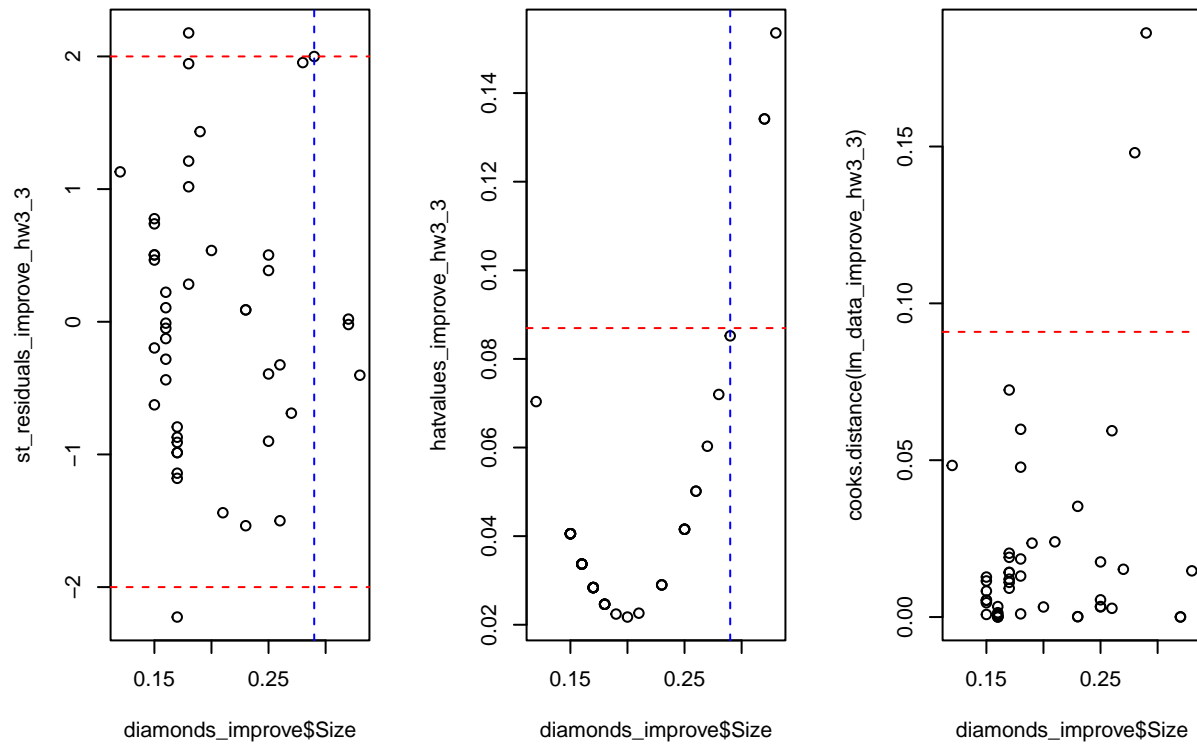


Then we can get this outcome. If we check the power of justification,

```
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)**

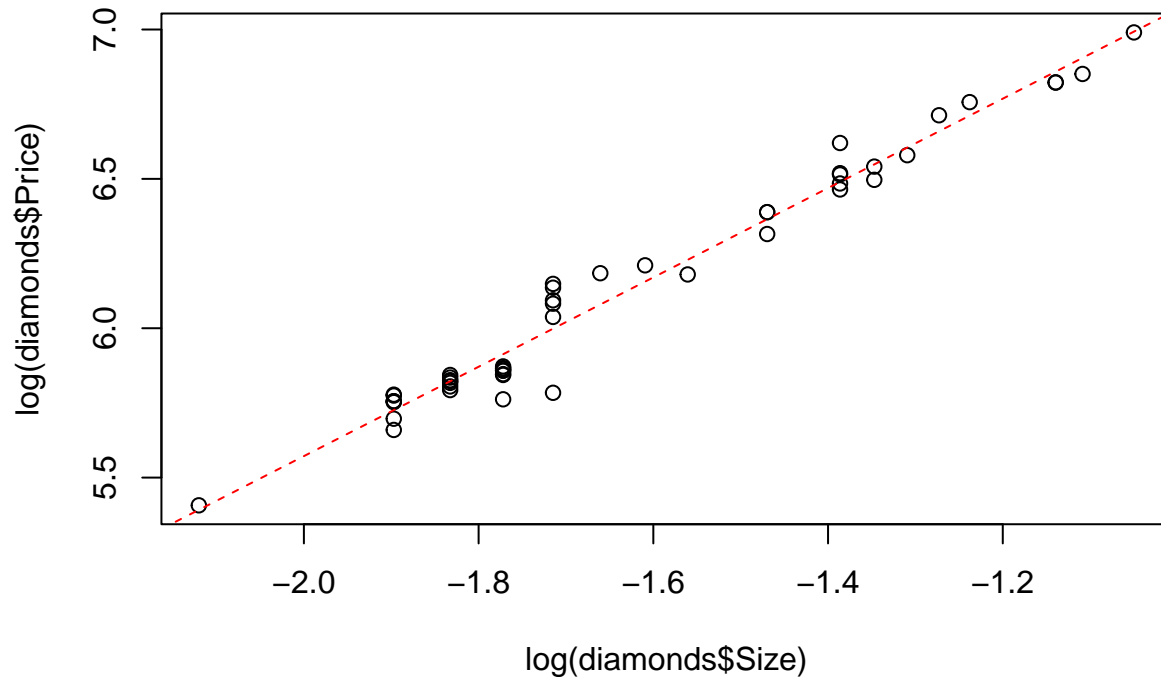
The number of data are small.

## 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(diamonds$Size) - 1))
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))

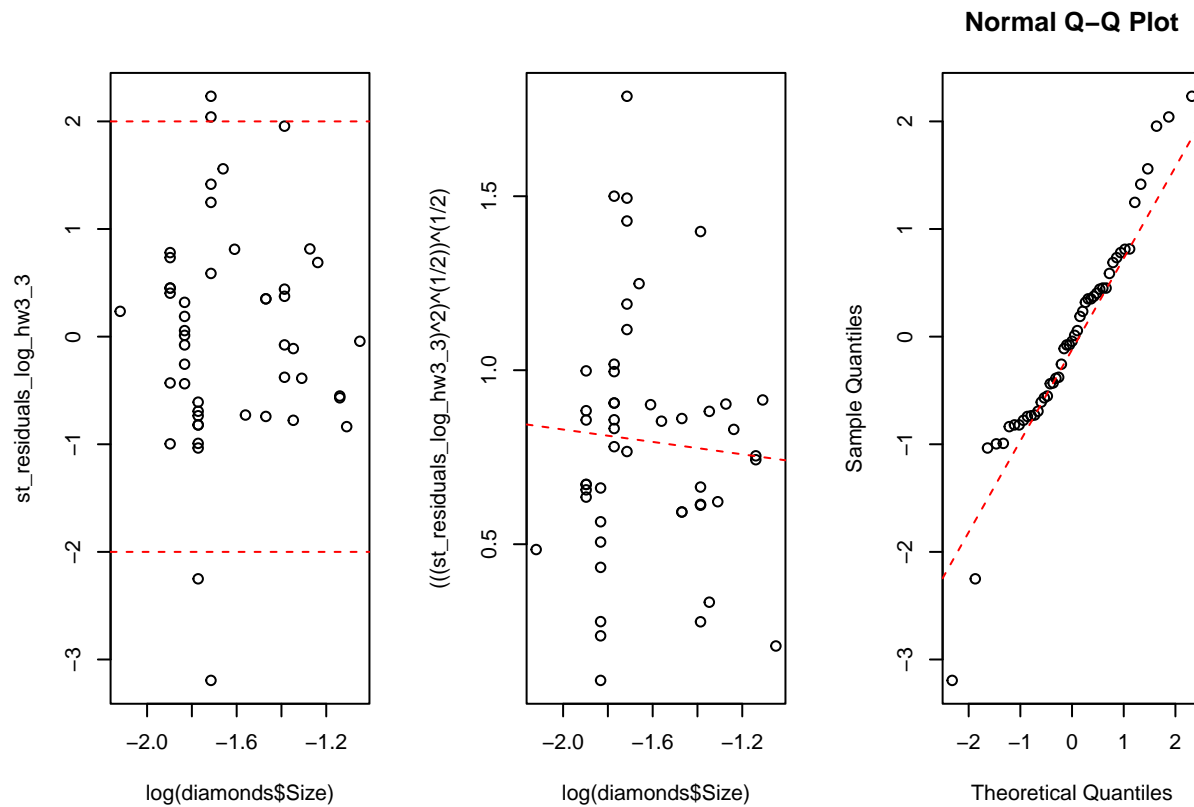
lm_data_resid_log_hw3_3 <- lm((((st_residuals_log_hw3_3)^2)^(1/2))^(1/2)~log(diamonds$Size), data=diamonds)

###

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$Size), (((st_residuals_log_hw3_3)^2)^(1/2))^(1/2))
abline(lm_data_resid_log_hw3_3$coefficients[1], lm_data_resid_log_hw3_3$coefficients[2], col='red', lty='dashed')

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

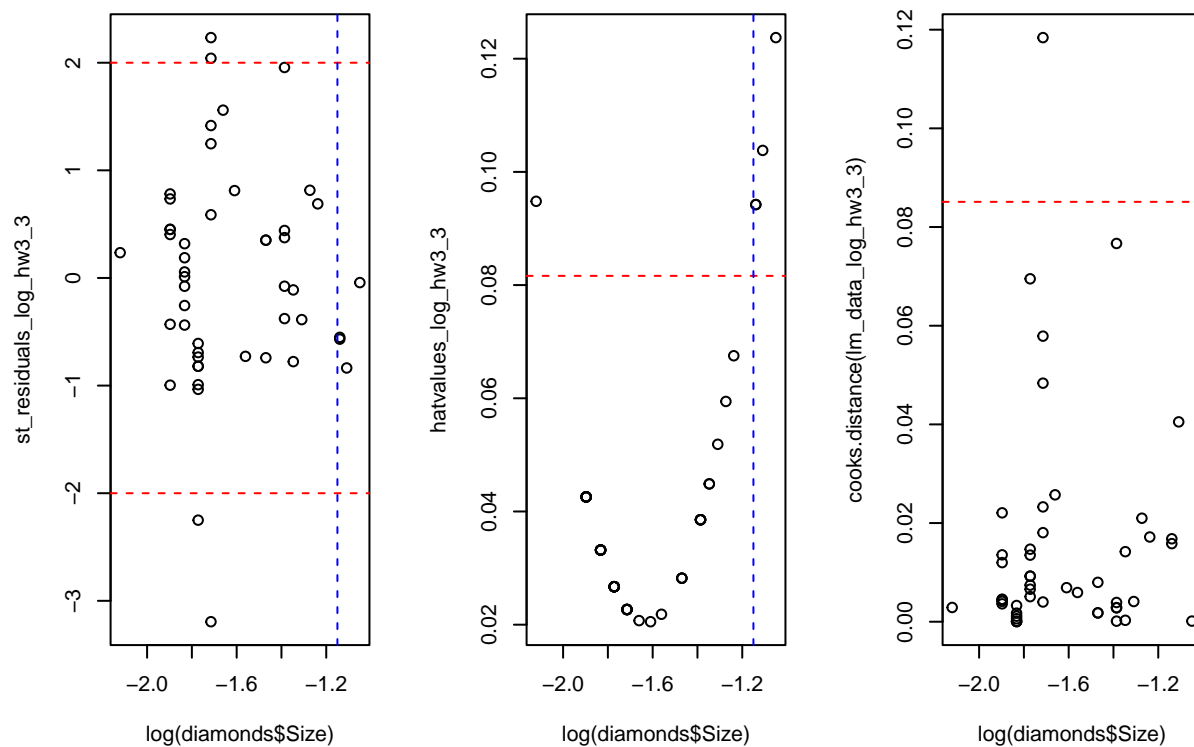


When we check the power of justification,

```
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
```

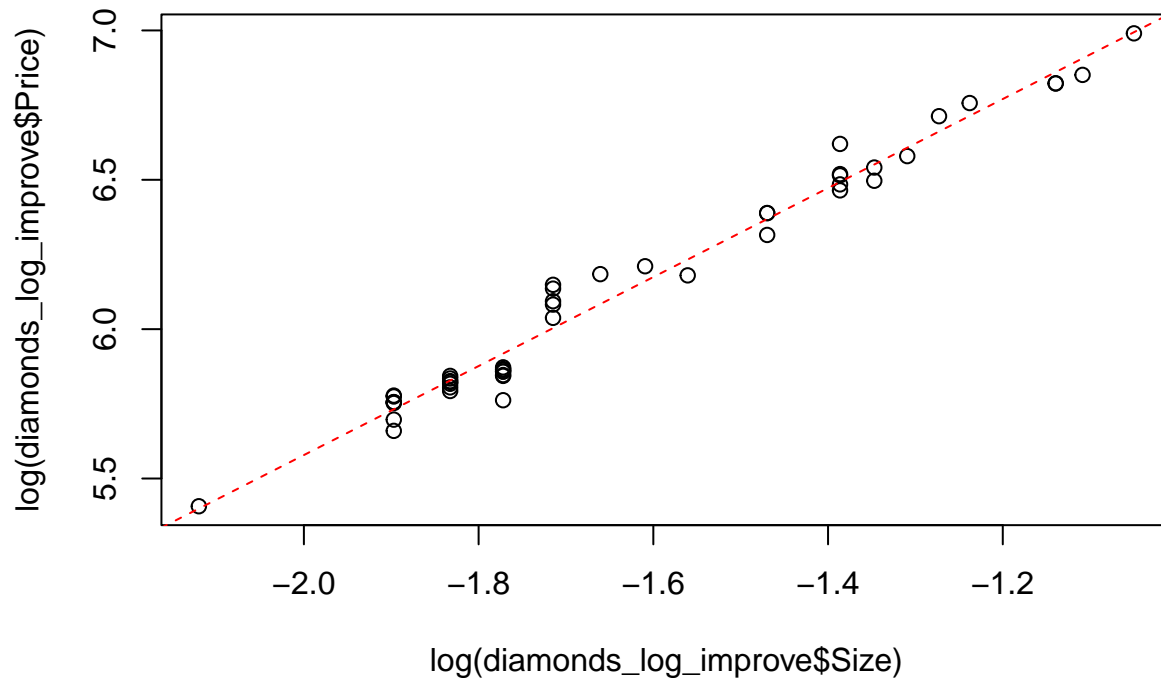
```
diamonds_log_improve <- diamonds[c(-4),]
```

```
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$residuals - mean(lm_data_log_improve_hw3_3$residuals))^2))^0.5
```

```
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-hatvalues_log_improve_hw3_3))
```

```
lm_data_residual_log_improve_hw3_3 <- lm((((st_residuals_log_improve_hw3_3^2)^(1/2))^(1/2)~log(diamonds_log_improve$Size)))
```

```
###
```

```
par(mfrow=c(1,3))
```

```
plot(log(diamonds_log_improve$Size), st_residuals_log_improve_hw3_3)
```

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

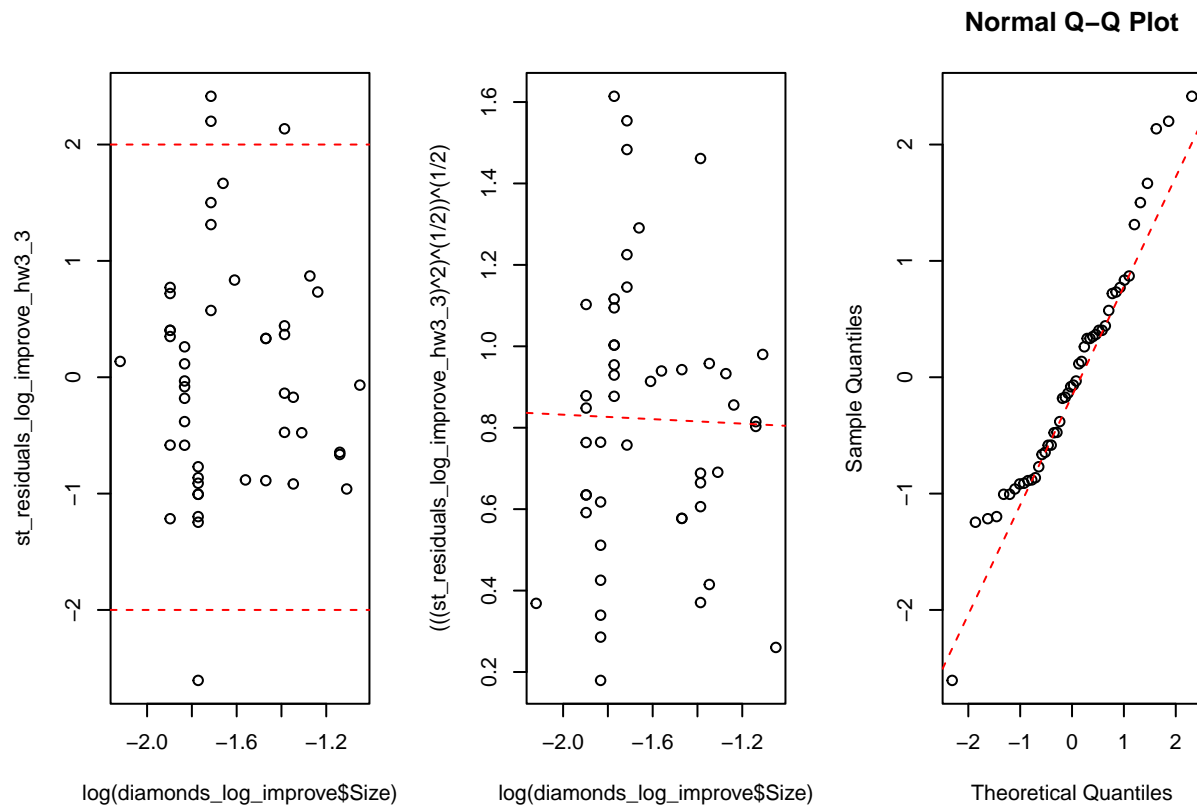
```
abline(-2,0,col='red', lty='dashed')
```

```
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$coefficients[1], lm_data_residual_log_improve_hw3_3$coefficients[2])
```

```
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.