Assignment2

Lien Dao

3/20/2022

Question 3.4.4

Tonnage data

```
library(datasets)
library(readr)
url1 <- "https://gattonweb.uky.edu/sheather/book/docs/datasets/glakes.txt"
tonnage <- read.table(url1, header = TRUE)
print(tonnage)
```

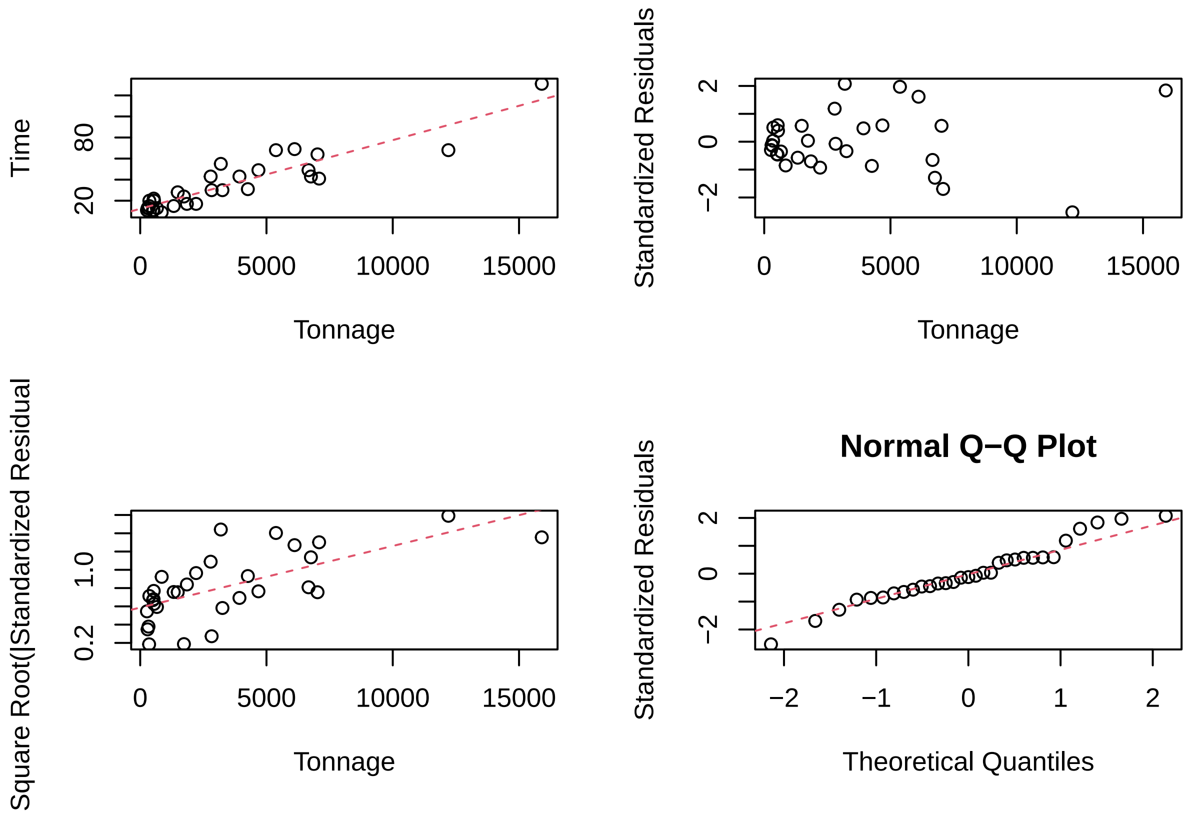
```
##    Case Tonnage Time
## 1     1    2213   17
## 2     2    3256   30
## 3     3   12203   68
## 4     4    7021   64
## 5     5     529   11
## 6     6    3192   55
## 7     7     547   20
## 8     8    4682   49
## 9     9    6112   69
## 10   10    5375   68
## 11   11    6666   49
## 12   12    3930   43
## 13   13    4263   31
## 14   14    1849   17
## 15   15     663   13
## 16   16     329   13
## 17   17    2790   43
## 18   18     353   15
## 19   19    2829   30
## 20   20     363   20
## 21   21    7084   41
## 22   22    1328   15
## 23   23     294   13
## 24   24     268   11
## 25   25    1732   24
## 26   26     507   11
## 27   27    1486   28
## 28   28     536   22
## 29   29     851    9
## 30   30    6760   43
## 31   31   15900  131
```

Simple linear regression model 1

```
tonnage_lm <- lm(Time ~ Tonnage, data = tonnage)
summary(tonnage_lm)
```

```
##
## Call:
## lm(formula = Time ~ Tonnage, data = tonnage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.882  -6.397  -1.261   5.931  21.850
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.344707   2.642633   4.671 6.32e-05 ***
## Tonnage      0.006518   0.000531  12.275 5.22e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.7 on 29 degrees of freedom
## Multiple R-squared:  0.8386, Adjusted R-squared:  0.833
## F-statistic: 150.7 on 1 and 29 DF,  p-value: 5.218e-13
```

```
par(mfrow=c(2,2))
plot(tonnage$Tonnage, tonnage$Time, xlab = "Tonnage", ylab = "Time")
abline(tonnage_lm, lty=2,col=2)
StanRes1 <- rstandard(tonnage_lm)
absrtsr1 <- sqrt(abs(StanRes1))
plot(tonnage$Tonnage, StanRes1, ylab="Standardized Residuals", xlab = "Tonnage")
plot(tonnage$Tonnage, absrtsr1, ylab="Square Root(|Standardized Residuals|)",
                                xlab = "Tonnage")
abline(lsfit(tonnage$Tonnage, absrtsr1), lty=2, col=2)
qqnorm(StanRes1, ylab="Standardized Residuals")
qqline(StanRes1, lty=2, col=2)
```

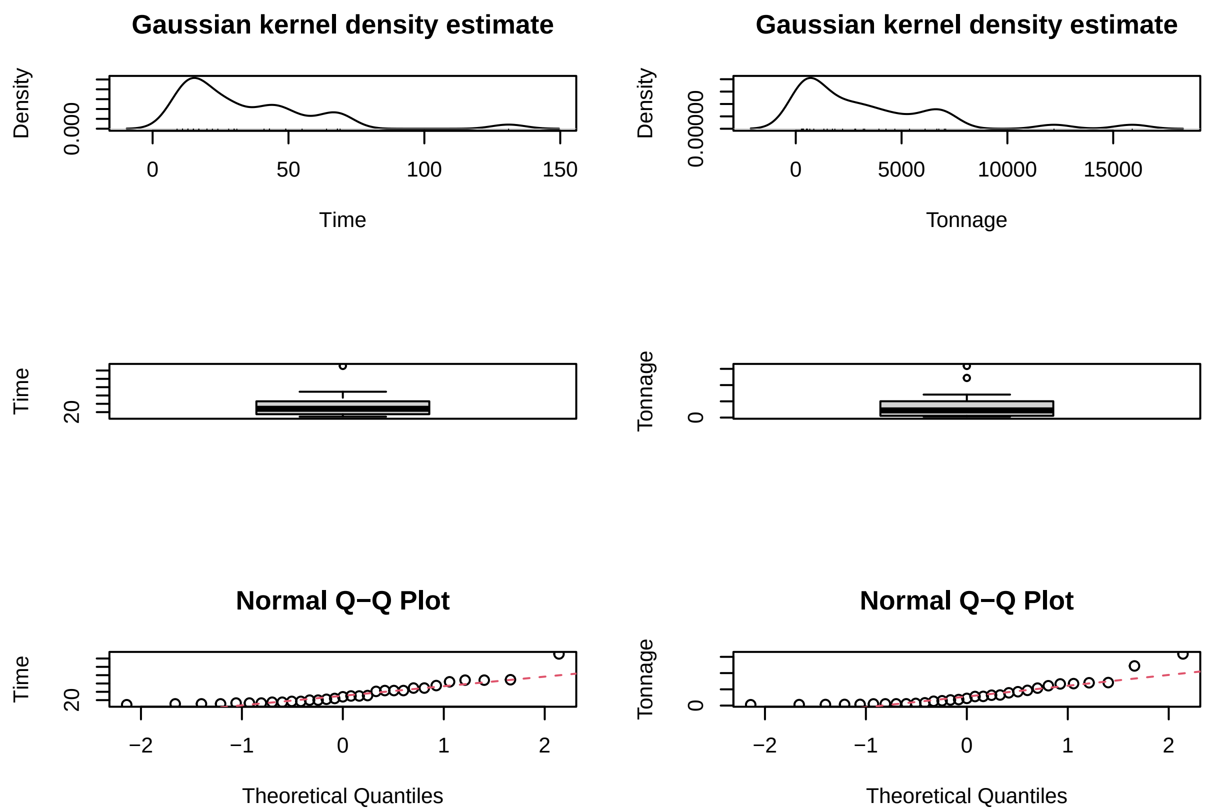


a1) The straight line regression model (3.8) seem to NOT fit the data well. The scatterplot shows that the time required increases as the volume of a ship's cargo increases, but the variance also increases. The variability in the standardized residuals tends to increase with the higher tonnages. We can also detect some outliers. There is a nonrandom pattern (is close to funnel shape and varies greatly to the right) evident in the plot of standardized residuals.

```
par(mfrow=c(3,2))
plot(density(tonnage$Time, bw="SJ", kern="gaussian"), type="l",
     main="Gaussian kernel density estimate", xlab="Time")
rug(tonnage$Time)
plot(density(tonnage$Tonnage, bw="SJ", kern="gaussian"), type="l",
     main="Gaussian kernel density estimate", xlab="Tonnage")
rug(tonnage$Tonnage)

boxplot(tonnage$Time ,ylab="Time")
boxplot(tonnage$Tonnage,ylab="Tonnage")

qqnorm(tonnage$Time, ylab = "Time")
qqline(tonnage$Time, lty = 2, col=2)
qqnorm(tonnage$Tonnage, ylab = "Tonnage")
qqline(tonnage$Tonnage, lty = 2, col=2)
```



b1) Calculate a prediction interval for Time when Tonnage = 10,000

```
pred_time <- predict(tonnage_lm, data.frame(Tonnage = 10000),
                     interval = "prediction")
pred_time
```

```
##        fit      lwr      upr
## 1 77.5234 54.17047 100.8763
```

The prediction interval for Tonnage = 10,000 is probably be too short, but might be valid.This is to be expected in this situation since on the original scale the data have variance which increases as the x-variable increases meaning that realistic prediction intervals will get wider as the x-variable increases.Generally, we expect the intervals for Time to be large for high tonnages and short for low tonnages.

Simple linear regression model 2

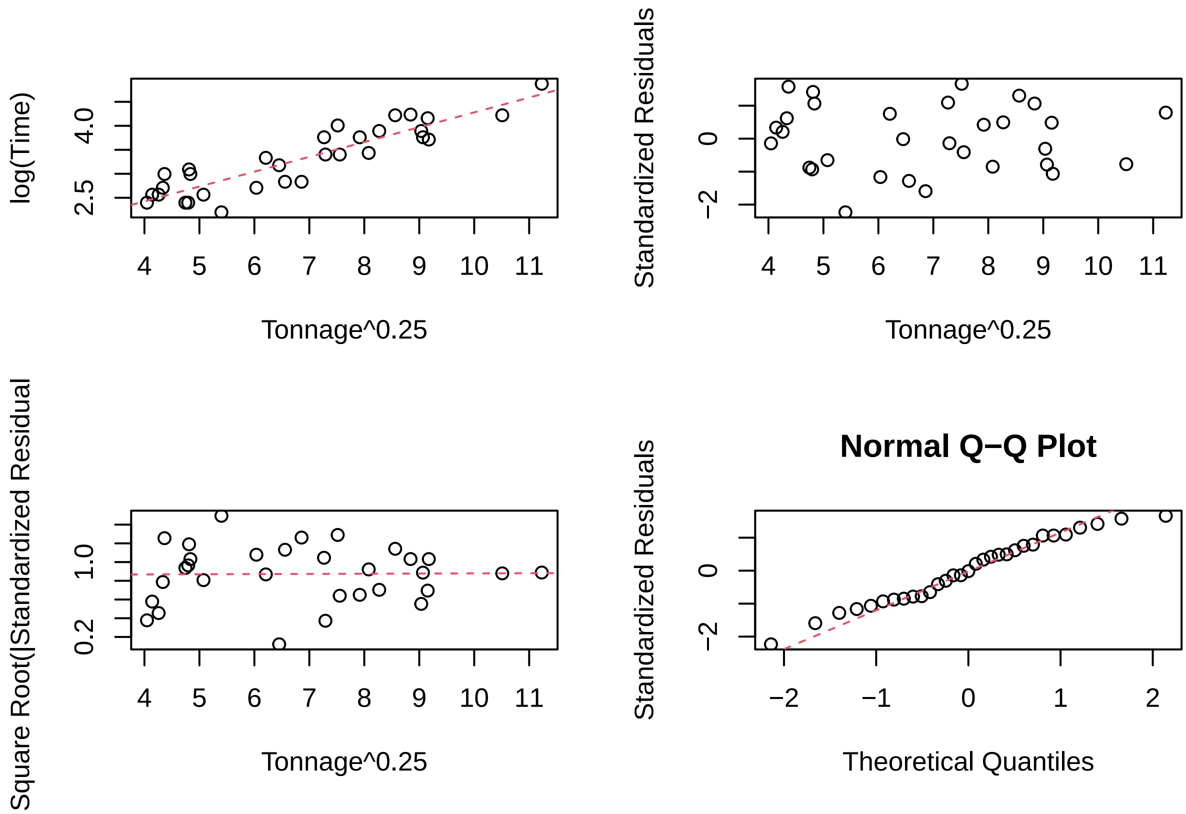
```
new_ton <- tonnage$Tonnage^0.25
new_time <- log(tonnage$Time)
tonnage_lm2 <- lm(new_time ~ new_ton, data = tonnage)
summary(tonnage_lm2)
```

```
##
## Call:
## lm(formula = new_time ~ new_ton, data = tonnage)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6607 -0.2410 -0.0044  0.2203  0.4956
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.18842    0.19468   6.105  1.2e-06 ***
## new_ton      0.30910    0.02728  11.332  3.6e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3034 on 29 degrees of freedom
## Multiple R-squared:  0.8158, Adjusted R-squared:  0.8094
## F-statistic: 128.4 on 1 and 29 DF,  p-value: 3.599e-12
```

a2) Even though model 2's R-squared is lower, the ability to predict time seems to improve. The residuals distribute more randomly. Prediction intervals for Time might align more with the volumes for model 2 because of the more random residuals and the values distribute more evenly along the regression line.

```r
par(mfrow=c(2,2))
plot(new_ton, new_time, xlab = "Tonnage^0.25", ylab = "log(Time)")
abline(tonnage_lm2, lty=2,col=2)
StanRes2 <- rstandard(tonnage_lm2)
absrtsr2 <- sqrt(abs(StanRes2))
plot(new_ton, StanRes2, ylab="Standardized Residuals", xlab = "Tonnage^0.25")
plot(new_ton, absrtsr2, ylab="Square Root(|Standardized Residuals|)",
                        xlab = "Tonnage^0.25")
abline(lsfit(new_ton, absrtsr2), lty=2, col=2)
qqnorm(StanRes2, ylab="Standardized Residuals")
qqline(StanRes2, lty=2, col=2)
```



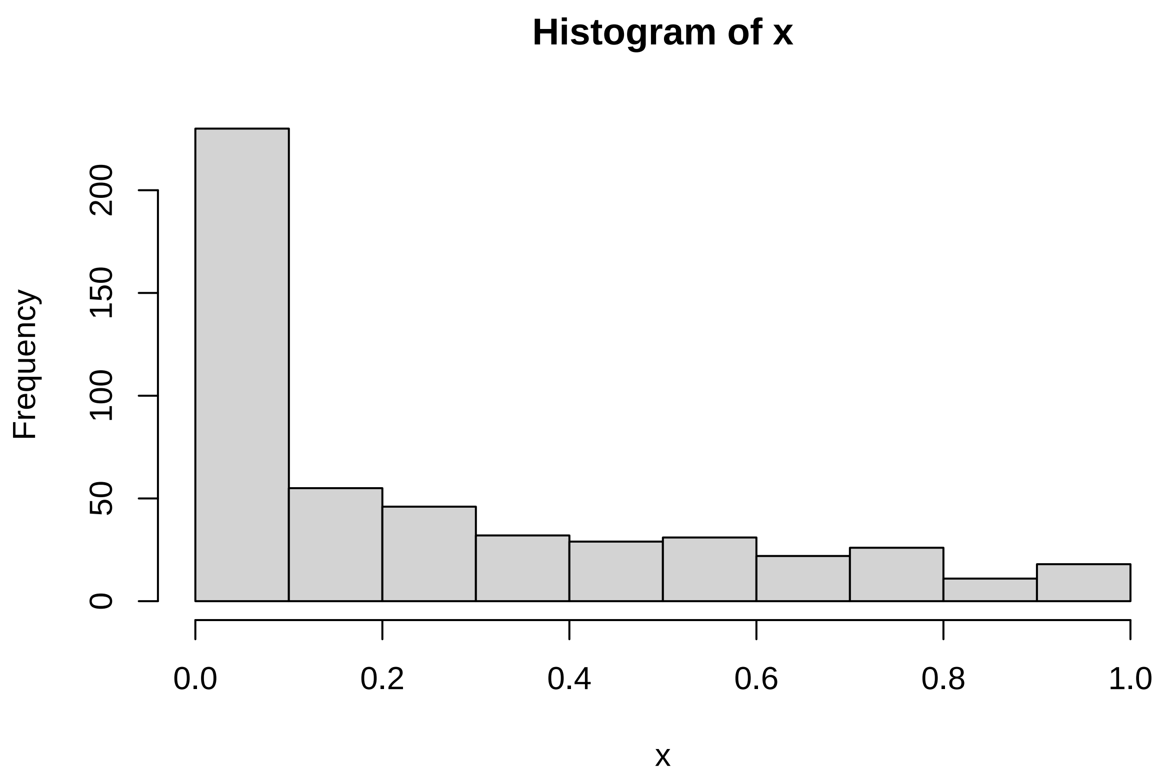
b2) Instead of having a funnel shape with smaller left tail, the distribution of the residuals seem to have a slight nonrandom pattern with a small right tail, which shows that the residuals' variability decreases for higher tonnages. There are also more values concentrating to the left of the distribution that is not fixed entirely after the transformation.

Question 3.4.6

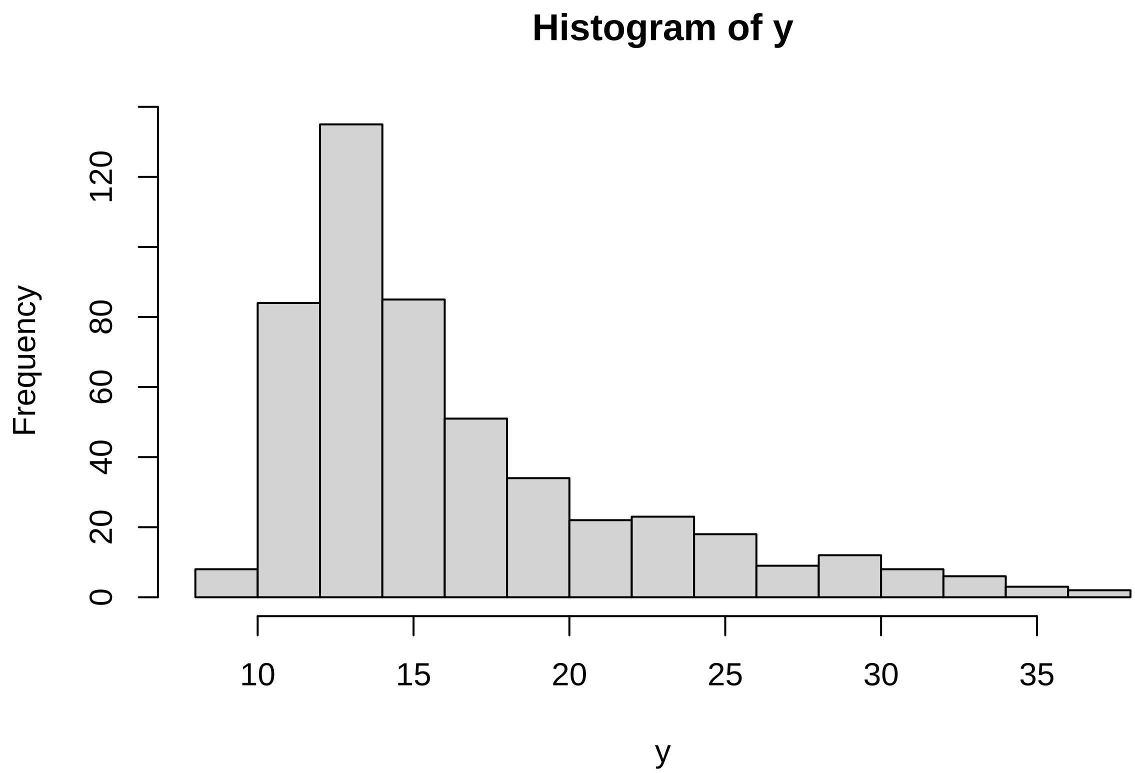
```
library(car)
```

```
## Loading required package: carData
```

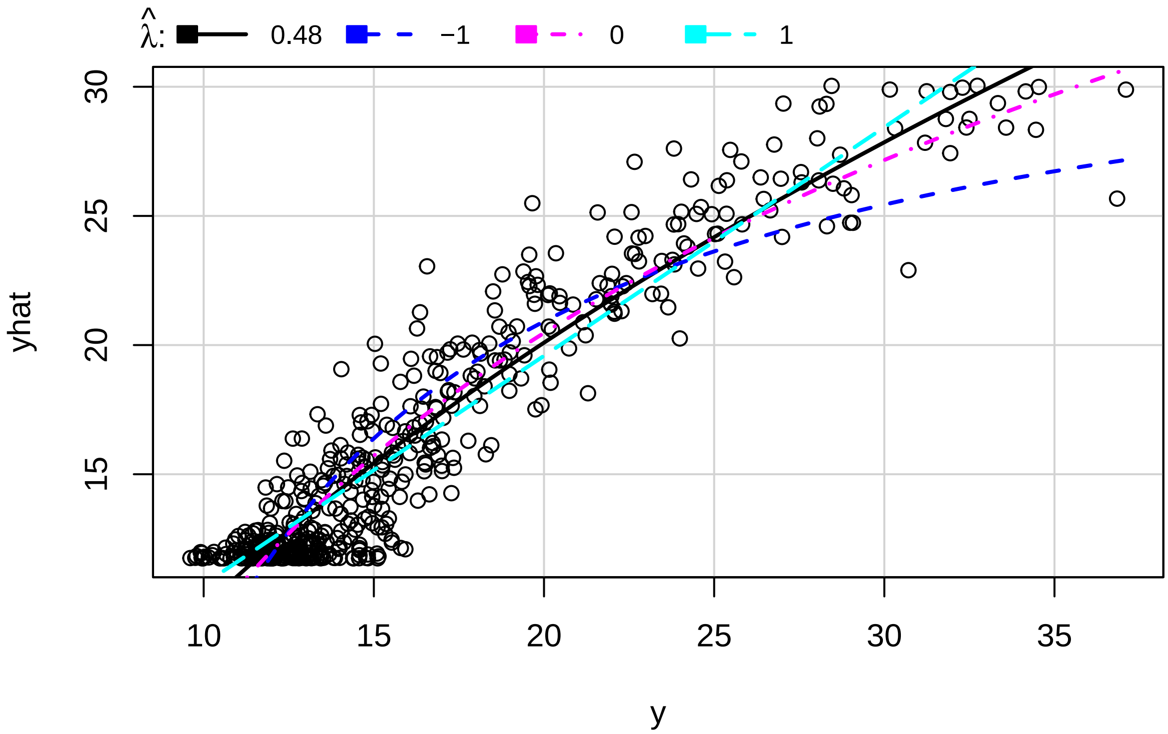
```
set.seed(3)
n <- 500
x <- runif(n,0,1)^3
e <- rnorm(n,0,0.1)
y <- exp(2.5 + 1*x + e)
hist(x)
```




```
hist(y)
```




```
sample_lm3 <- lm(y ~ x)
par(mfrow=c(1,1))
inverseResponsePlot(sample_lm3,key=TRUE)
```




```
##       lambda      RSS
## 1  0.475978 1460.463
## 2 -1.000000 2254.016
## 3  0.000000 1549.348
## 4  1.000000 1569.923
```

For the inverse plot of y and yhat to give an estimate of lambda that is close to the correct value of lamdba for this model, the distribution of Y needs to be skewed and the distribution of x needs to be symmetric or is approximately normally distributed. In this situation, x is assumed to be highly skewed, which leads to an incorrect lambda estimate.

Question 3.4.8

Diamond data

```
url2 <- "https://gattonweb.uky.edu/sheather/book/docs/datasets/diamonds.txt"
diamond <- read.table(url2, header = TRUE)
print(diamond)
```

```
##    Size Price
## 1  0.17   355
## 2  0.16   328
## 3  0.17   350
## 4  0.18   325
## 5  0.25   642
## 6  0.16   342
## 7  0.15   322
## 8  0.19   485
## 9  0.21   483
## 10 0.15   323
## 11 0.18   462
## 12 0.28   823
## 13 0.16   336
## 14 0.20   498
## 15 0.23   595
## 16 0.29   860
## 17 0.12   223
## 18 0.26   663
## 19 0.25   750
## 20 0.27   720
## 21 0.18   468
## 22 0.16   345
## 23 0.17   352
## 24 0.16   332
## 25 0.17   353
## 26 0.18   438
## 27 0.17   318
## 28 0.18   419
## 29 0.17   346
## 30 0.15   315
## 31 0.17   350
## 32 0.32   918
## 33 0.32   919
## 34 0.15   298
## 35 0.16   339
## 36 0.16   338
## 37 0.23   595
## 38 0.23   553
## 39 0.17   345
## 40 0.33   945
## 41 0.25   655
## 42 0.35  1086
## 43 0.18   443
## 44 0.25   678
## 45 0.25   675
## 46 0.15   287
```

```
## 47 0.26   693
## 48 0.15   316
## 49 0.15   316
```

Simple linear regression model 3

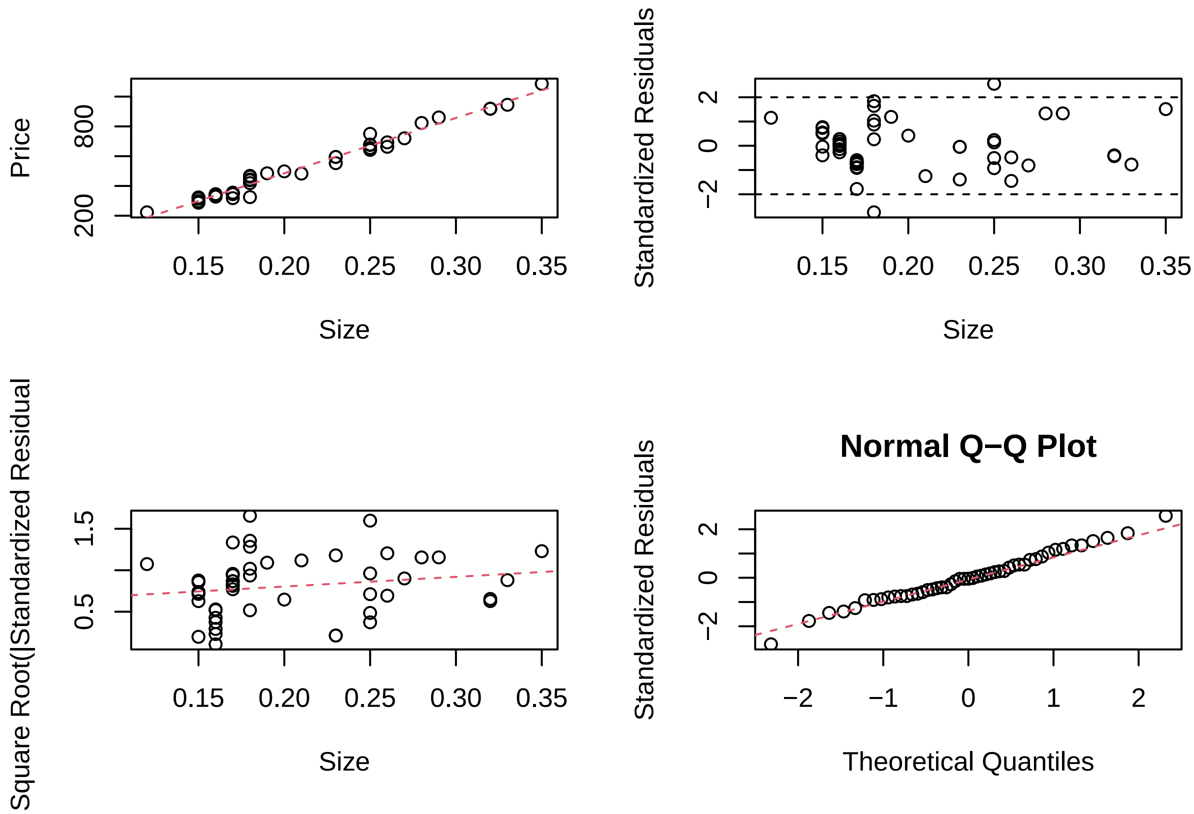
```
diamond_lm3 <- lm(Price ~ Size, data = diamond)
summary(diamond_lm3)
```

```
##
## Call:
## lm(formula = Price ~ Size, data = diamond)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.654 -21.503  -1.203  16.797  79.295
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -258.05      16.94  -15.23   <2e-16 ***
## Size         3715.02      80.41   46.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.6 on 47 degrees of freedom
## Multiple R-squared:  0.9785, Adjusted R-squared:  0.978
## F-statistic:  2135 on 1 and 47 DF,  p-value: < 2.2e-16
```

Part 1

a) The model is given by Price = -258.05 + 3715.02*Size We choose to build a simple linear regression model because there is clearly a linear relationship between Size and Price based on the scatter plot.

```
par(mfrow=c(2,2))
plot(diamond$Size, diamond$Price, xlab = "Size", ylab = "Price")
abline(diamond_lm3, lty=2,col=2)
StanRes3 <- rstandard(diamond_lm3)
absrtsr3 <- sqrt(abs(StanRes3))
plot(diamond$Size, StanRes3, ylab="Standardized Residuals", xlab = "Size")
abline(h=2,lty=2)
abline(h=-2,lty=2)
plot(diamond$Size, absrtsr3, ylab="Square Root(|Standardized Residuals|)",
                             xlab = "Size")
abline(lsfit(diamond$Size, absrtsr3), lty=2, col=2)
qqnorm(StanRes3, ylab="Standardized Residuals")
qqline(StanRes3, lty=2, col=2)
```

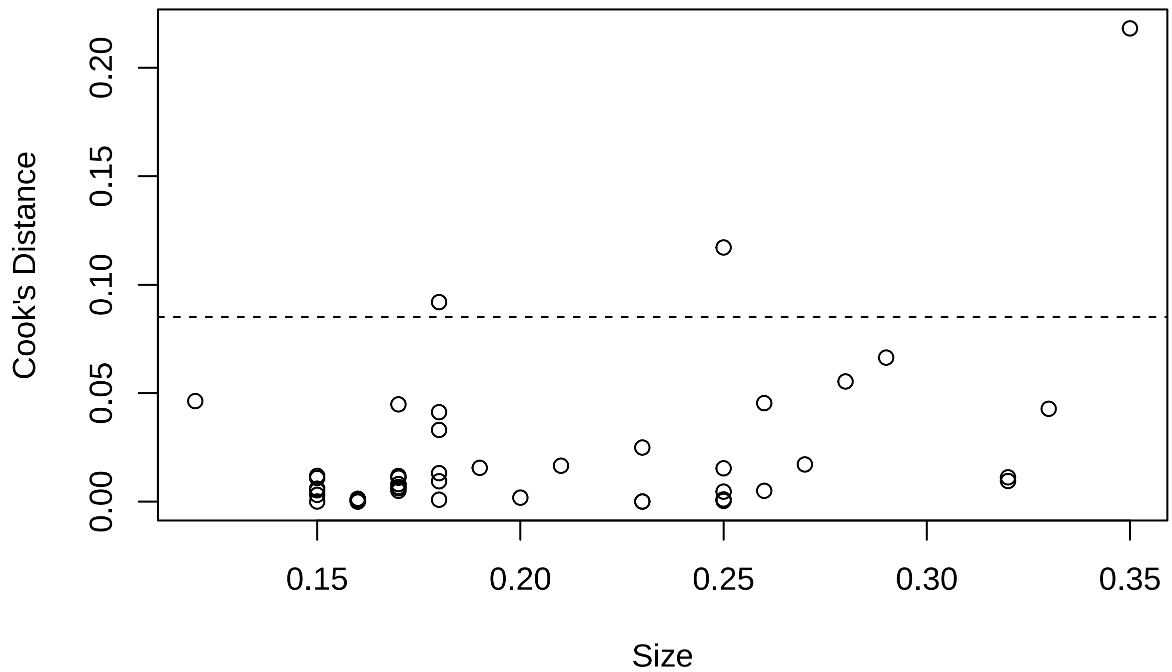


b) Overall, the model has strong statistics. The p-value of variable Size is small and significant, which indicates that there is a relationship between Size and Price. The model also has a high R-squared, which shows that the model fits the data well. However, if we look at the standardized residuals plot, there are two leverage points (outliers) because they lie outside of [-2,2]. The distribution of data points in the normal Q-Q plot is slightly not linear. Moreover, the residuals don't have a constant variance since they vary greatly. Therefore, we can examine the leverage points/outliers or apply transformations to improve the model and avoid non-constant variance.

Part 2

a) Cook's Distance to evaluate the leverage points

```
N <- 49
cd <- cooks.distance(diamond_lm3)
plot(diamond$Size, cd, xlab="Size", ylab="Cook's Distance")
abline(h=4/(N-2),lty=2)
```




```
lev_points <- ifelse(cd > 4/(N-2), diamond$Size, NA)
df_lev <- as.data.frame(lev_points)
subset(df_lev, df_lev$lev_points != "NA")
```

```
##    lev_points
## 4        0.18
## 19       0.25
## 42       0.35
```

There are 3 leverage points. The 0.18 and 0.25 carat diamonds' prices seem not too unrealistic from prices of other diamonds of the similar sizes. However, the 0.35 carat diamond is sold with a much higher price so it needs to be checked and confirmed. We build a linear model without the three leverage points identified earlier.

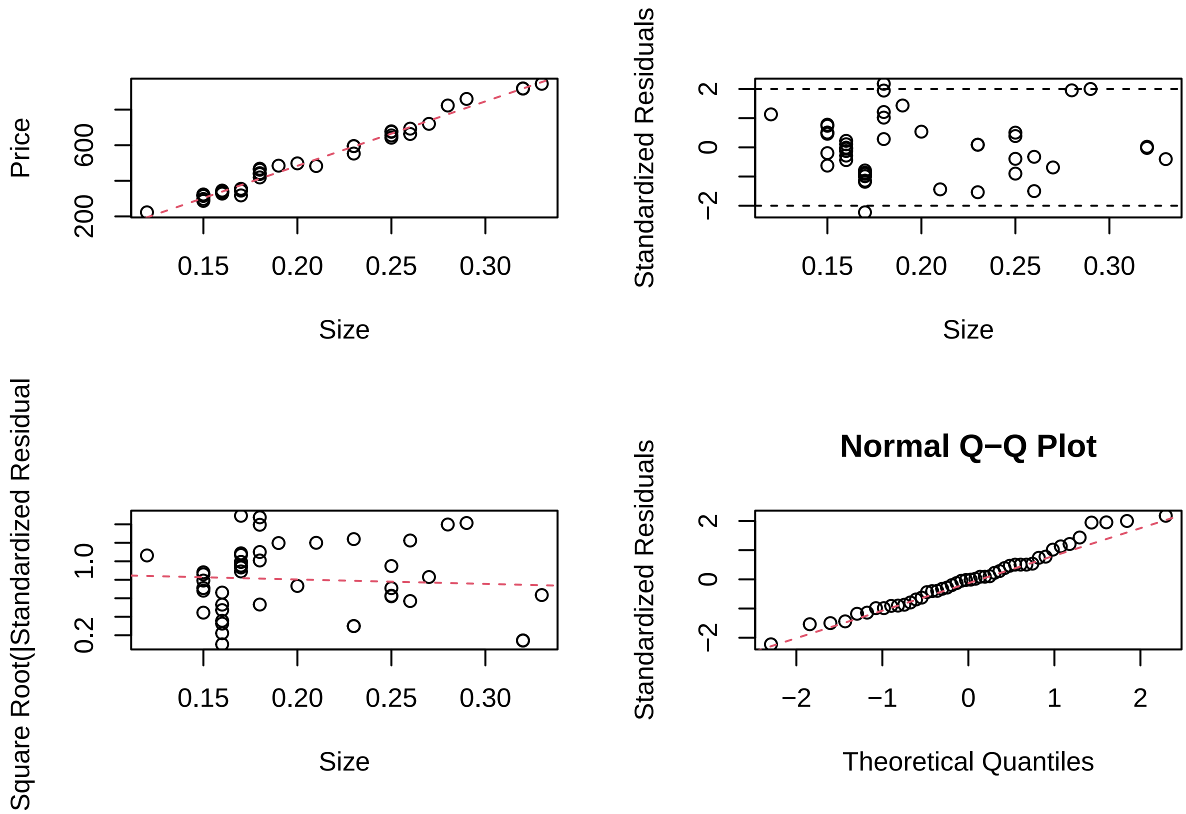
```
new_diamond <- diamond[c(-4,-19,-42),]
diamond_lm4 <- lm(Price ~ Size, data = new_diamond)
summary(diamond_lm4)
```

```
##
## Call:
## lm(formula = Price ~ Size, data = new_diamond)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.471 -19.727  -0.388  12.934  56.327
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -239.97      15.02  -15.98   <2e-16 ***
## Size         3620.25      72.80   49.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.2 on 44 degrees of freedom
## Multiple R-squared:  0.9825, Adjusted R-squared:  0.9821
## F-statistic:  2473 on 1 and 44 DF,  p-value: < 2.2e-16
```

The model has a much higher R-squared.

```
par(mfrow=c(2,2))
plot(new_diamond$Size, new_diamond$Price, xlab = "Size", ylab = "Price")
abline(diamond_lm4, lty=2,col=2)
StanRes4 <- rstandard(diamond_lm4)
absrtsr4 <- sqrt(abs(StanRes4))
plot(new_diamond$Size, StanRes4, ylab="Standardized Residuals", xlab = "Size")
abline(h=2,lty=2)
abline(h=-2,lty=2)
plot(new_diamond$Size, absrtsr4, ylab="Square Root(|Standardized Residuals|)",
                                 xlab = "Size")
abline(lsfit(new_diamond$Size, absrtsr4), lty=2, col=2)
qqnorm(StanRes4, ylab="Standardized Residuals")
qqline(StanRes4, lty=2, col=2)
```



b) Even though the value of R-squared improves, the model has new leverage points after omitting the previous leverage points. Also, the variance of residuals is not so constant since they cluster to the left of the distribution and vary greatly.

Part 3

After the modification, the model B has stronger statistics than model A. For model B, if we look at the standardized residuals plot, the new leverage points are closer to [-2,2]. The distribution of data points in the normal Q-Q plot is really close to being diagonal. Moreover, the residuals' variance improve and appear to be more stable than model A's.