Lab 4 in R: Analysis of Forestry.csv

Erin Canada

October 25, 2018

Forestry.CSV Setting up variables:

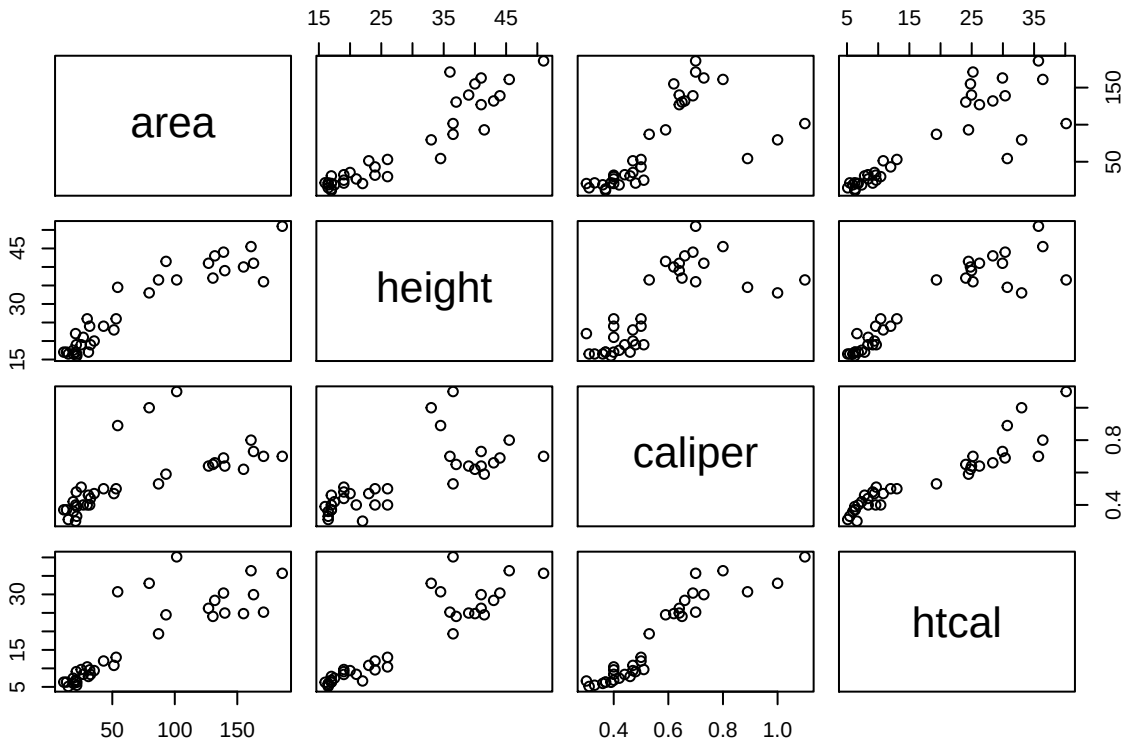
This dataset records the needle areaand several other variables for 35 coniferous trees.

Specific Variables: .area: the total needle area of the tree .height: the height of the tree .caliper: a measure of the tree's trunk size .htcal: the product of heightand caliper(a measure of the tree's trunk volume)

```
getwd()
```

```
## [1] "C:/Users/Erin Canada/Documents/USF/Fall 2018/Regression/Lab"
```

```
setwd("C:/Users/Erin Canada/Documents/USF/Fall 2018/Regression/Lab")

forest <- read.csv("forestry.csv", header = T)

pairs(forest)
```




```
cor(forest)
```

```
##              area    height   caliper     htcal
## area    1.0000000 0.9359487 0.6824739 0.8659820
```

```
## height  0.9359487 1.0000000 0.7362026 0.9252236
## caliper 0.6824739 0.7362026 1.0000000 0.9304803
## htcal   0.8659820 0.9252236 0.9304803 1.0000000
```

```
#Setting Variables
y <- forest$area
x1 <- forest$height
x2 <- forest$caliper
x3 <- forest$htcal
```

Part A

(a)Fit a multiple linear regression model relating area to the three explanatory variables listed above and construct the following residual plots:

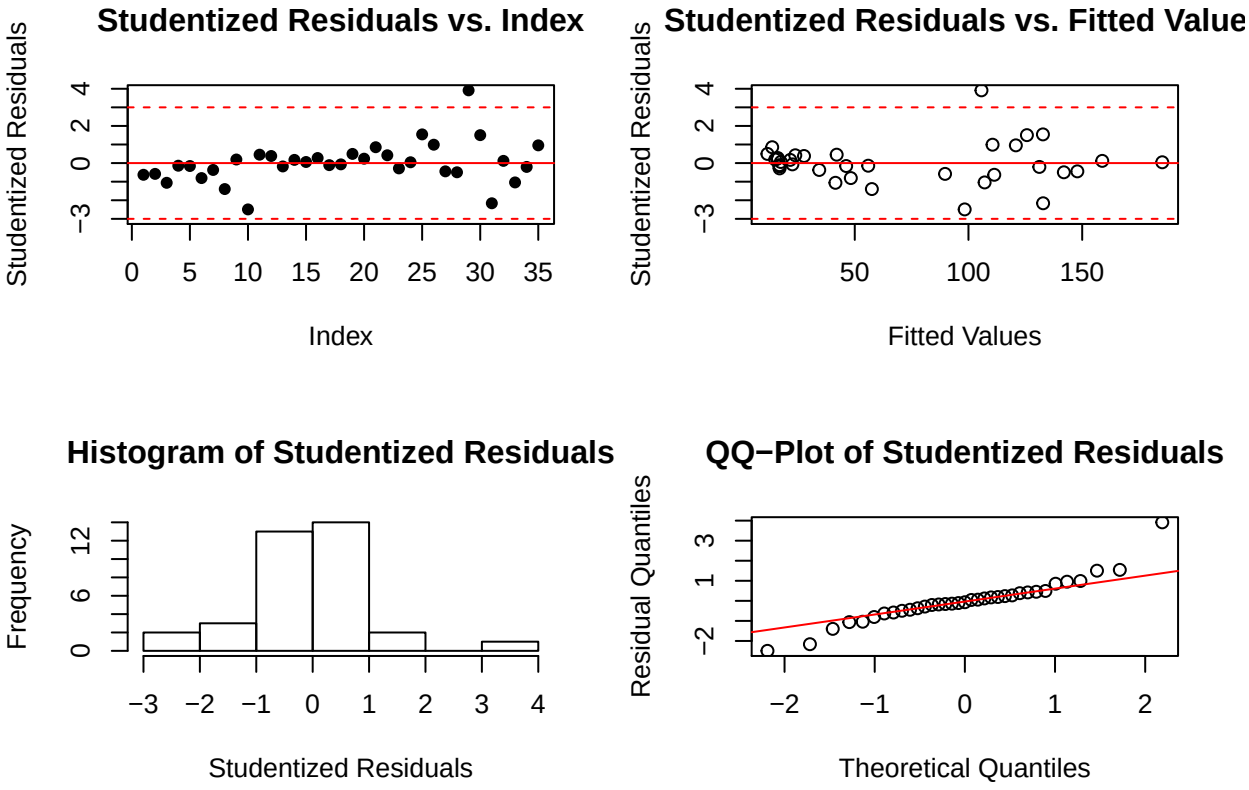
i.Studentized Residuals vs. Index ii.Studentized Residuals vs. Fitted Values iii.Histogram of Studentized Residuals iv.QQ-plot of Studentized Residuals

```
#Creating MLR model relating area to three explanatory variables

model <- lm(formula = y~x1+x2+x3, data = forest)

#Summary of model
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = forest)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.034  -9.353  -1.396   8.219  65.281
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -33.177     55.096  -0.602   0.5514
## x1             3.735      1.952   1.914   0.0649 .
## x2           -74.117    114.012  -0.650   0.5204
## x3             2.234      3.539   0.631   0.5324
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.81 on 31 degrees of freedom
## Multiple R-squared:  0.8777, Adjusted R-squared:  0.8658
## F-statistic: 74.14 on 3 and 31 DF,  p-value: 3.078e-14
```

```
library(MASS)

par(mfrow = c(2,2))

#Studentized Residuals vs. Index
plot(studres(model), pch = 16, xlab = "Index",ylim = c(min(-3, min(studres(model))),max(3, max(studres(
abline(h=c(0,3,-3), lty = c(1,2,2), col = "red")

#Studentized Residuals vs Fitted Values
plot(model$fitted.values, studres(model), ylim = c(min(-3, min(studres(model))),max(3, max(studres(model
```

```r
abline(h=c(0,3,-3), lty = c(1,2,2), col = "red")

#Histogram of Studentized Residuals
hist(studres(model), xlab = "Studentized Residuals", main = "Histogram of Studentized Residuals")

#QQ-Plot of Studentized Residuals
qqnorm(studres(model), main = "QQ-Plot of Studentized Residuals", ylab = "Residual Quantiles")
qqline(studres(model), col = "red")
```



Part B

Based on the plots in (a), answer "Yes" or "No" to the following questions and give a one sentence justification.

i.Do the residuals appear to be independent?

In examining the residuals vs index plot, we are seeing some patterns at index between 5-10 and 30-35, suggesting that there may not be independece.

ii.Do the residuals appear to have constant variance?

In examining the residuals vs fitted values plot, we see no visable patterns or funnels between residuals and fitted values, indicating that we have a constant variance.

iii.Do the residuals appear to be normally distributed?

By looking at the historgram and qq-plot, we can see that it seems to have heavy tails in the qq plot and not a symmetric bell curve on the histogram, indicating that the residuals are not normally distributed.

iv.Do the residuals suggest the existence of an outlier?

Based on the plotted residuals, it does appear that our model may have an outlier where the studentized residuals are greater than 3. From the residuals vs index and residuals vs fitted values, it seems to be an observation in the y dimension.

Part c

Which observation has the largest Studentized residual?

```
which(studres(model)>3)
```

```
## 29
## 29
```

At Studentized Residuals where greater than 3, observation 29 is an outlier.

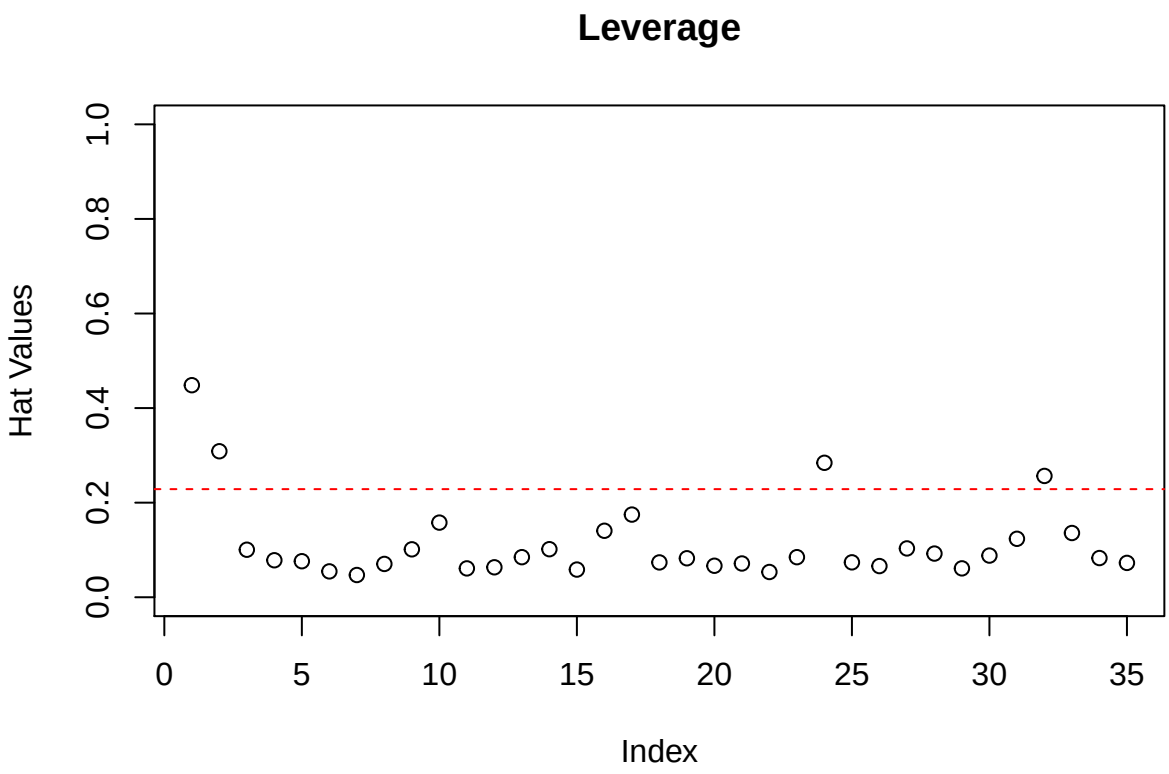
Part D

Calculate the leverage for each observation and construct a plot of them vs. their index. Which observations have 'high'leverage(i.e., leverage larger than twice the average leverage)?

```
#Calculate leverage for each observation (i.e diagonal values of hat matrix)
hatvalues(model)
```

```
##          1          2          3          4          5          6
## 0.44838774 0.30870904 0.10045117 0.07821682 0.07614714 0.05467466
##          7          8          9         10         11         12
## 0.04703267 0.07039175 0.10129502 0.15782718 0.06101001 0.06332837
##         13         14         15         16         17         18
## 0.08475604 0.10160375 0.05851252 0.14058356 0.17481524 0.07363143
##         19         20         21         22         23         24
## 0.08240691 0.06658612 0.07138849 0.05336468 0.08475604 0.28433074
##         25         26         27         28         29         30
## 0.07381084 0.06590437 0.10324293 0.09233541 0.06103058 0.08810671
##         31         32         33         34         35
## 0.12357671 0.25651243 0.13581279 0.08286254 0.07259760
```

```
#Plot leverage values vs index
plot(hatvalues(model), main = "Leverage", ylab = "Hat Values", ylim = c(0,1))
abline(h = 2*mean(hatvalues(model)), col = "red", lty = 2)
```




```
#High leverage
which(hatvalues(model)==max(hatvalues(model)))
```

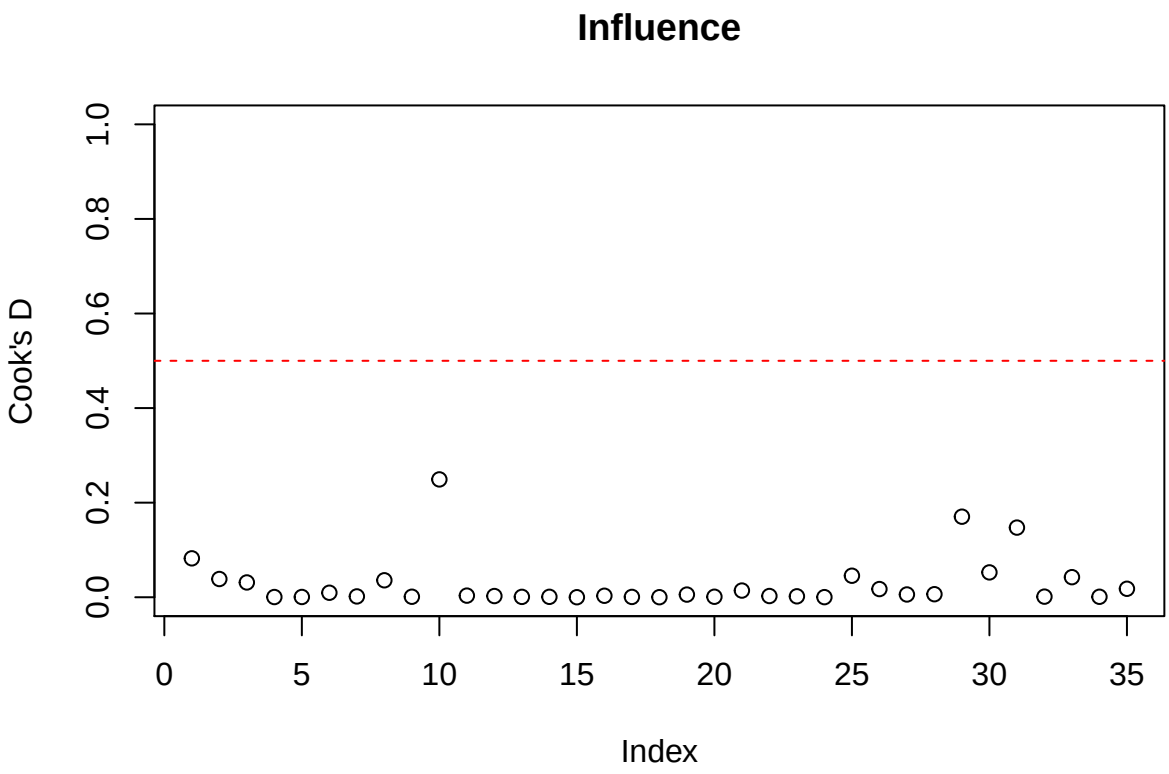
```
## 1
## 1
```

Observation 1 has 'high leverage'

Part E

Calculate Cook's D-statistic for each observation and construct a plot of them vs. their index. List the top three most influential points.

```
#Calculate Cook's D-statistic
cooks <- cooks.distance(model)

#Plot statistics vs indes
plot(cooks.distance(model), main = "Influence", ylab = "Cook's D", ylim = c(0,1))
abline(h = 0.5, col = "red", lty = 2)
```




```
#Top 3 Most influential points
influential <- order(cooks.distance(model), decreasing=TRUE)[1:3]

influential
```

```
## [1] 10 29 31
```

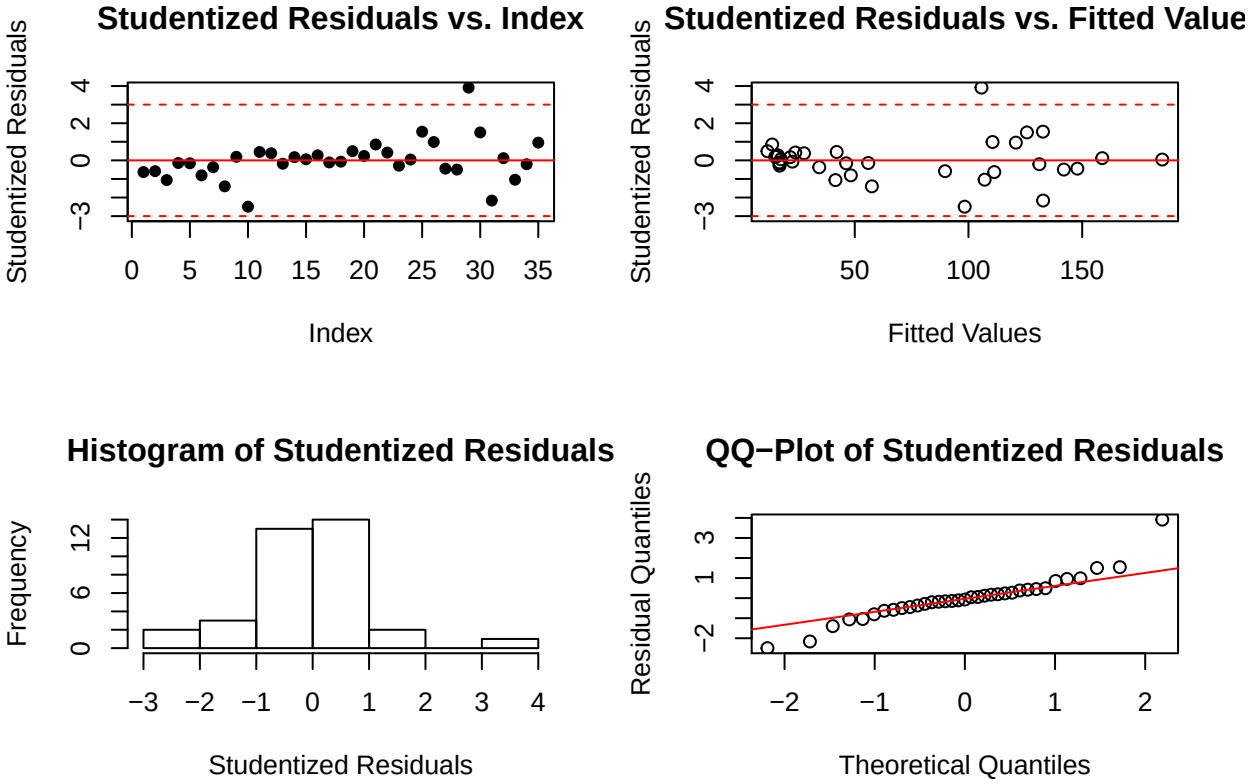
The top 3 most influential points for the cooks are the 10th, 29th and 31st observation.

Part F

Repeat part (a) but with observations 10 and 29 deleted.

```
#Remove observation 10, 29


forest_new <- forest[c(-10,-29),]
model2 <- lm(formula = y~ x1+x2+x3, data = forest_new)

par(mfrow = c(2,2))

#Studentized Residuals vs. Index
plot(studres(model2), pch = 16, xlab = "Index",ylim = c(min(-3,
min(studres(model2))),max(3, max(studres(model2)))),
ylab = "Studentized Residuals", main = "Studentized Residuals vs. Index")
abline(h=c(0,3,-3), lty = c(1,2,2), col = "red")

#Studentized Residuals vs Fitted Values
```

```r
plot(model2$fitted.values, studres(model2), ylim = c(min(-3,
min(studres(model2))),max(3, max(studres(model2)))),
main = "Studentized Residuals vs. Fitted Values",
ylab = "Studentized Residuals", xlab = "Fitted Values")
abline(h=c(0,3,-3), lty = c(1,2,2), col = "red")

#Histogram of Studentized Residuals
hist(studres(model2), xlab = "Studentized Residuals",
     main = "Histogram of Studentized Residuals")

#QQ-Plot of Studentized Residuals
qqnorm(studres(model2), main = "QQ-Plot of Studentized Residuals",
       ylab = "Residual Quantiles")
qqline(studres(model2), col = "red")
```



Part G

Consider the plot of the Studentized Residuals vs. Fitted Values from (f). Do these residuals appear to have constant variance? Answer "Yes" or "No" with a one sentence justification.

No, these residuals do not appear to have constant variance because of the funnel shape starting to happen as the fitted values increase.

Part H

Fit a multiple linear regression model relating log(area) (i.e., the natural logarithm of area) to the three explanatory variables, excluding observations 10 and 29 as you did in (f).

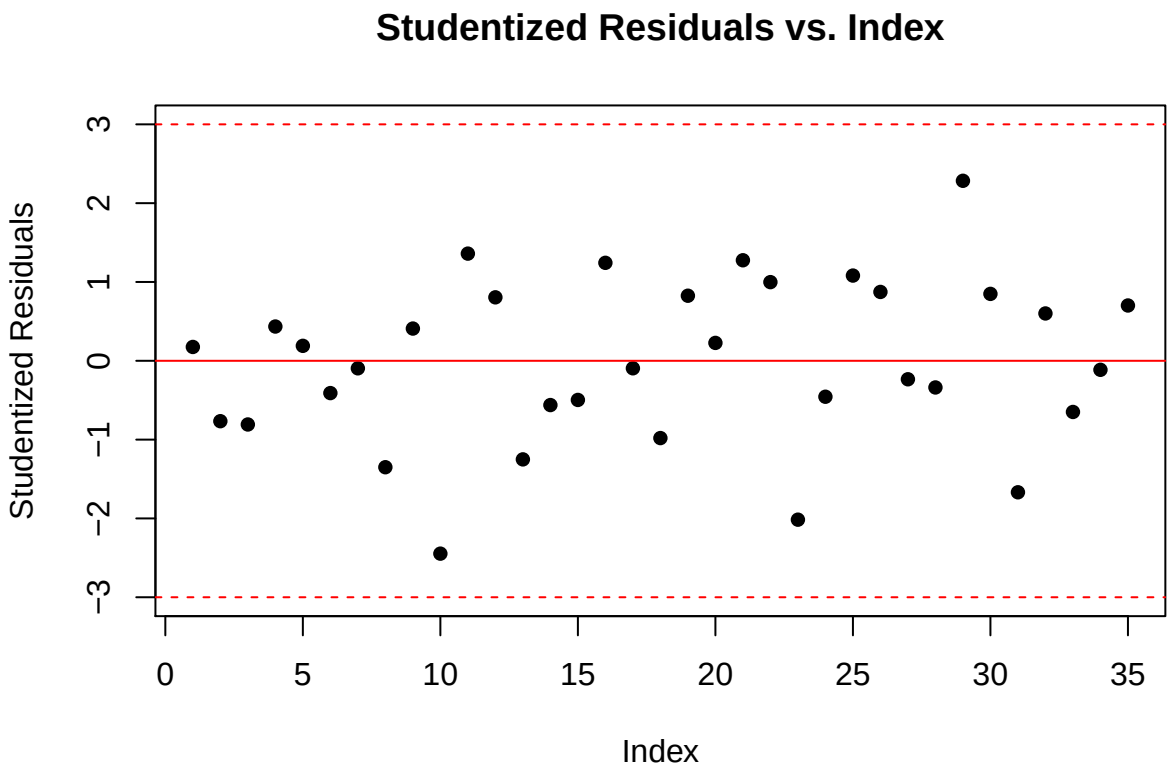
```
#Fitting MLR
log.model <- lm(formula = log(y)~x1+x2+x3, data = forest_new)
summary(log.model)
```

```
##
## Call:
## lm(formula = log(y) ~ x1 + x2 + x3, data = forest_new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55344 -0.15236 -0.02333  0.20962  0.55137
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.21239    0.70315  -0.302  0.76463
## x1           0.13369    0.02491   5.367  7.5e-06 ***
## x2           4.32655    1.45506   2.973  0.00566 **
## x3          -0.12059    0.04516  -2.670  0.01196 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2655 on 31 degrees of freedom
## Multiple R-squared:  0.9176, Adjusted R-squared:  0.9096
## F-statistic: 115.1 on 3 and 31 DF,  p-value: < 2.2e-16
```

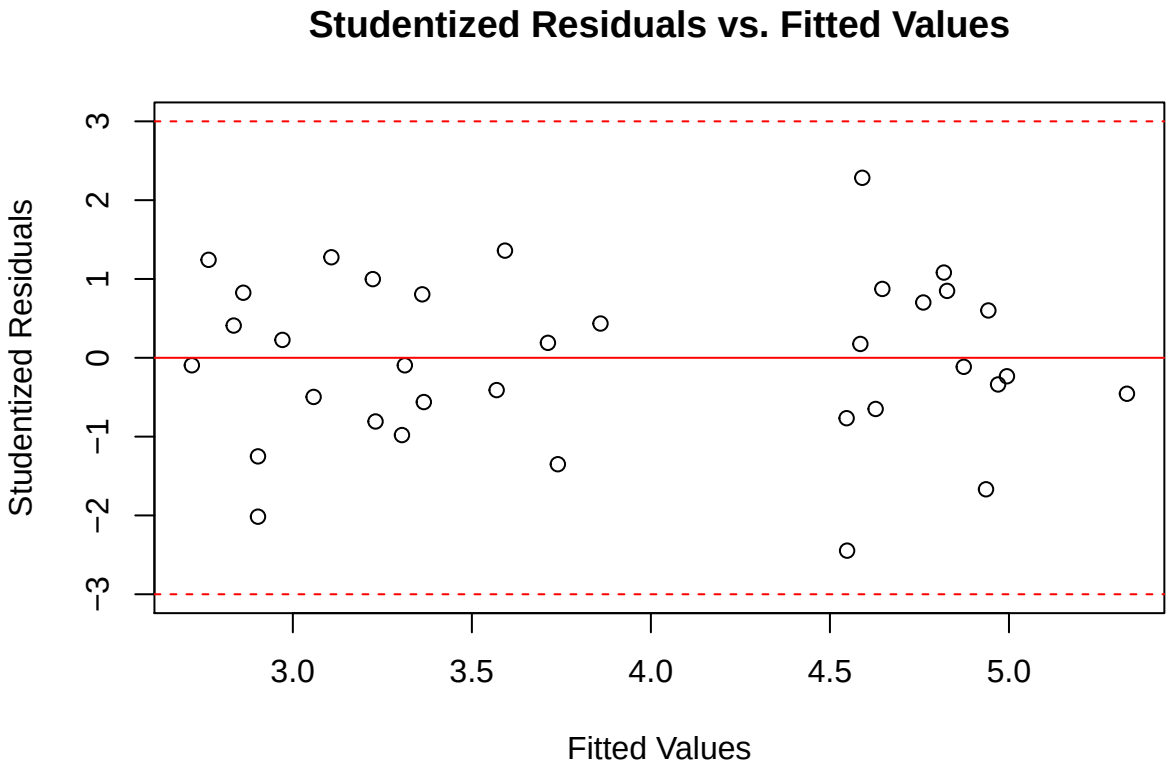
Part I

Construct a plot of the Studentized Residuals vs. Fitted Values for the model in (h). Does it appear as though the log-transformation has stabilized the variability of the residuals relative to what was observed in (g)? Which model - the one from (g) or the one from (h) - do you feel is most appropriate?

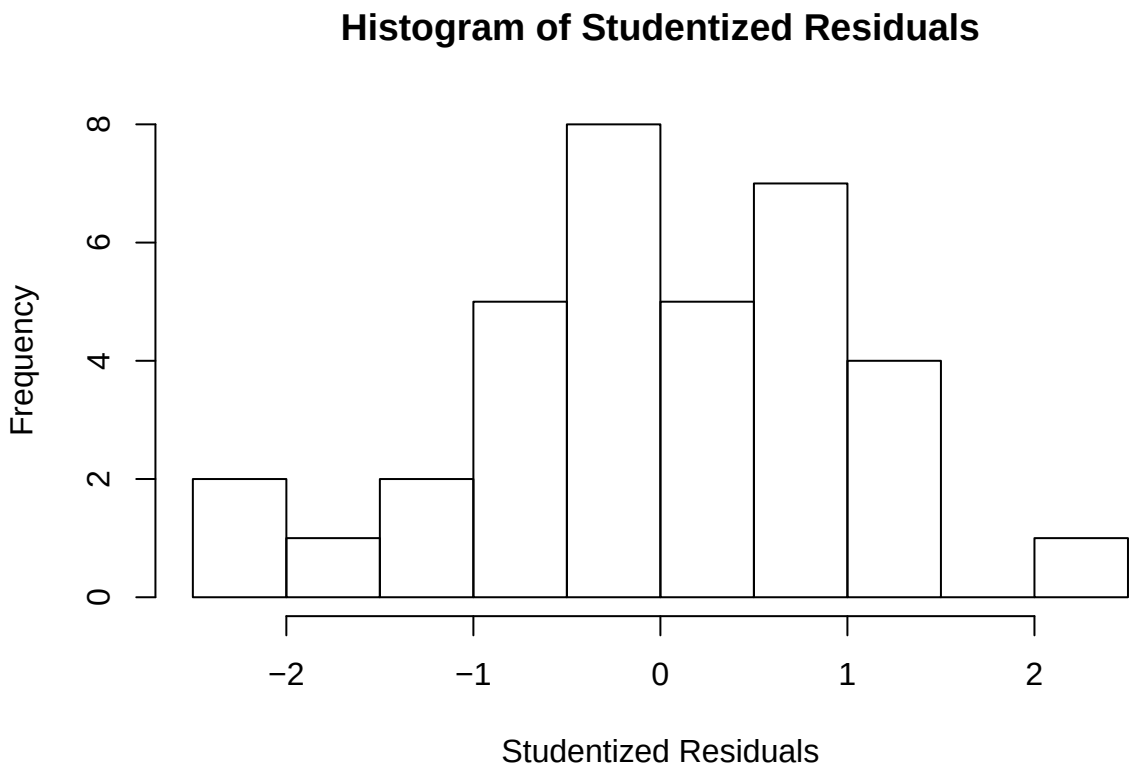
```
#Studentized Residuals vs. Index (Area)
plot(studres(log.model), pch = 16, xlab = "Index",ylim = c(min(-3,
      min(studres(log.model))),max(3, max(studres(log.model)))),
     ylab = "Studentized Residuals", main = "Studentized Residuals vs. Index")
abline(h=c(0,3,-3), lty = c(1,2,2), col = "red")
```



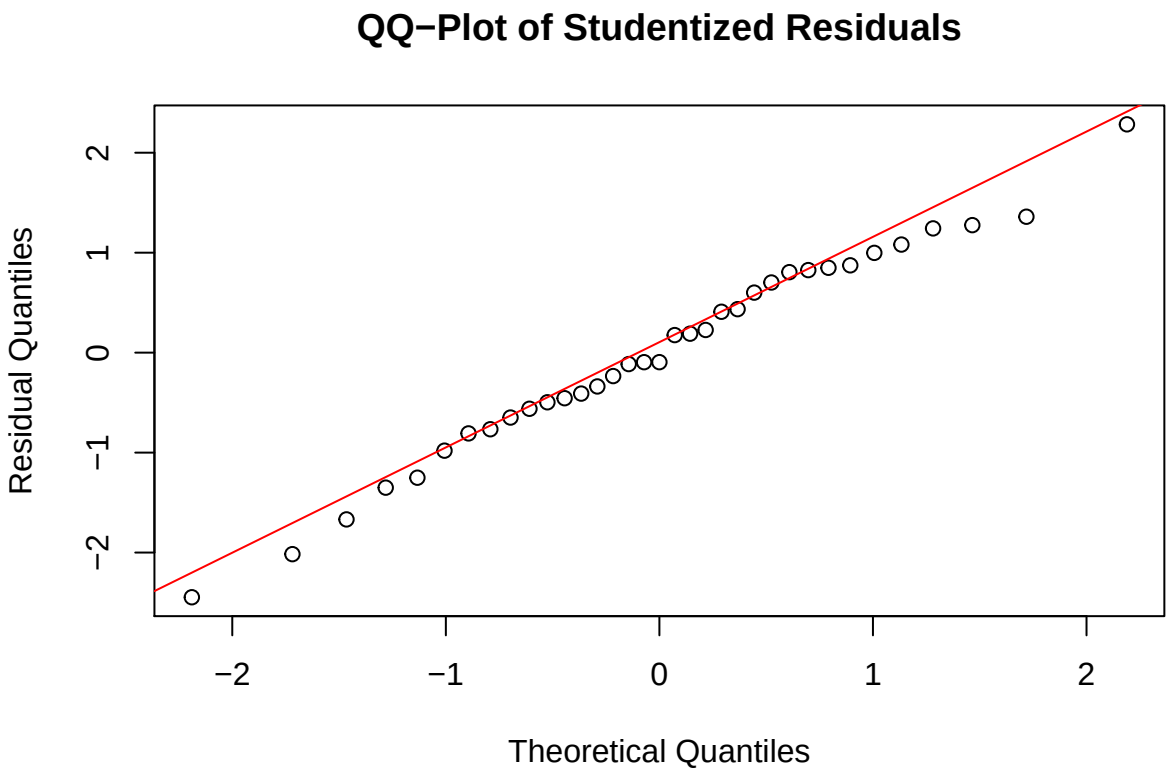

```
#Studentized Residuals vs Fitted Values (Log)
plot(log.model$fitted.values, studres(log.model), ylim = c(min(-3,
    min(studres(log.model))),max(3, max(studres(log.model)))),
     main = "Studentized Residuals vs. Fitted Values",
    ylab ="Studentized Residuals", xlab = "Fitted Values")
abline(h=c(0,3,-3), lty = c(1,2,2), col = "red")
```




```
#Histogram of Studentized Residuals (Log)
hist(studres(log.model), xlab = "Studentized Residuals", main = "Histogram of Studentized Residuals")
```




```r
#QQ-Plot of Studentized Residuals (Log)
qqnorm(studres(log.model), main = "QQ-Plot of Studentized Residuals", ylab = "Residual Quantiles")
qqline(studres(log.model), col = "red")
```



Comparing the plots from both (g) and (h), the most appropriate model is the one from (h) because we can see our assumptions can be better validated. Both Studentized Residuals vs (Index/Fitted Values) show no specific patterns indicating independence and constant variance where as both the histogram and qq plot show good signs of normality. There is a one outlier but nothing too worrisome.