Homework4

Datasets: Ollinger, Dodson

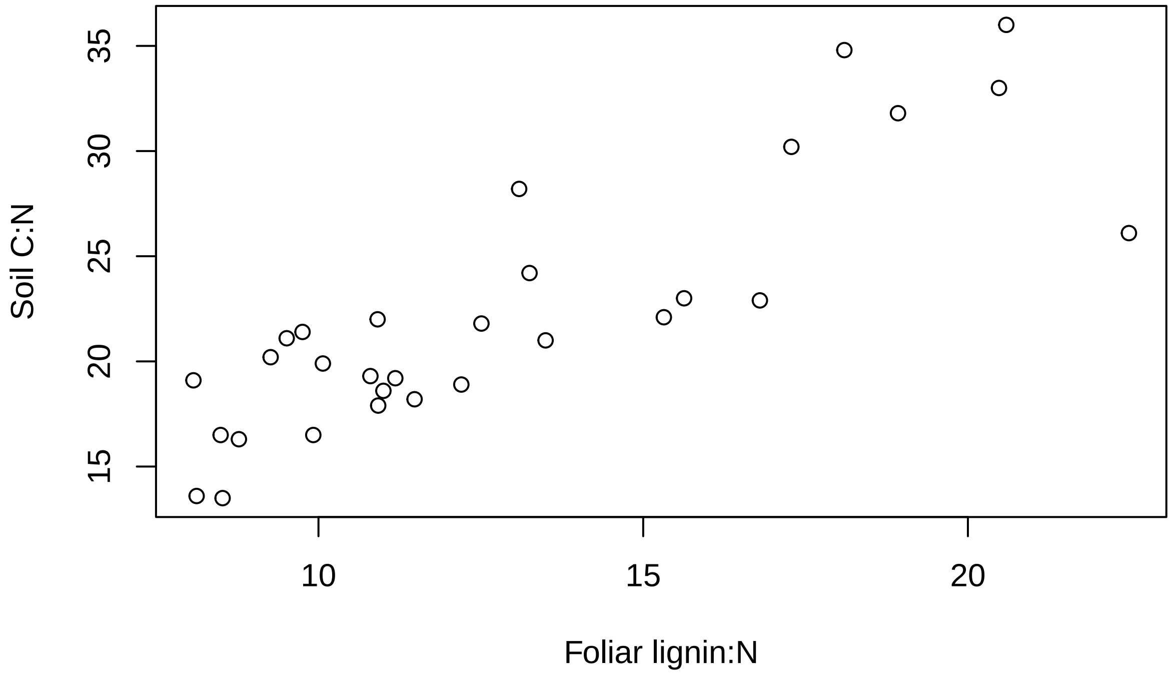
```
# Load dataset
ollinger <- readxl::read_excel("/Users/kanoalindiwe/Downloads/Projects/playground/R/Quantitative Ecolog
```

Question 1:

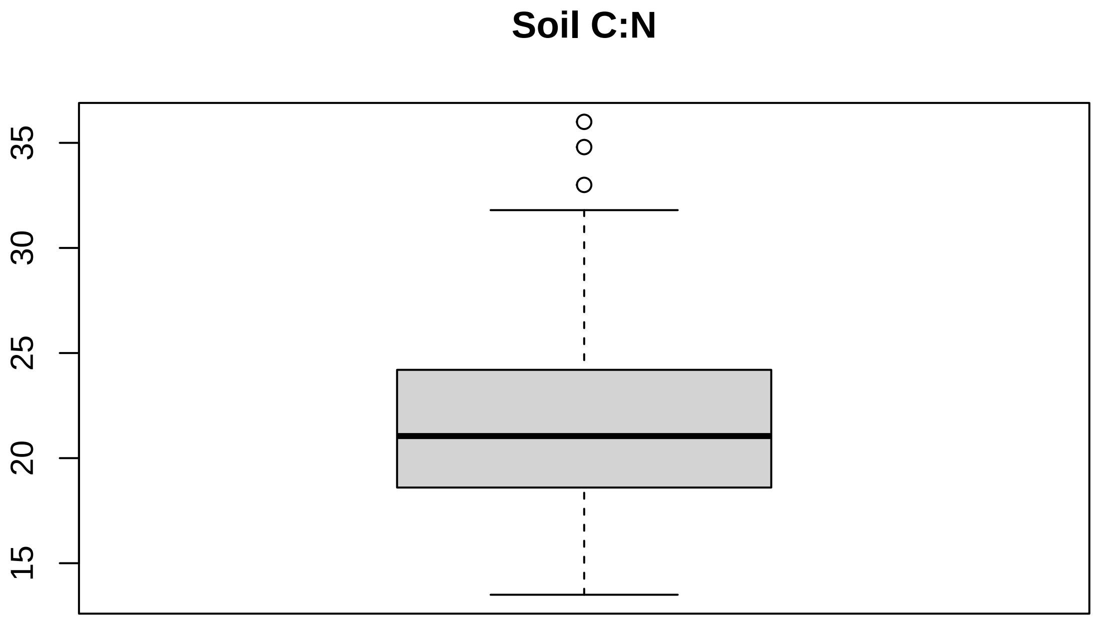
Ollinger et al. (2002) studied regional variation in the relationships between canopy chemistry, soil C:N ratios and soil N transformations in a forest in New Hampshire, USA. They sampled 30 plots in the forest, each plot with a different history of logging and a combination of one or more tree species (e.g. American beech, balsam fir, eastern hemlock etc.). For each plot, they measured the % lignin and % N from upper and mid-canopy foliage from trees of each species and combined these into single values for N and lignin. They also measured C:N ratios of replicate soil cores from each plot, as well as rates of N mineralization and nitrification and soil pH. Our interest is in the nature of the relationship between soil C:N ratios and foliar lignin:N ratios in the canopy. The latter can be measured across large spatial scales using various new remote sensing techniques (e.g. AVIRIS: Airborne Visible and InfraRed Imaging Spectrometer), whereas soil characteristics require time-consuming on-ground sampling, so predicting soil C:N from canopy lignin:N would be very useful.

a) Draw a scatterplot of soil C:N ratio against foliar lignin:N ratio, with boxplots and/ or histograms included for each variable (histograms for the y variable). Is there evidence of strong non-normality for either variable or nonlinearity in the relationship?

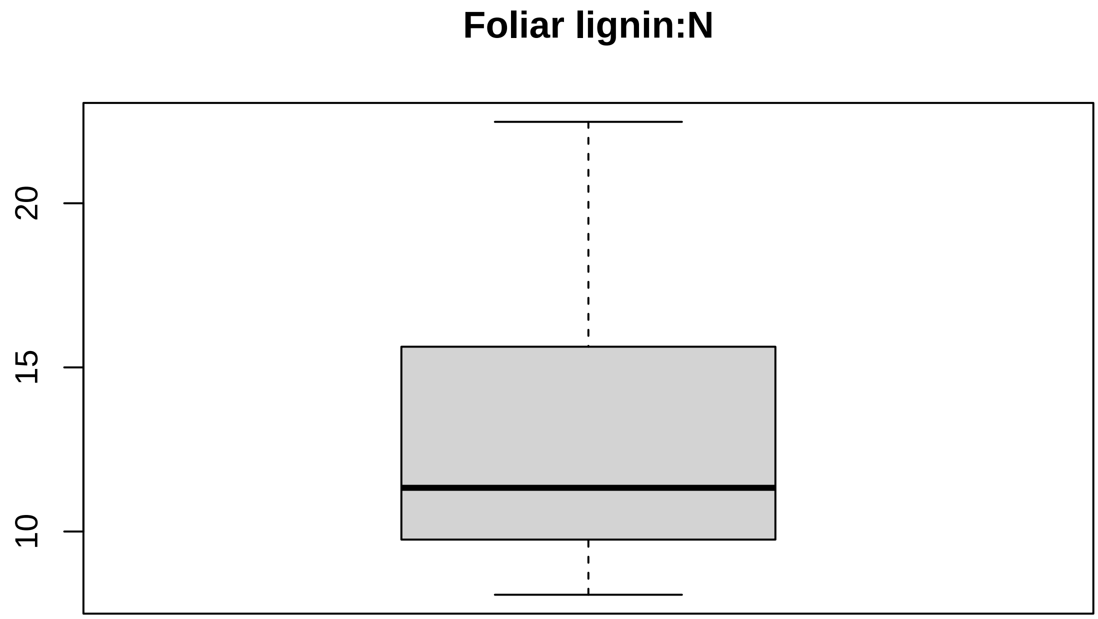
```
# Check assumptions
cn <- ollinger$SoilCN
ln <- ollinger$FoliarLigN

# Normality
plot(ln, cn, xlab = "Foliar lignin:N", ylab = "Soil C:N")
```



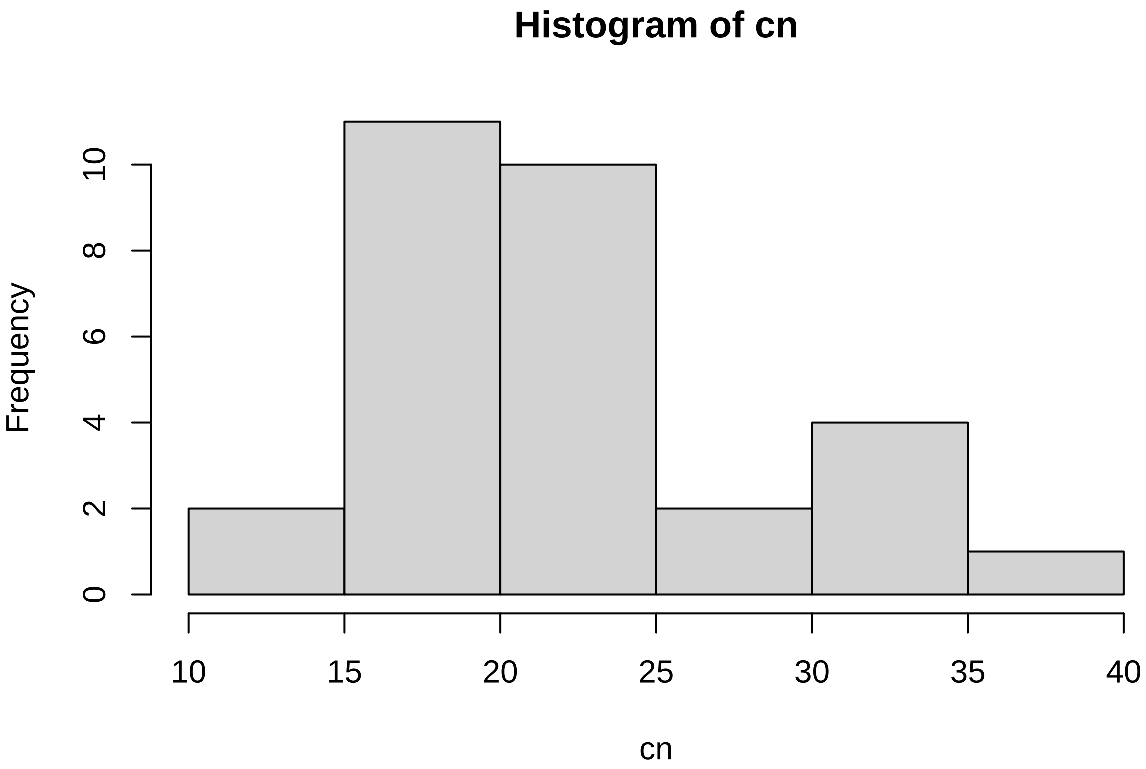
```
boxplot(cn, main="Soil C:N")
```

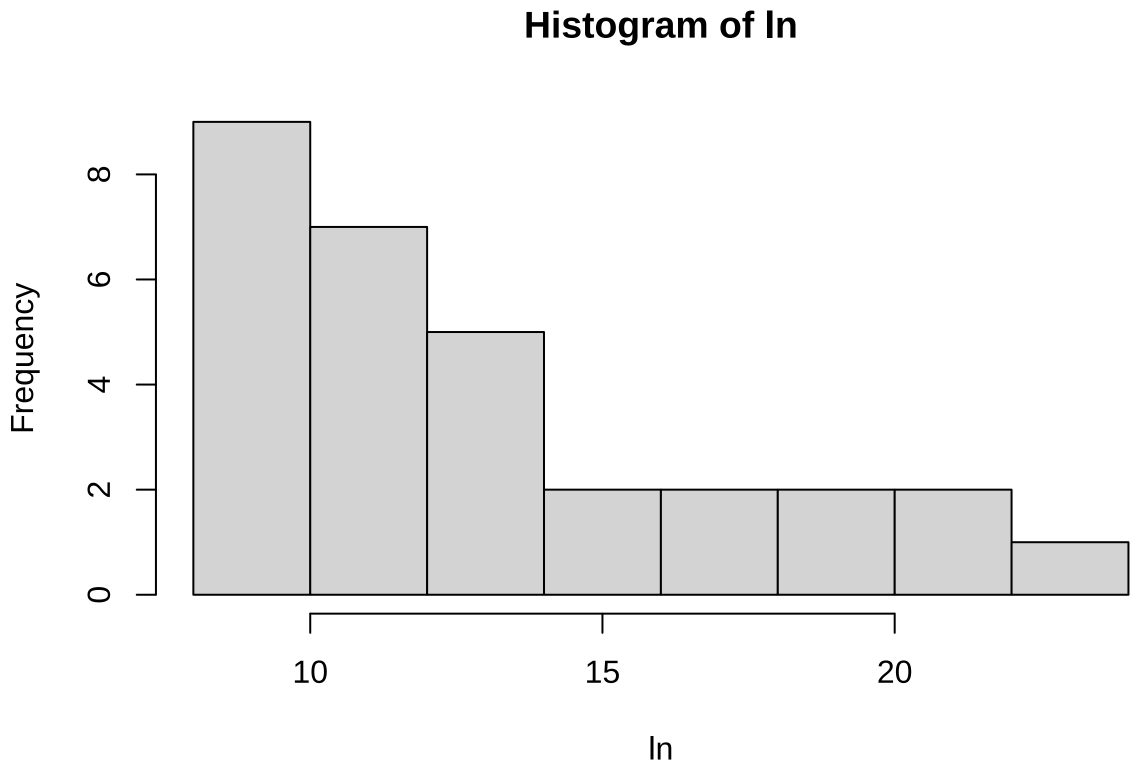


```r
boxplot(ln, main="Foliar lignin:N")
```

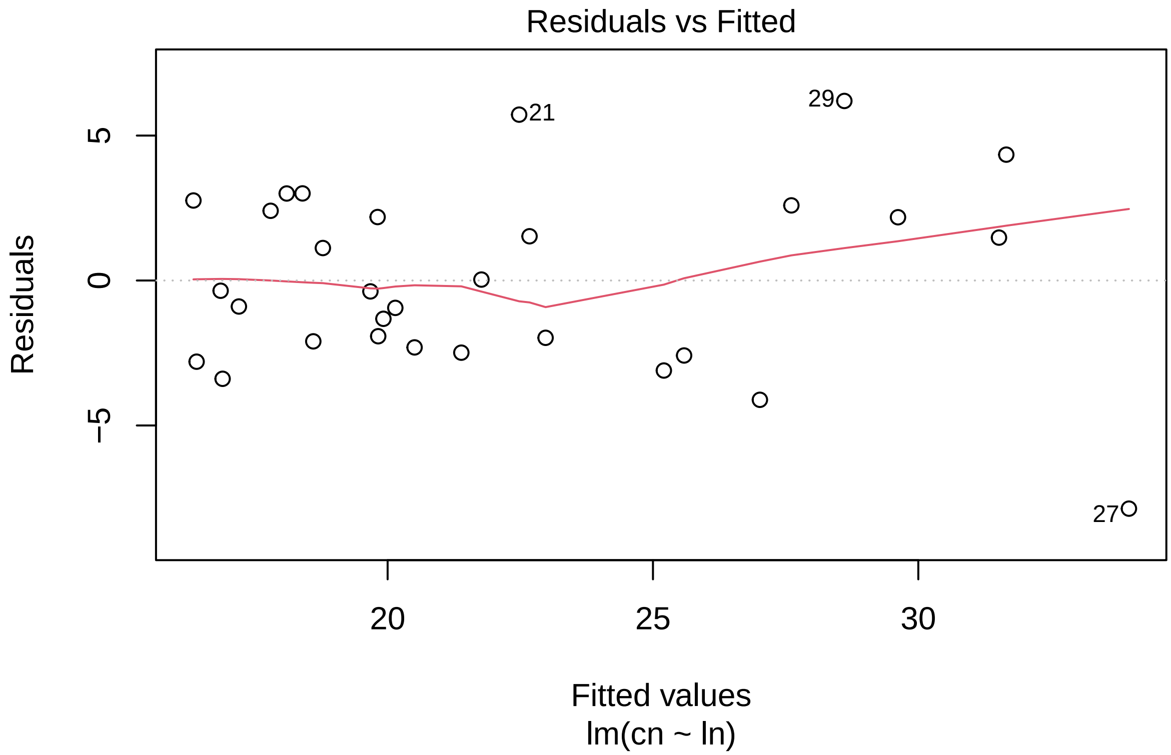


```
hist(cn); hist(ln)
```

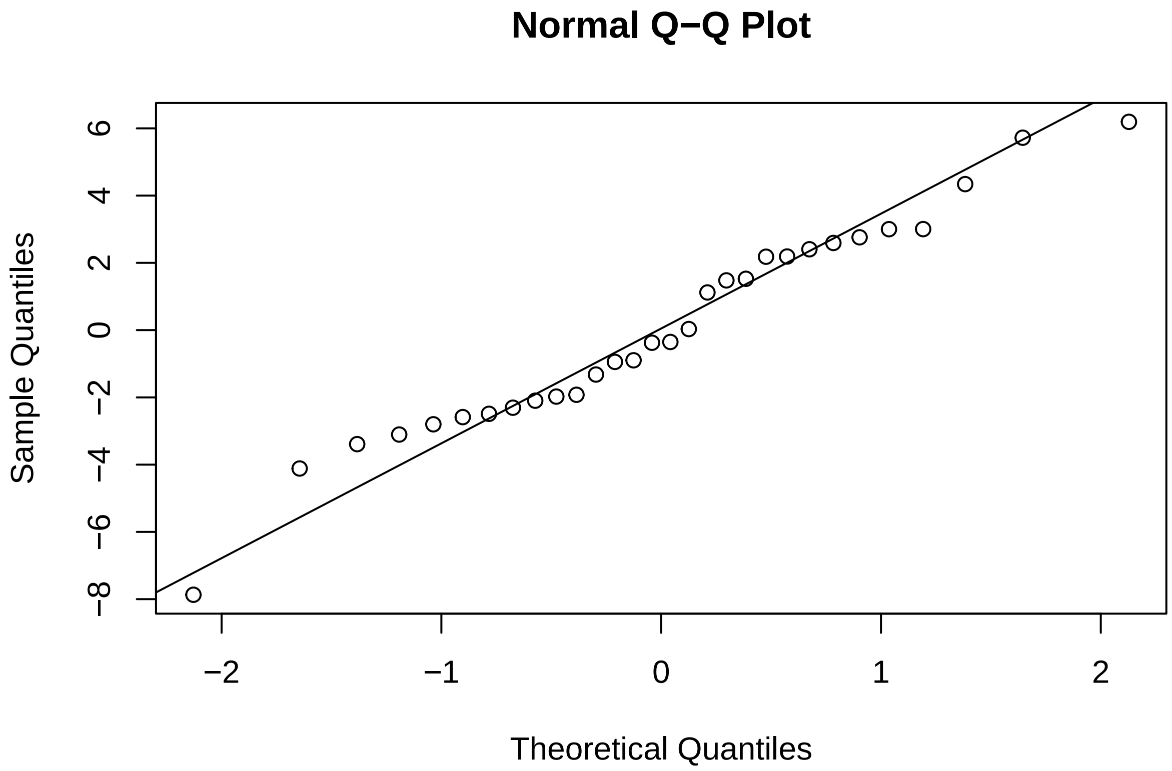

Histogram of cn
Frequency
0
2
4
6
8
10
10
15
20
25
30
35
40
cn



```r
# Nonlinearity
linearity_model <- lm(cn ~ ln)
plot(linearity_model, which = 1)
```



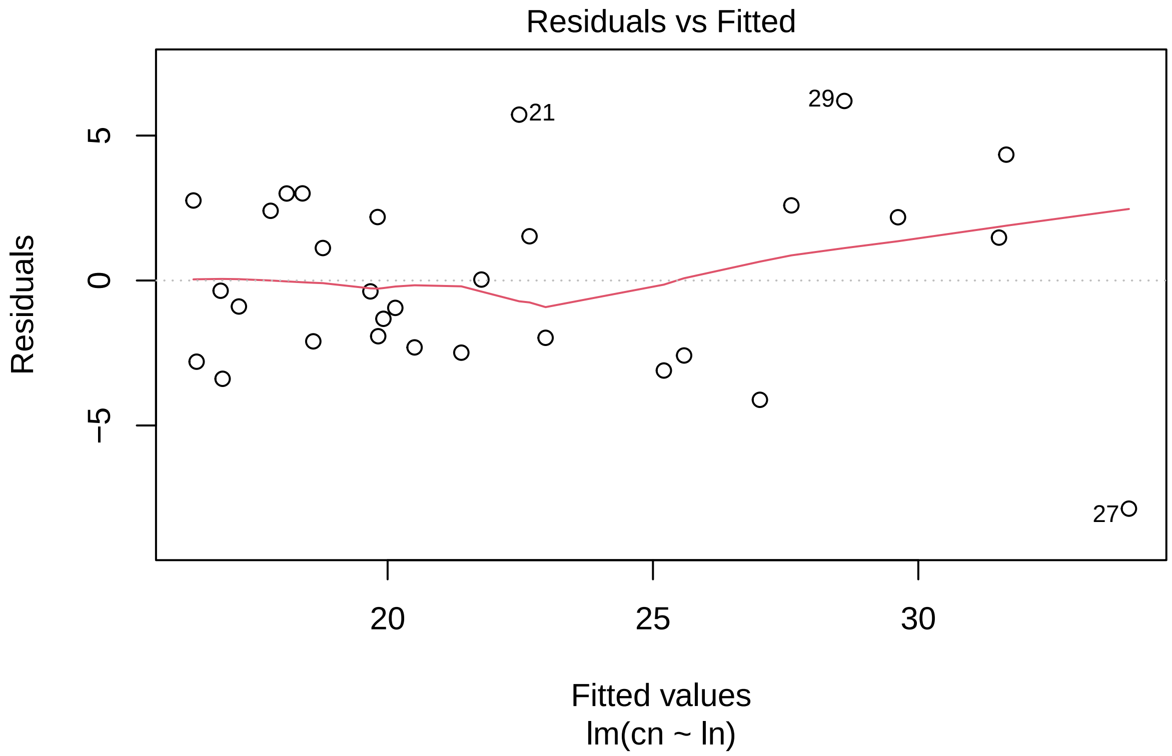
```r
qqnorm(residuals(linearity_model)); qqline(residuals(linearity_model))
```

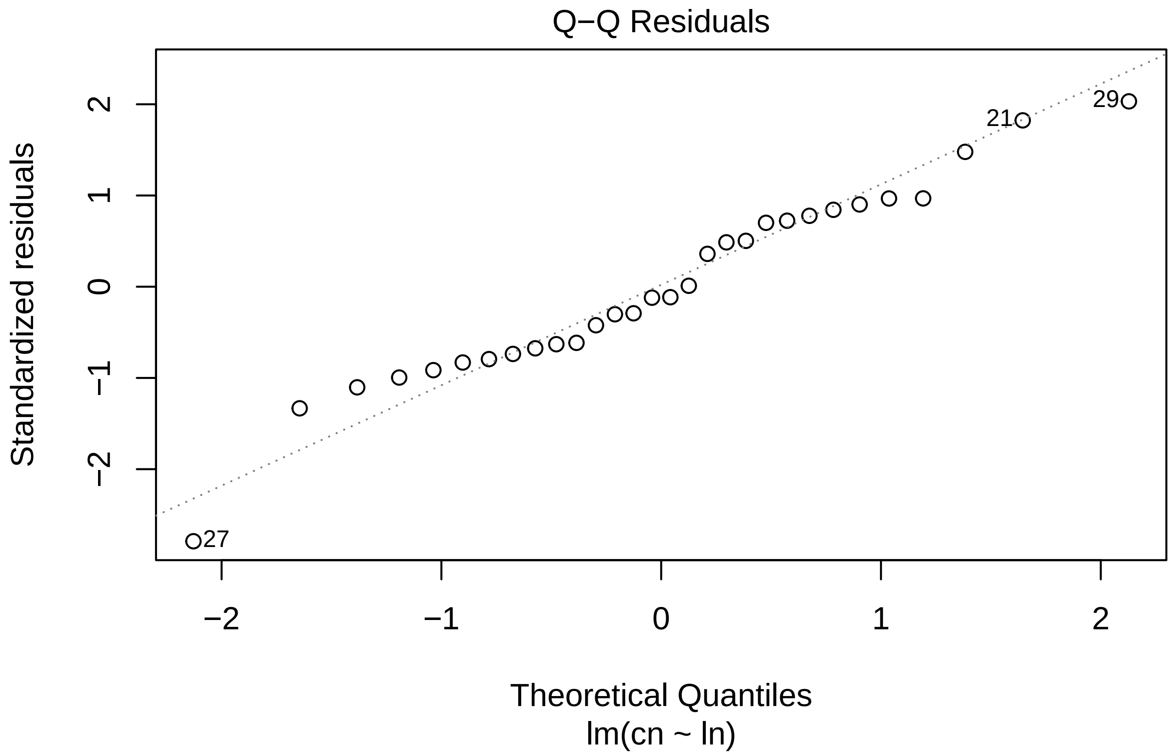


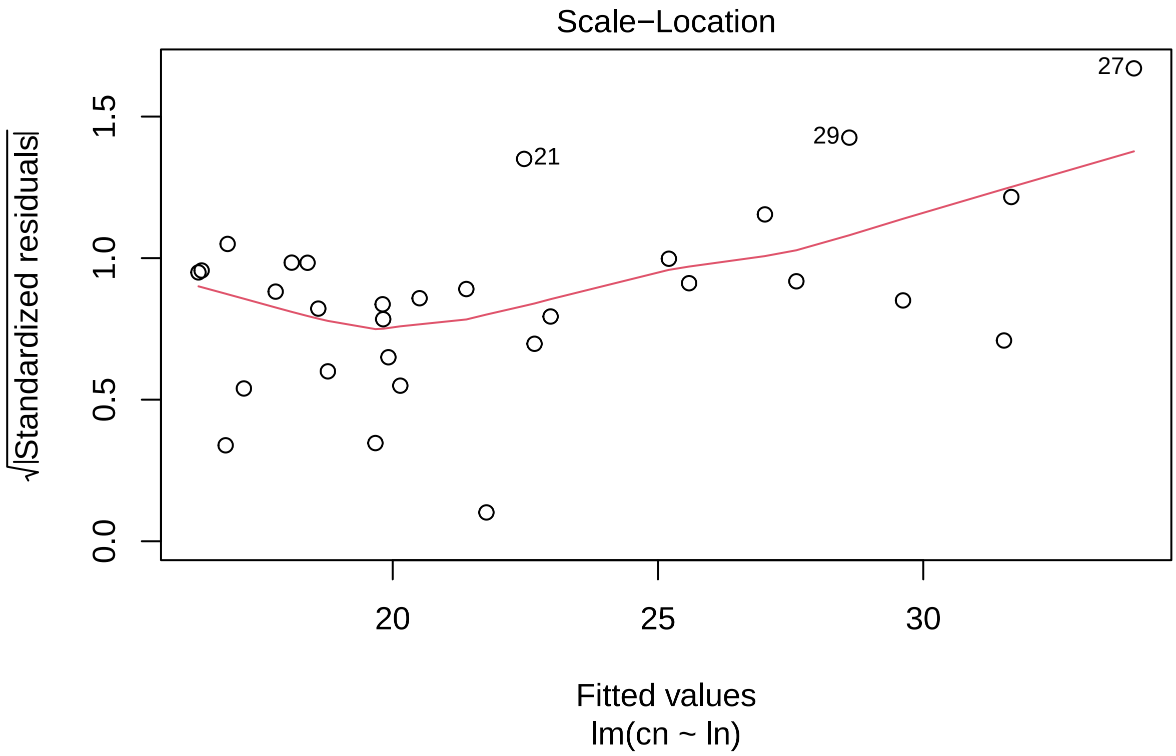
These data are not normal but regression does not require normality of input data. The residual plot does not show much correlation meaning that the variables are presumed to have a linear relationship.

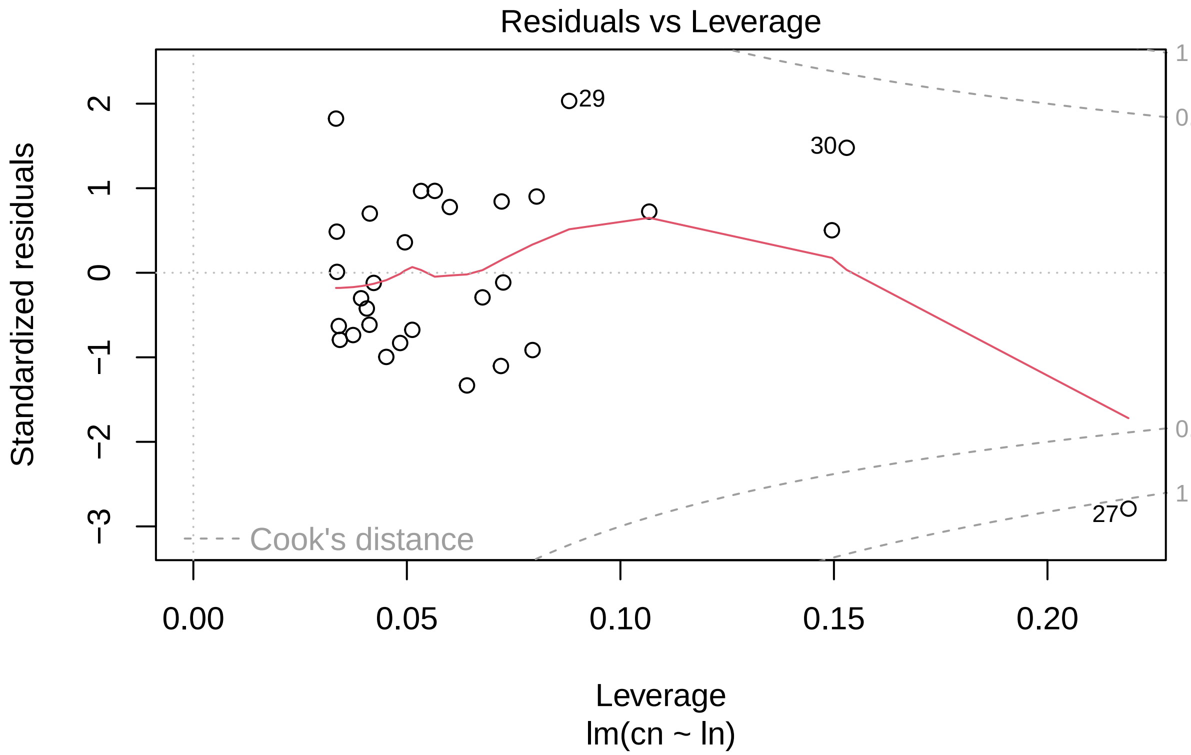
b) Now fit a linear regression model to these data, with soil C:N ratio as the response variable and foliar lignin:N ratio as the predictor. Examine the residuals from this model – any obvious problems? Any outliers or influential observations identified?

```r
#Build model
model <- lm(cn ~ ln)
plot(model)
```


Residuals vs Fitted
Residuals
5
0
−5
20
25
30
Fitted values
lm(cn ~ ln)
21
29
27


Q−Q Residuals
Standardized residuals
−2
−1
0
1
2
27
21
29
−2
−1
0
1
2
Theoretical Quantiles
lm(cn ~ ln)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
20
25
30
Fitted values
lm(cn ~ ln)
21
29
27



The residuals appear approximately normally distributed. Therefore, the normality assumption for linear regression is reasonably satisfied. We can setup our null and alternative hypotheses as: H0: slope = 0, no relationship between lignin:N and soil C:N; and H1: slope != 0, there is a relationship between lignin:N and soil C:N.

c) Summarize the results of the linear regression. Include the full regression model with confidence intervals on the parameter estimates and the test of the null hypothesis of zero slope.

```
# Get model results
summary(model)
```

```
##
## Call:
## lm(formula = cn ~ ln)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8690 -2.2553 -0.3643  2.3511  6.1934
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.4588     1.9405   3.328  0.00246 **
## ln            1.2238     0.1435   8.528 2.86e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3.191 on 28 degrees of freedom
## Multiple R-squared:  0.722,  Adjusted R-squared:  0.7121
## F-statistic: 72.72 on 1 and 28 DF,  p-value: 2.859e-09
```

The model shows that an increase in C:N ratio increased significantly with lignin:N ratio (slope=1.22 ± 0.1, t(28) = 8.5, p<0.001). The equation is as follows: Soil C:N = 6.46 + 1.22 × lignin:N. The null hypothesis (H0: slope = 0, no relationship between lignin:N and soil C:N) was rejected in favor of the alternative (H1: slope != 0). The model explained 72% of the variance in soil C:N (Rˆ2 = 0.72, adjusted Rˆ2 = 0.71), and the overall regression was highly significant (F(1, 28) = 72.7, p < 0.001).

d) What biological conclusions would you draw from this analysis? Can you predict soil C:N from canopy lignin:N? These results indicate that lignin:N in the canopy is a good predictor of soil C:N.

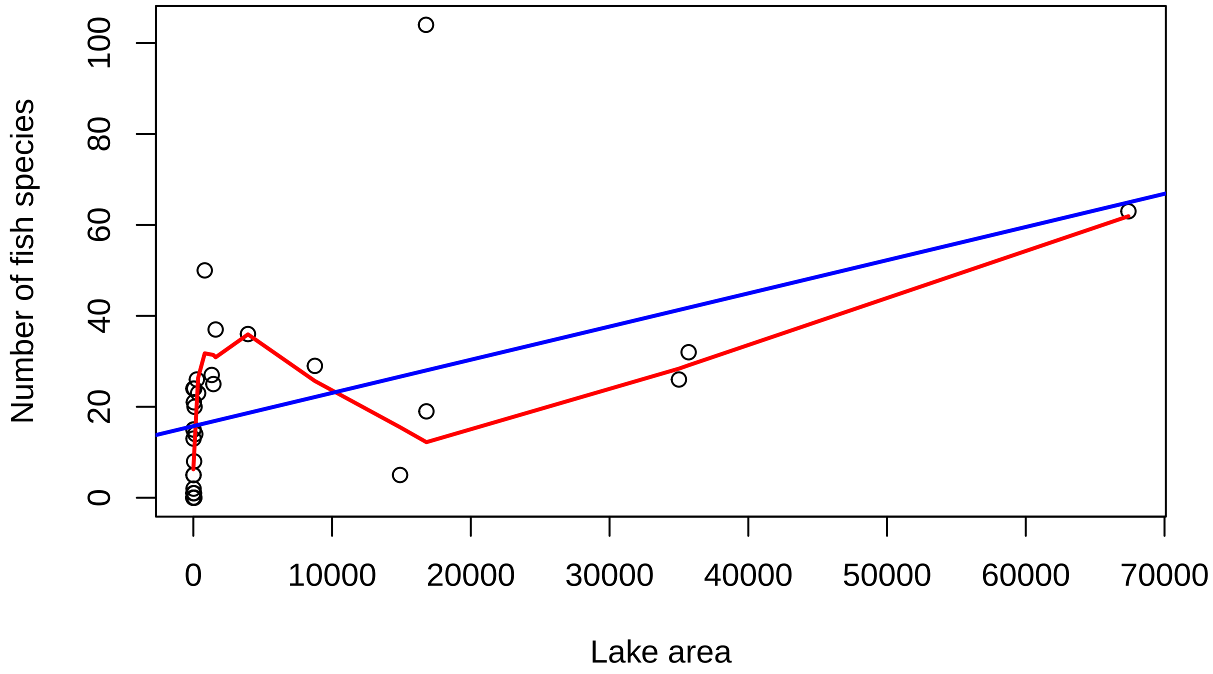
Question 2:

A positive relationship between number of species and area sampled (the species-area relationship) has long been considered one of ecology's few "laws". Using the data from Dodson et al. (2000), we will examine the relationships between the number of species of fish and copepods and lake area.

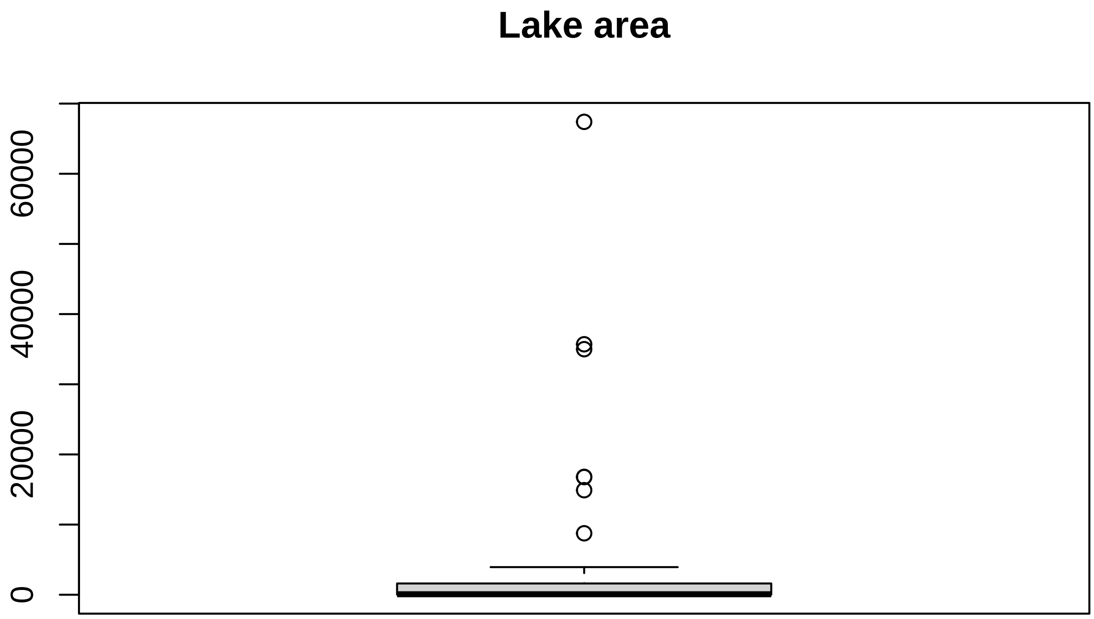
```r
# Import data
dodson <- readxl::read_excel("/Users/kanoalindiwe/Downloads/Projects/playground/R/Quantitative Ecology/
```

a) Draw a scatterplot of number of fish species against lake area, with boxplots and histograms for relevant variables. Also fit a Lowess smoothing function to the data. Is there evidence of non-normality for either variable or nonlinearity in the relationship?

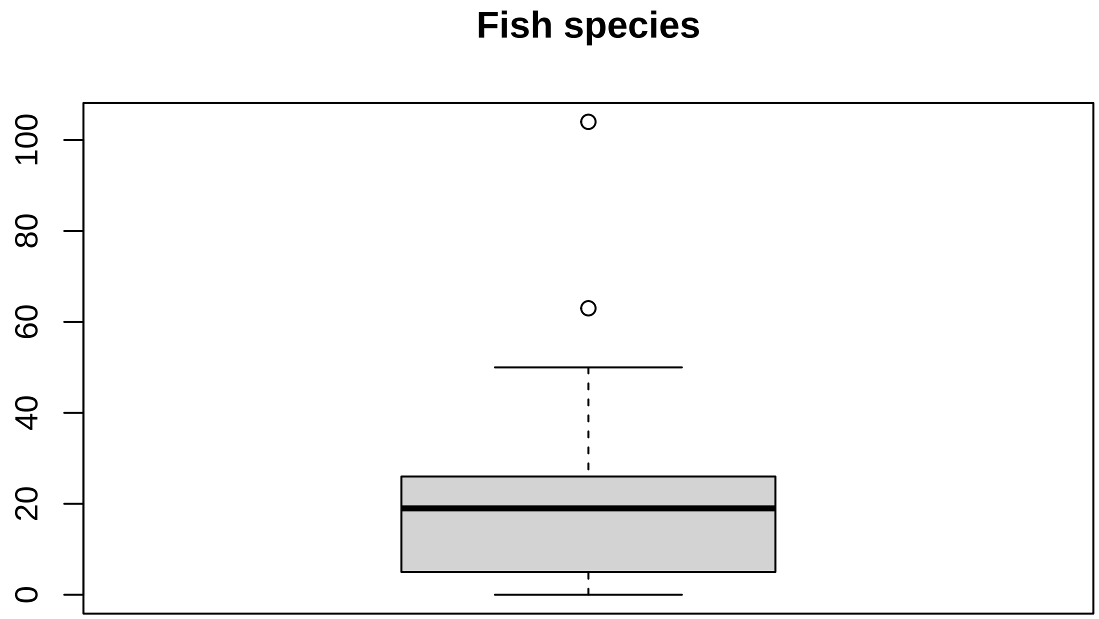
```r
# Get vars
dodson.species <- dodson$Fish
dodson.area <- dodson$Area

# Scatterplot with lowess for nonlinearity
plot(dodson.area, dodson.species,
     xlab = "Lake area",
     ylab = "Number of fish species")
lines(lowess(dodson.area, dodson.species), col = "red", lwd = 2)
abline(lm(dodson.species ~ dodson.area), col = "blue", lwd = 2)
```



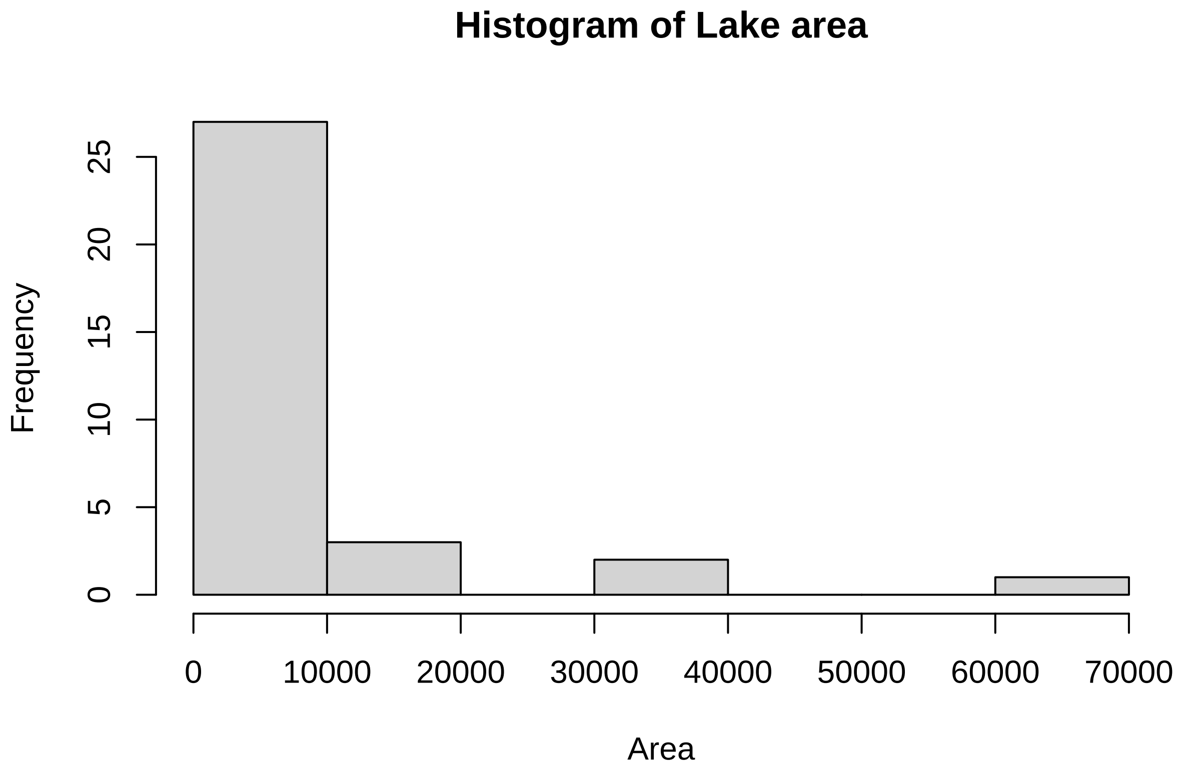
```
# Boxplots
boxplot(dodson.area, main="Lake area")
```



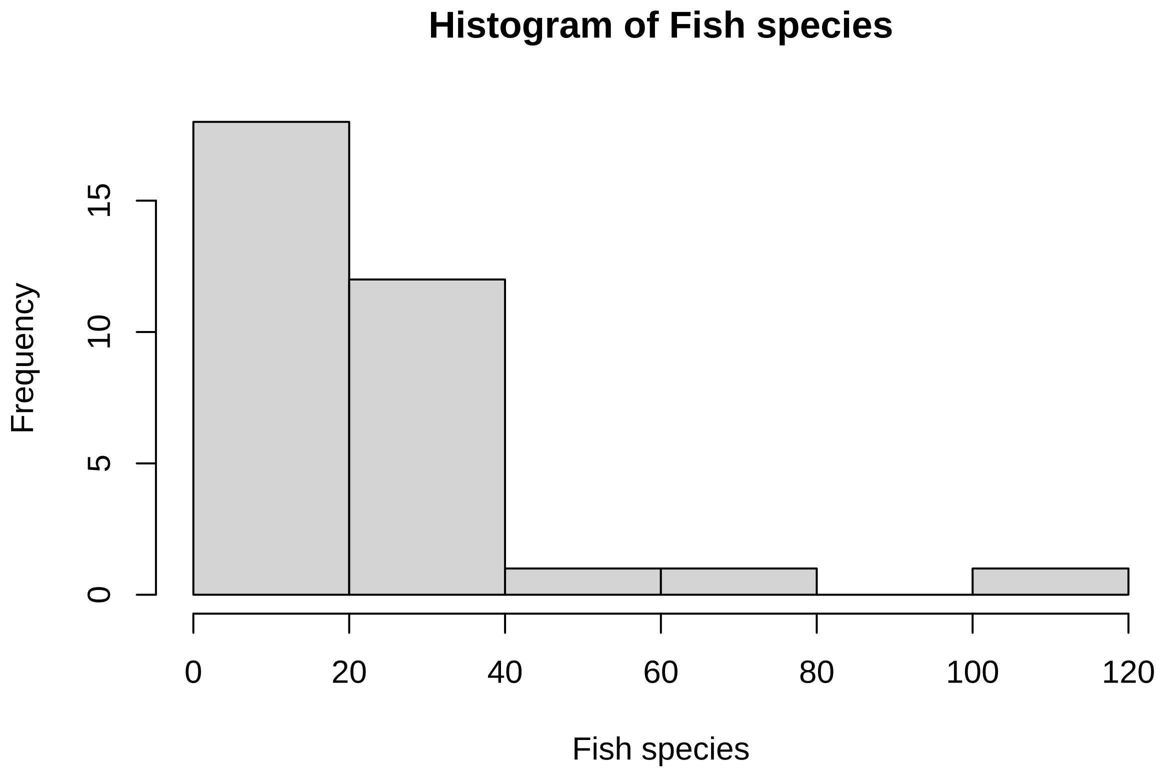
```
boxplot(dodson.species, main="Fish species")
```



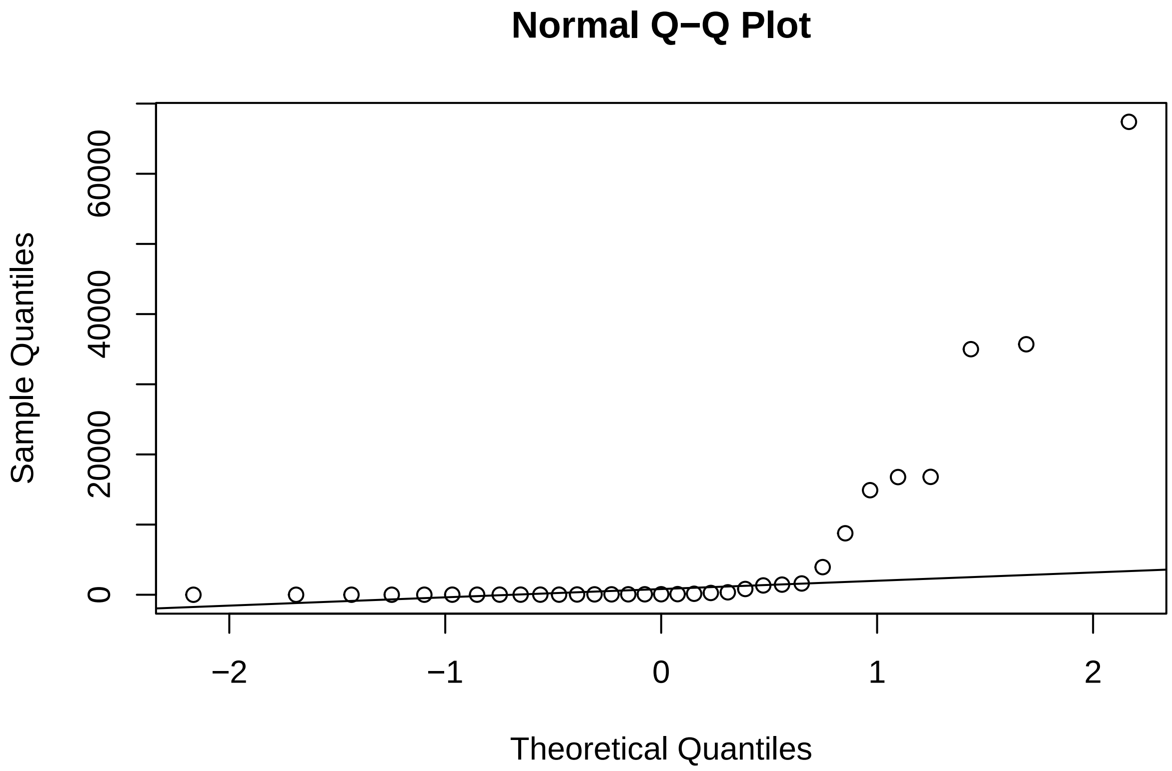
```
# Histograms
hist(dodson.area, main="Histogram of Lake area", xlab="Area")
```



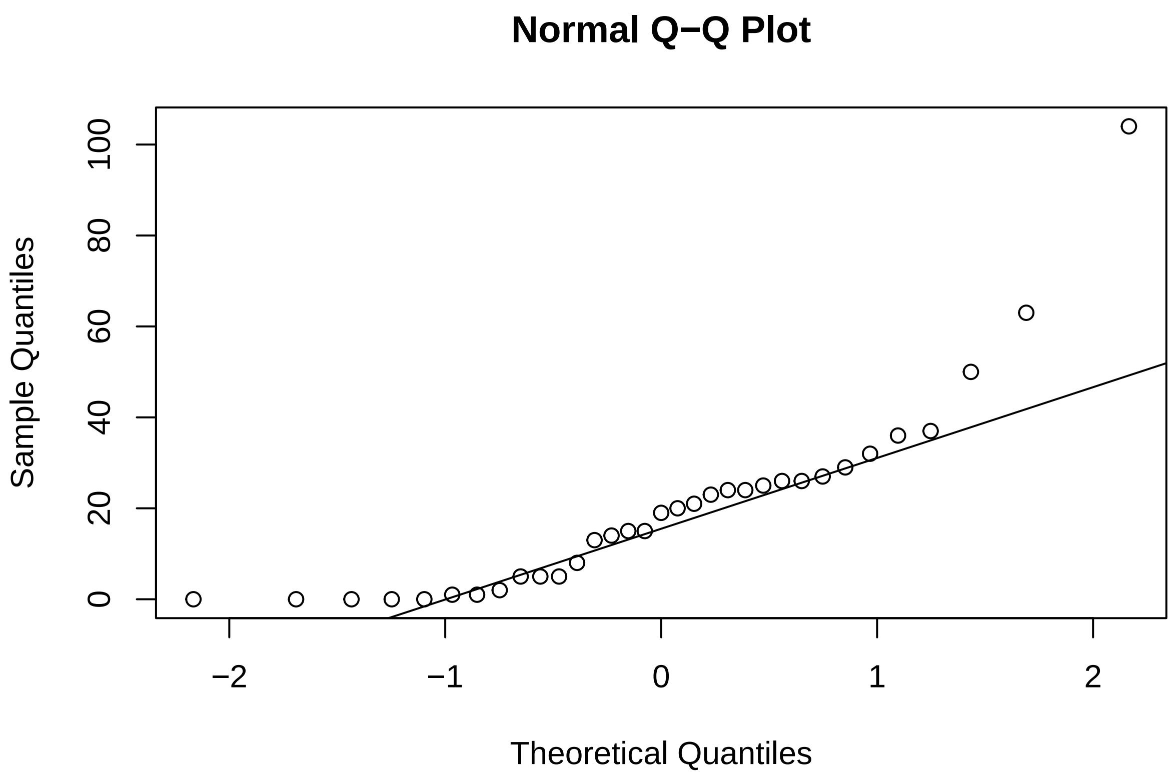
```r
hist(dodson.species, main="Histogram of Fish species", xlab="Fish species")
```



```
# QQ plots
qqnorm(dodson.area); qqline(dodson.area)
```



```r
qqnorm(dodson.species); qqline(dodson.species)
```

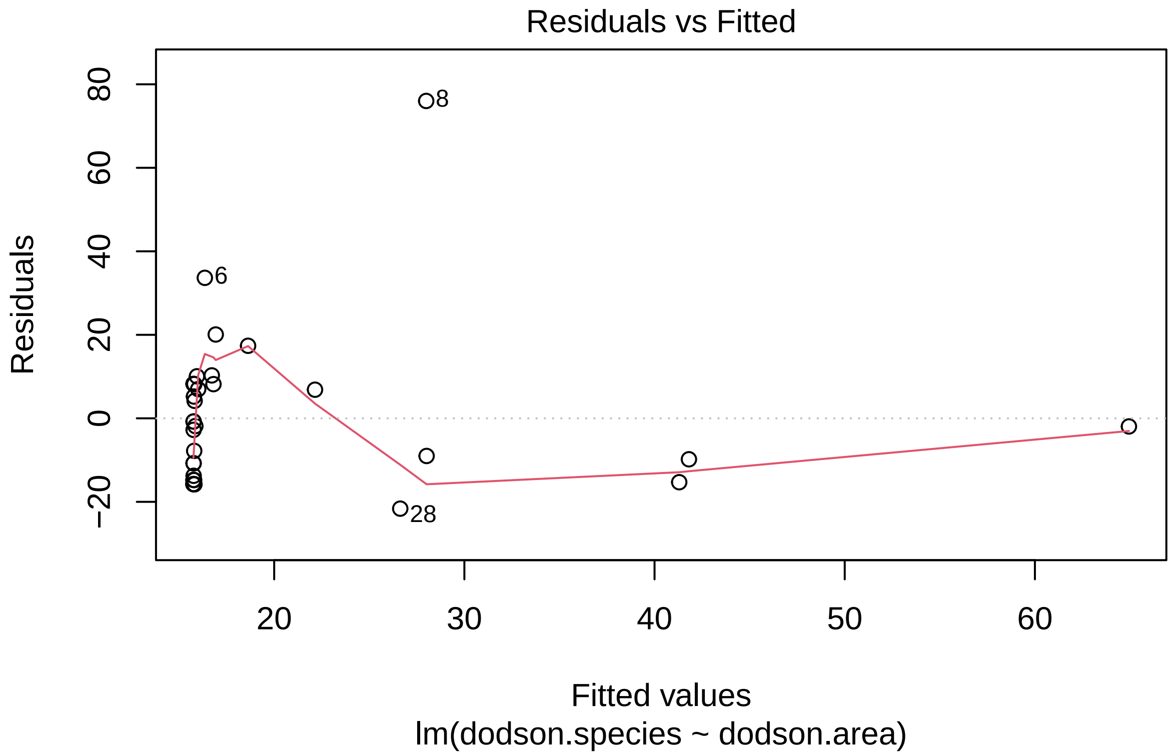


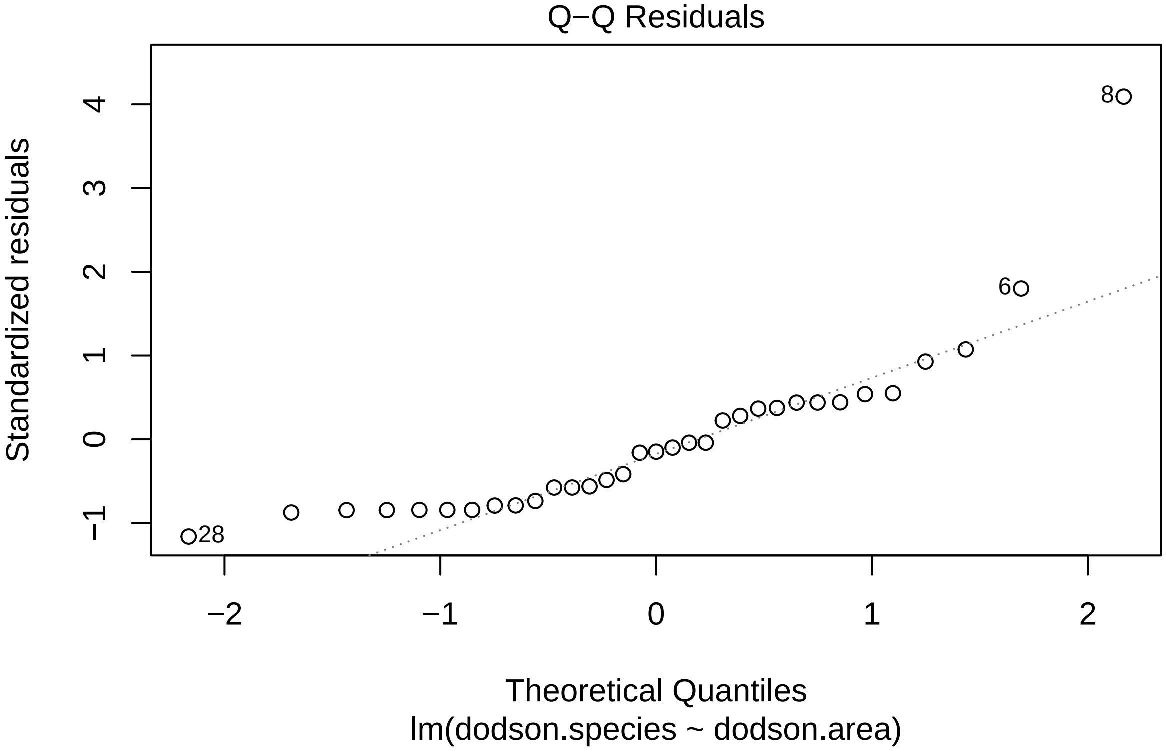
```
# Log will be explored below
```

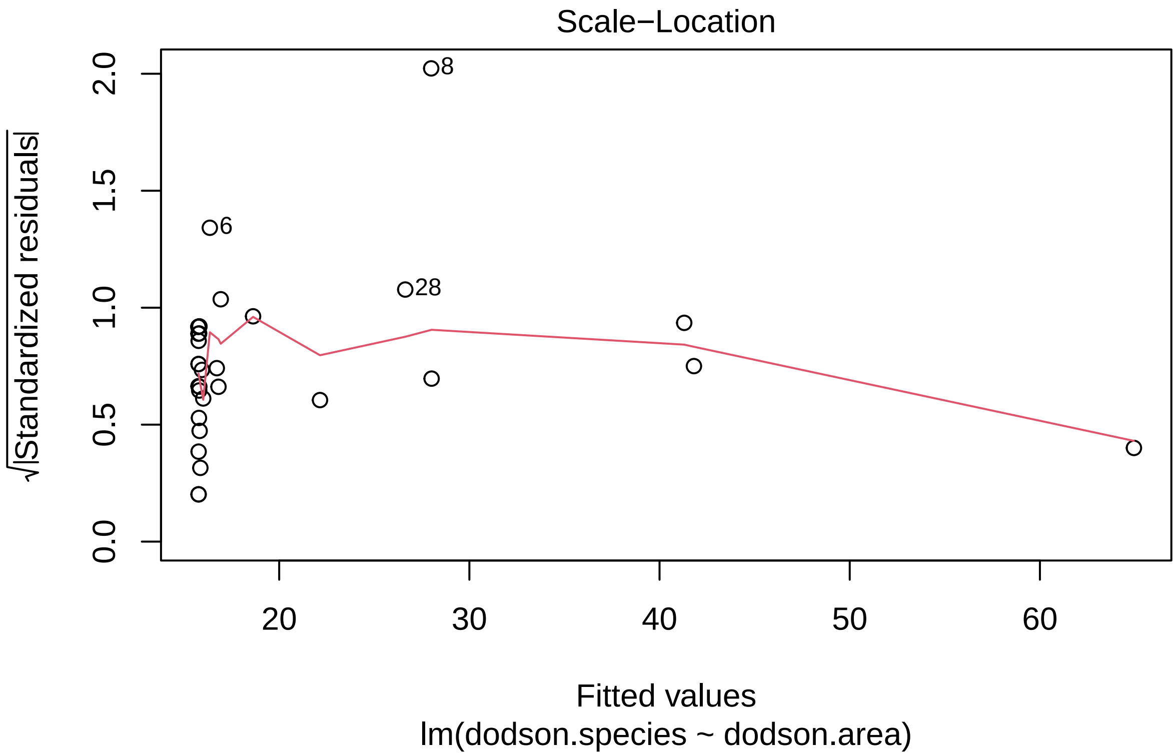
The qqplot indicates non-normal data, however, linear regression does not need normal data. Lowess indicates nonlinearity in the data as the two line deviate from eachother. After log-log transformation, the species-area relationship appears more normal. This indicates diminishing gains in species richness with increasing lake area.

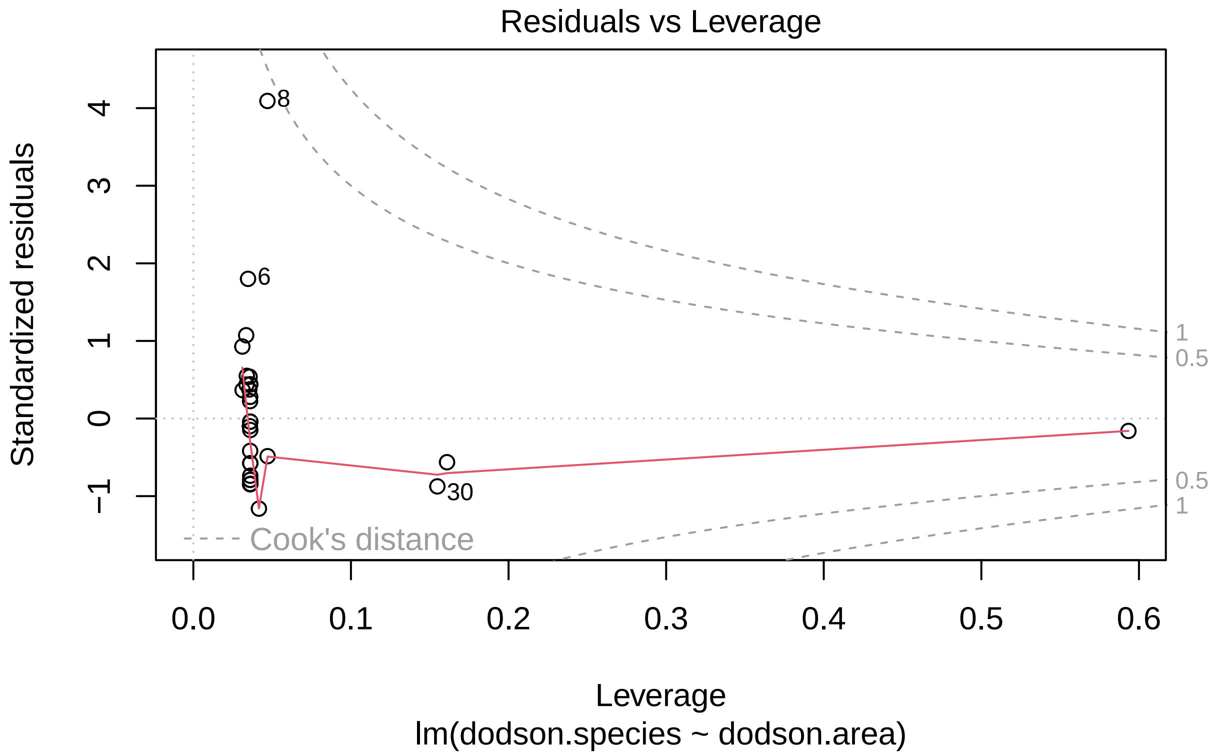
b) Now fit a linear regression model to these data, with number of fish species as the response variable and lake area as the predictor. Examine the residuals from this model – any obvious problems? Any outliers or influential observations identified?

```
# Make linear model
model <- lm(dodson.species ~ dodson.area)
plot(model)
```


Residuals vs Fitted
Residuals
80
60
40
20
0
−20
8
6
28
20
30
40
50
60
Fitted values
lm(dodson.species ~ dodson.area)


Q−Q Residuals
Standardized residuals
−1
0
1
2
3
4
−2
−1
0
1
2
Theoretical Quantiles
lm(dodson.species ~ dodson.area)
8
6
28


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
20
30
40
50
60
8
6
28
Fitted values
lm(dodson.species ~ dodson.area)



```
summary(model)
```

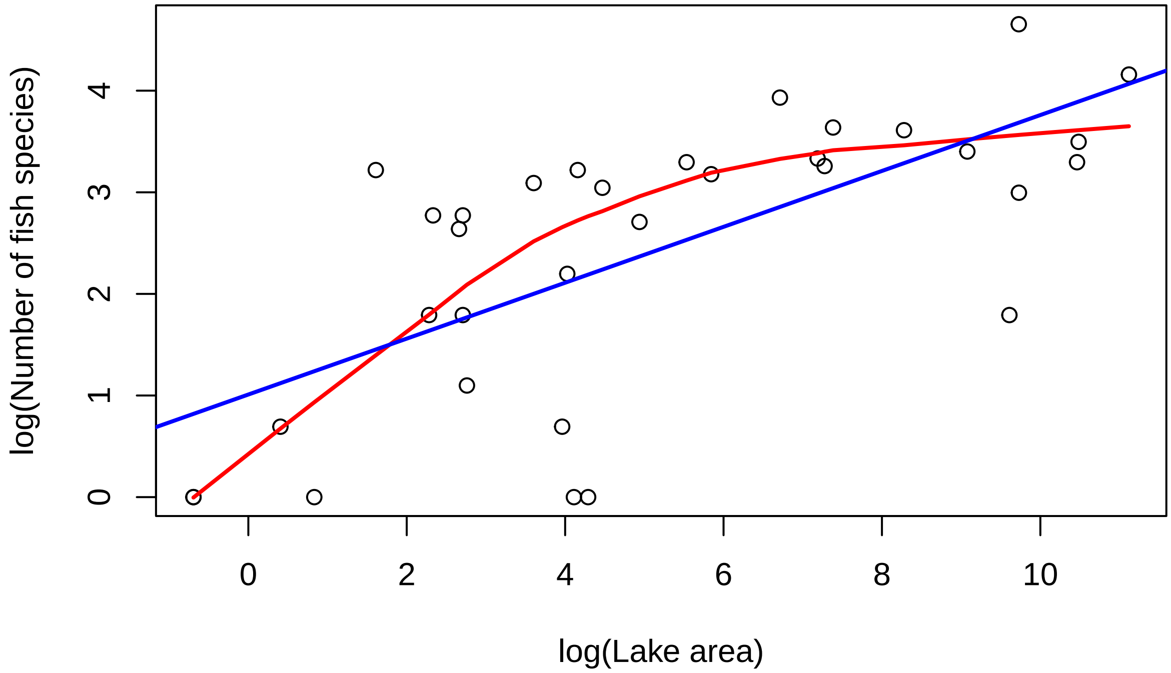
```
##
## Call:
## lm(formula = dodson.species ~ dodson.area)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.628 -14.754  -1.945   8.192  76.007
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.575e+01  3.618e+00   4.355 0.000135 ***
## dodson.area 7.298e-04  2.334e-04   3.127 0.003823 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.03 on 31 degrees of freedom
## Multiple R-squared:  0.2398, Adjusted R-squared:  0.2153
## F-statistic: 9.778 on 1 and 31 DF,  p-value: 0.003823
```

These results indicate that a linear model may not fit the data appropriately, and, given that the data did not conform to nonlinearity, this makes sense. It seems that the model does not predict the the number of fish species in lakes of small or large sizes. There are a handful of outlines on qq residual plot that are easy to spot.

These assessments of the data and adequacy of the model probably convinced you that transformation of both variables might be appropriate. For ex, Dodson decided to log transform.

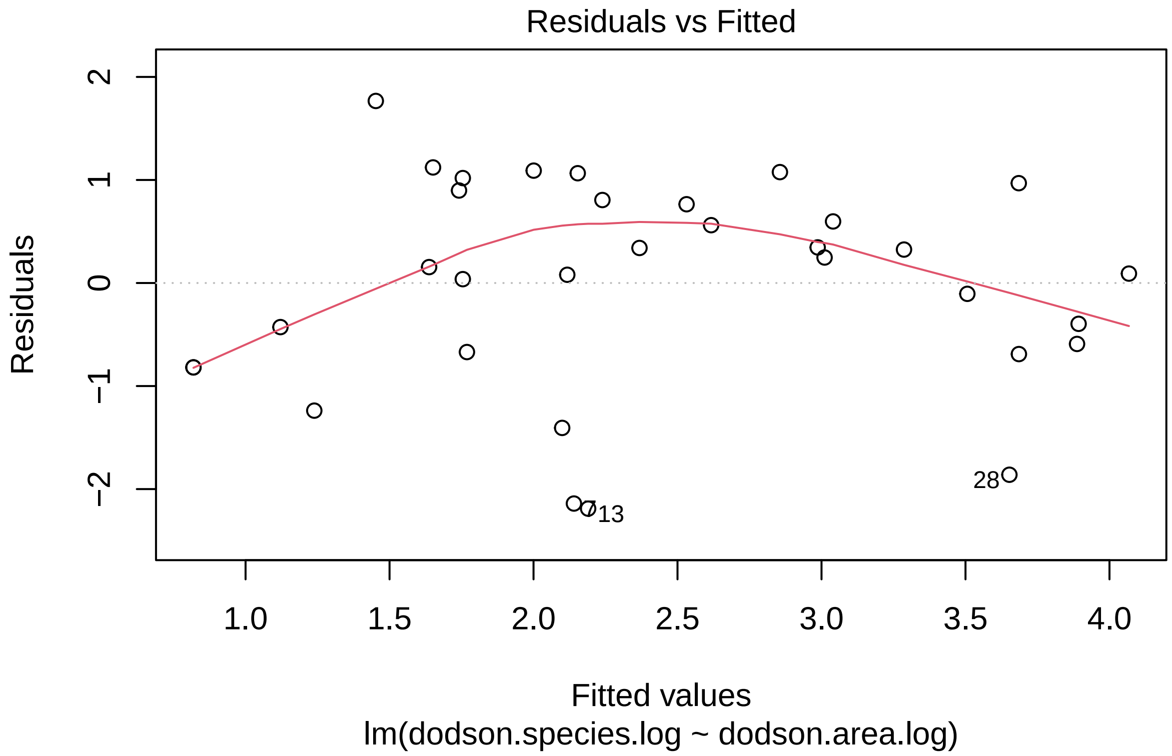
c) Redo the scatterplot and Lowess smoother using transformed variables – do the boxplots appear more symmetrical and a linear relationship more plausible?

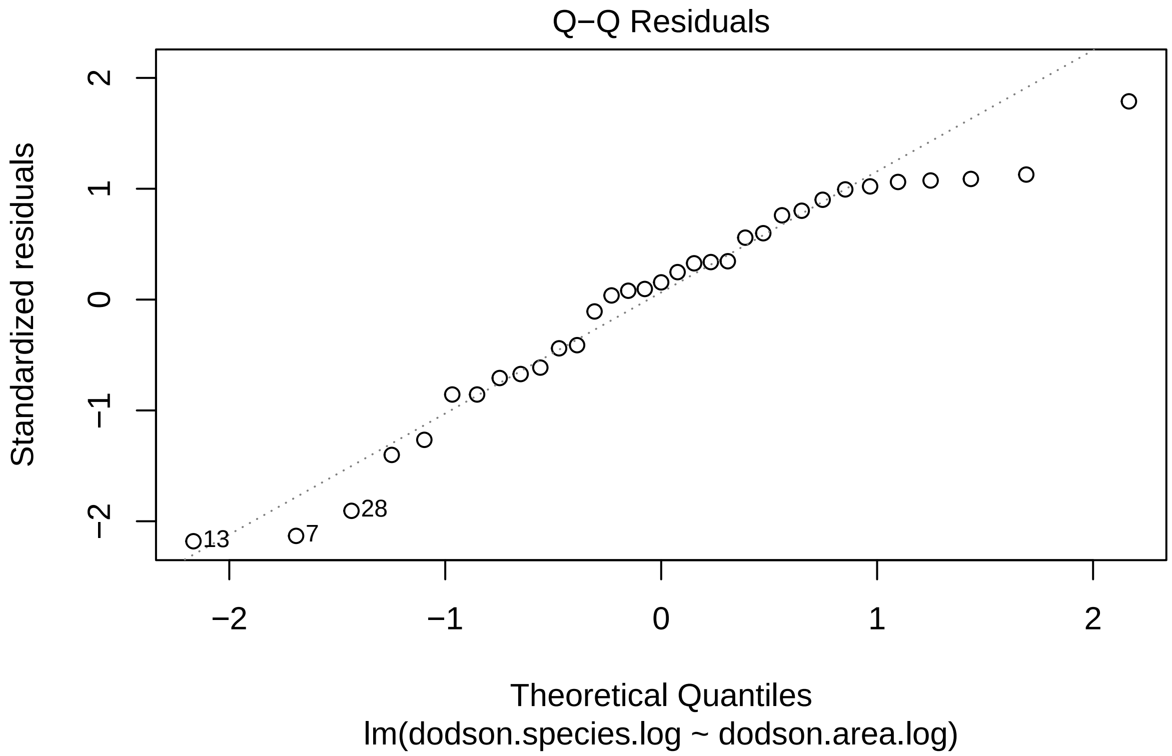
```
# Test log transform
dodson.species.log <- log1p(dodson$Fish)
dodson.area.log <- log(dodson$Area)

plot(dodson.area.log, dodson.species.log,
     xlab = "log(Lake area)",
     ylab = "log(Number of fish species)")
lines(lowess(dodson.area.log, dodson.species.log), col="red", lwd=2)
abline(lm(dodson.species.log ~ dodson.area.log), col="blue", lwd=2)
```

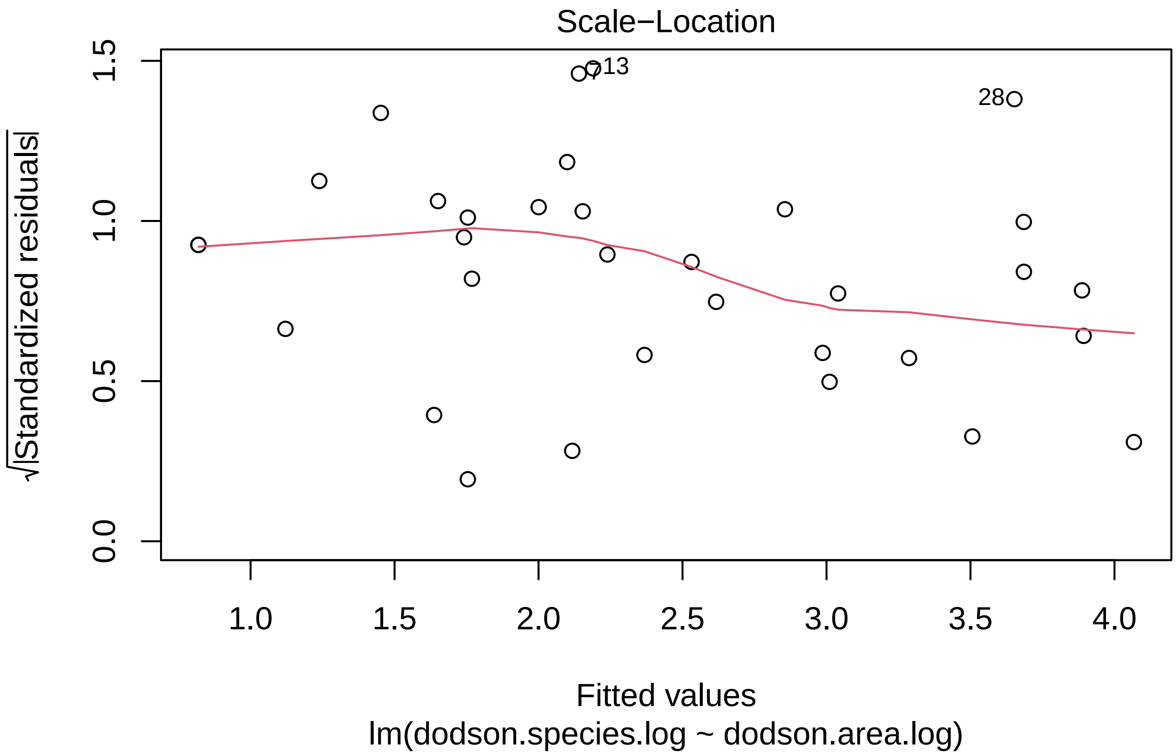


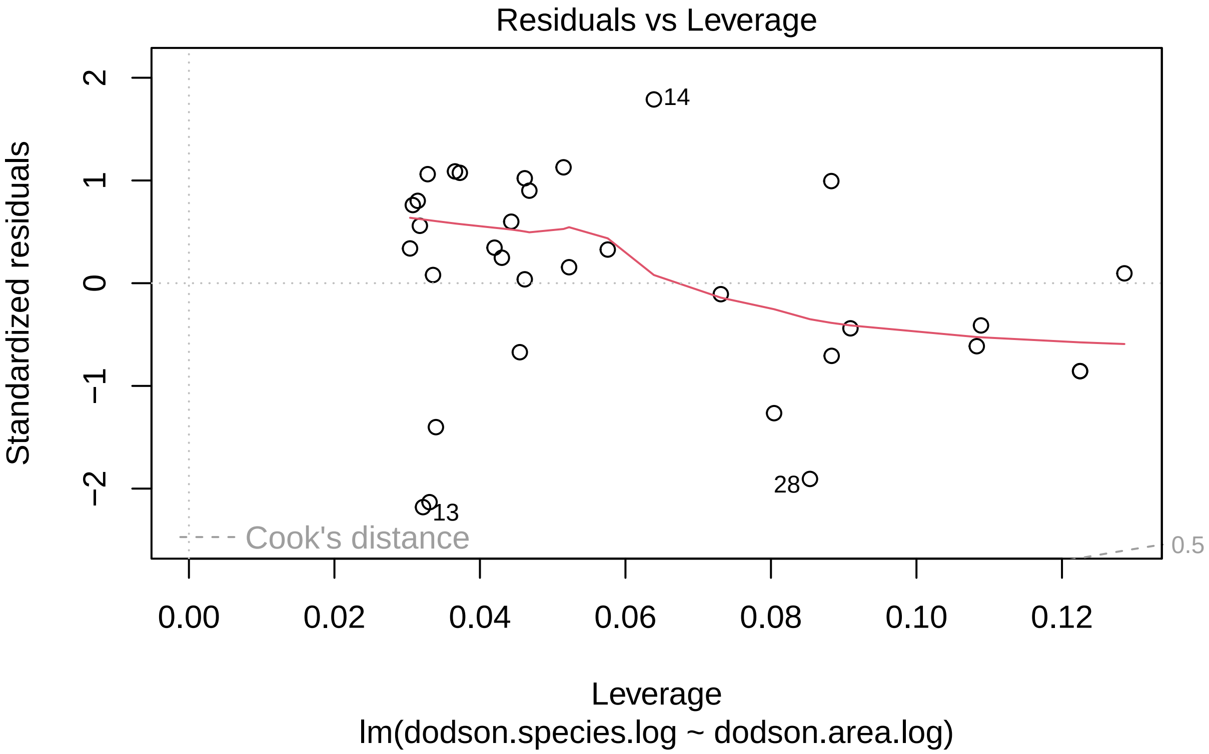
d) Refit a linear regression model using transformed data and examine the residuals. Are they more reasonable than they were with untransformed variables?

```
# Create model for log data
model <- lm(dodson.species.log ~ dodson.area.log )
plot(model)
```


Residuals vs Fitted
Residuals
2
1
0
−1
−2
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Fitted values
lm(dodson.species.log ~ dodson.area.log)
7
13
28


Q−Q Residuals
Standardized residuals
2
1
0
−1
−2
13
7
28
−2
−1
0
1
2
Theoretical Quantiles
lm(dodson.species.log ~ dodson.area.log)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Fitted values
lm(dodson.species.log ~ dodson.area.log)
7
13
28



Performing a log-log transform greatly reduces the residuals on the high and low ends of lake sizes. This makes for a better model.

e) Write out the results from your linear regression analysis, including the test of the null hypothesis of zero slope.

The non transformed linear regression revealed a significant positive relationship between the number of fish species and lake area. The slope was 328.5 ± 105.1 (t(31) = 3.13, p = 0.0038), indicating that lake area increases with species richness. The fitted equation was lake area = −436.4 + 328.5 × fish species. The model explained about 24% of the variation in lake area (Rˆ2 = 0.2398, adjusted Rˆ2 = 0.2153), and the overall regression was significant (F(1, 31) = 9.78, p = 0.0038), leading us to reject the null hypothesis of zero slope.

```
summary(model)
```

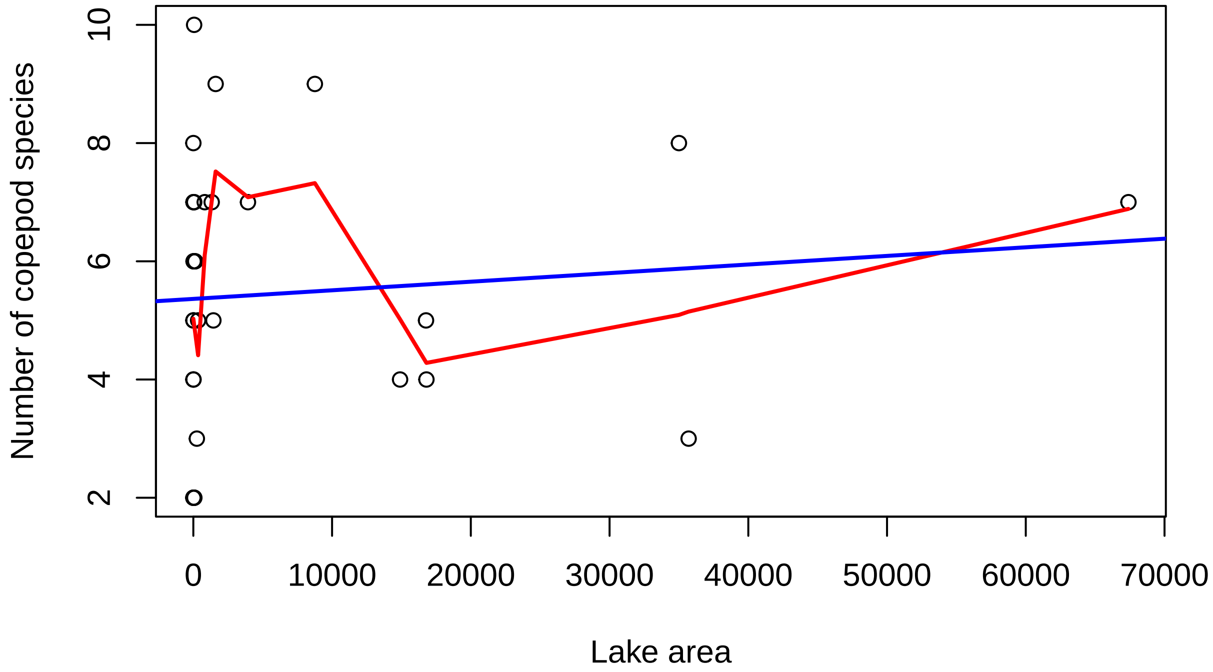
```
##
## Call:
## lm(formula = dodson.species.log ~ dodson.area.log)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1896 -0.6701  0.1544  0.8054  1.7666
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.00959    0.32572   3.100   0.0041 **
## dodson.area.log  0.27504    0.05334   5.157 1.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.021 on 31 degrees of freedom
## Multiple R-squared:  0.4617, Adjusted R-squared:  0.4444
## F-statistic: 26.59 on 1 and 31 DF,  p-value: 1.37e-05
```

f) What conclusions would you draw from the regression analysis using transformed variables?

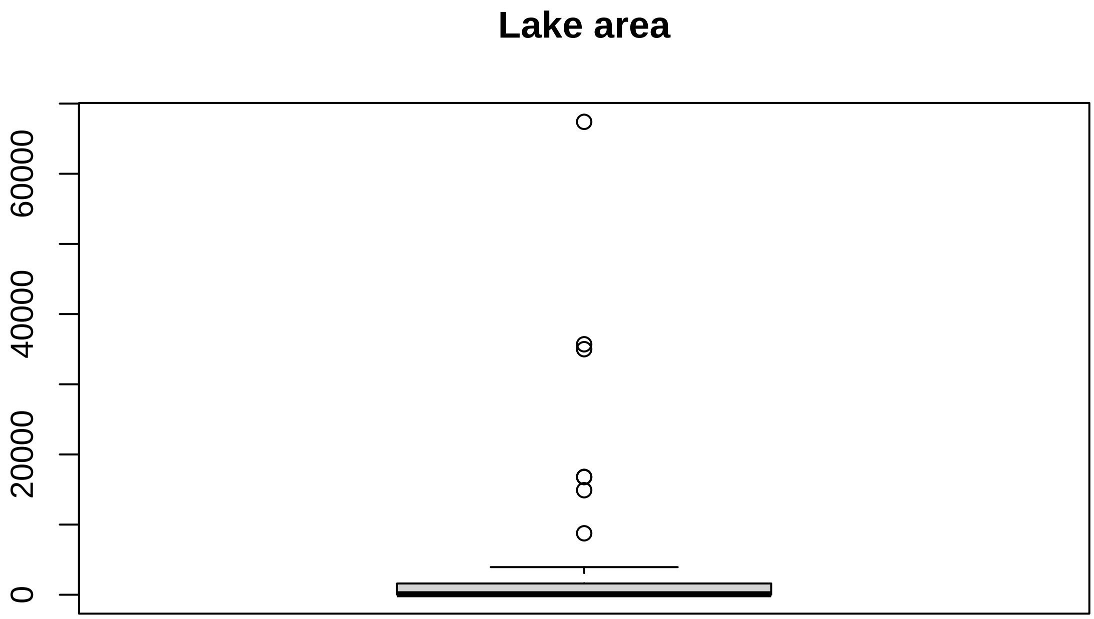
With log-transformed variables, the model fits much better (R^2 = 0.46 vs. 0.24). The slope (0.275 ± 0.053, $p < 0.001$) shows that species richness increases with lake area but at a slowing rate—larger lakes add species more gradually. This supports the idea of diminishing returns in the species–area relationship. Additionally, the Q-Q residual plot shows a better normality indicating a better model fit.

Question 3: Follow the same procedure as above (a-f) using number of species of copepods, rather than fish, as the response variable. What conclusions can you draw from the analysis?

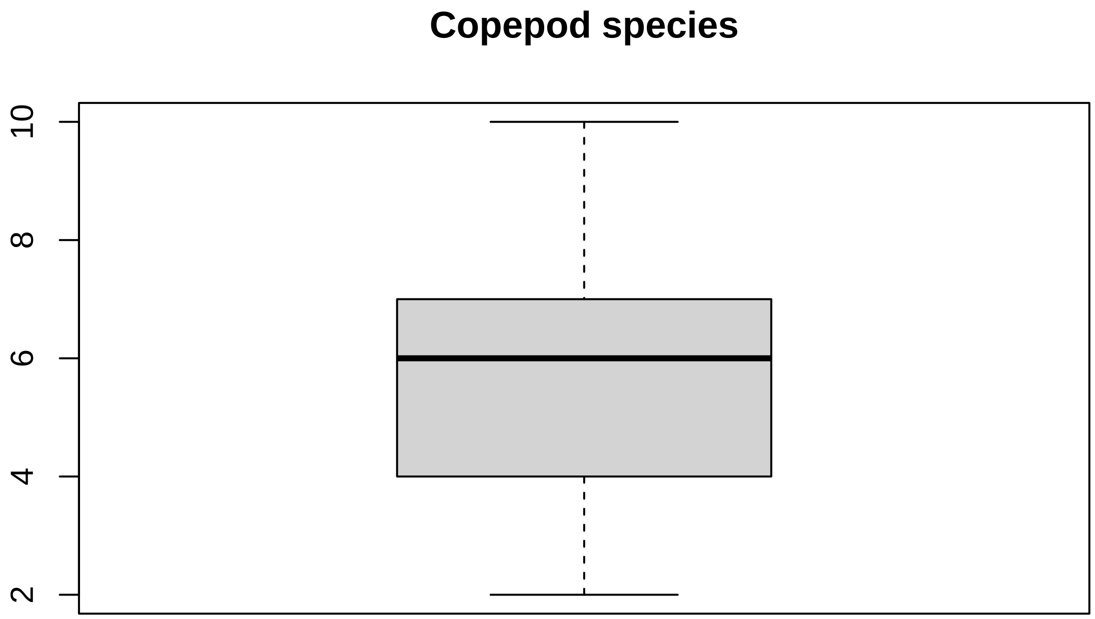
```
# Get vars
dodson.species <- dodson$Copepods
dodson.area <- dodson$Area

# Scatterplot with lowess for nonlinearity
plot(dodson.area, dodson.species,
     xlab = "Lake area",
     ylab = "Number of copepod species")
lines(lowess(dodson.area, dodson.species), col = "red", lwd = 2)
abline(lm(dodson.species ~ dodson.area), col = "blue", lwd = 2)
```



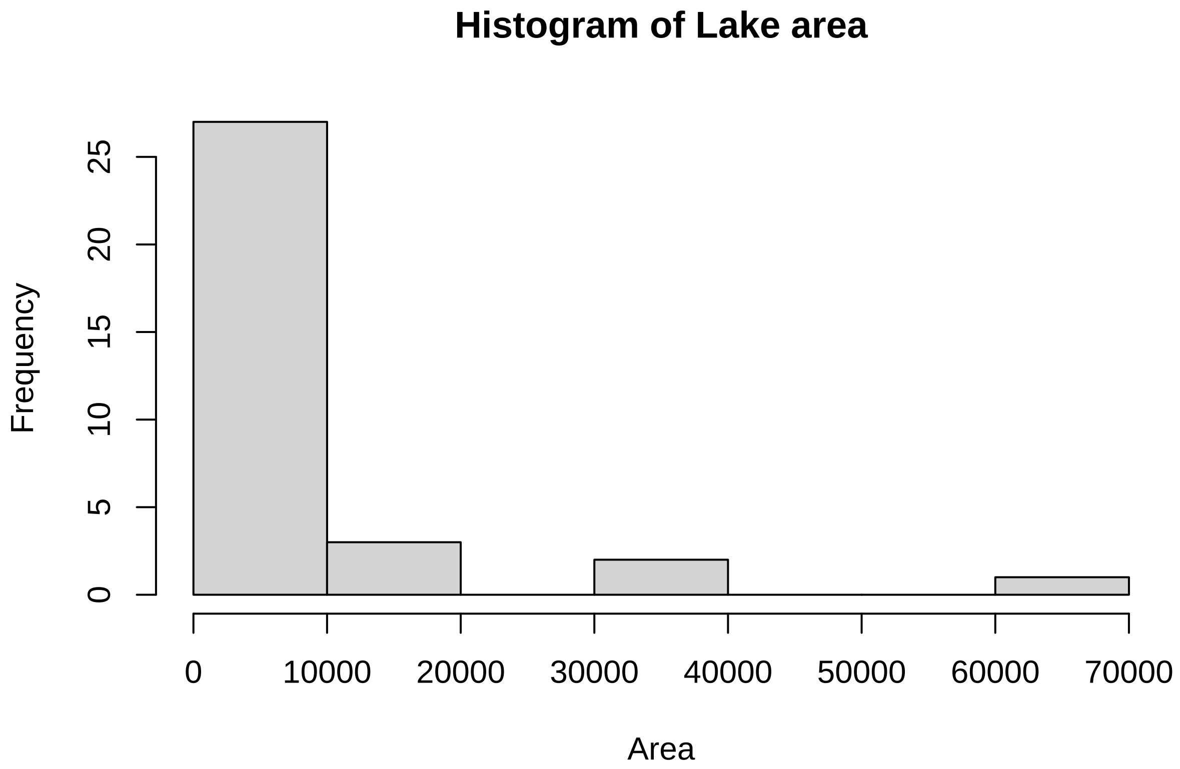
```
# Boxplots
boxplot(dodson.area, main="Lake area")
```



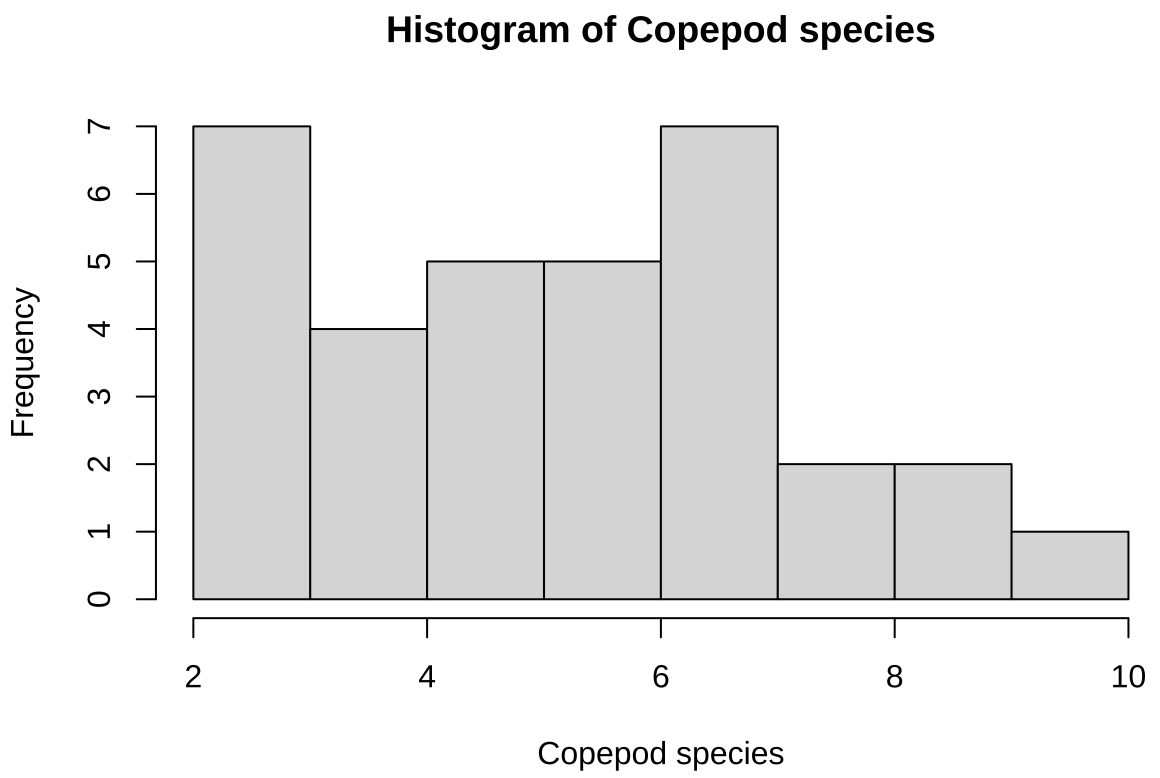
```
boxplot(dodson.species, main="Copepod species")
```



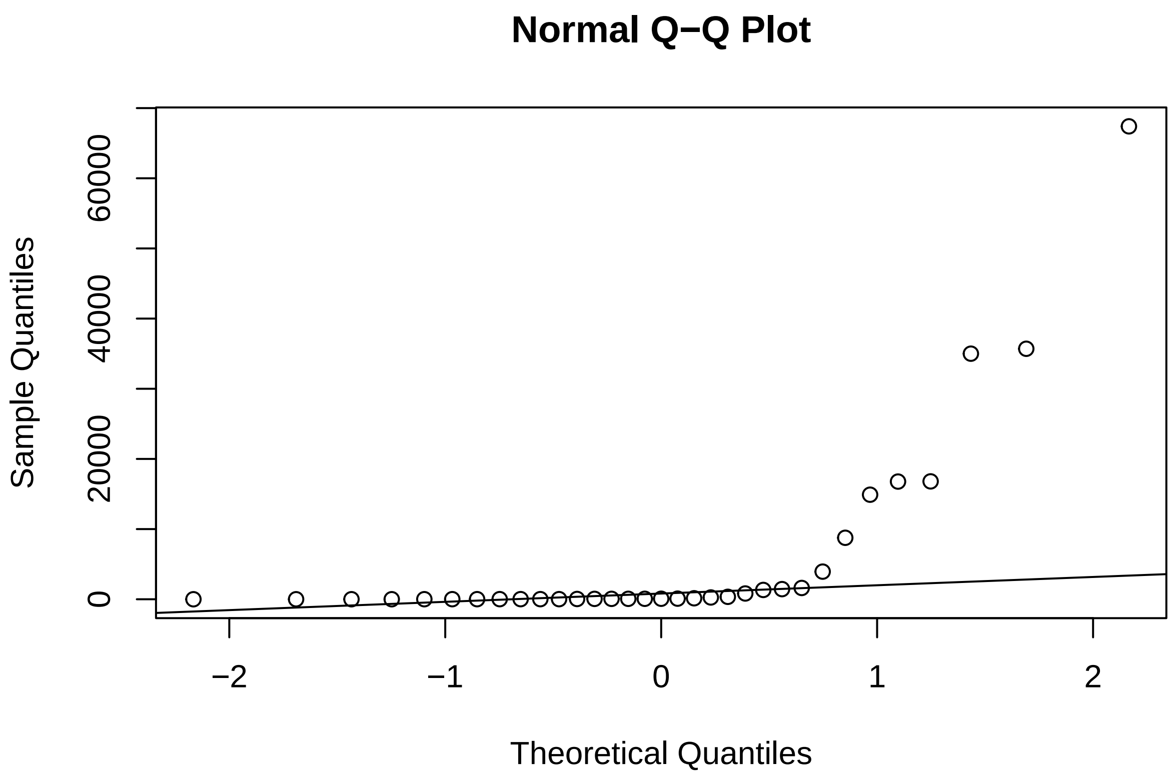
```
# Histograms
hist(dodson.area, main="Histogram of Lake area", xlab="Area")
```



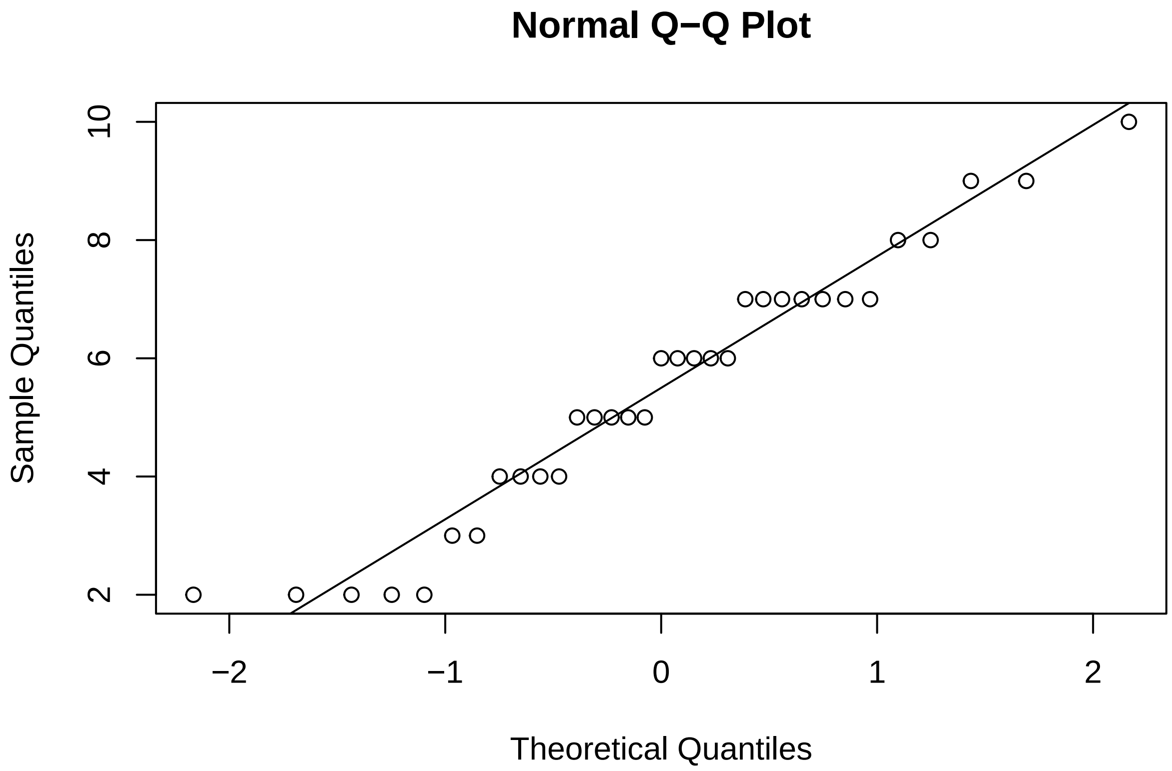
```r
hist(dodson.species, main="Histogram of Copepod species", xlab="Copepod species")
```



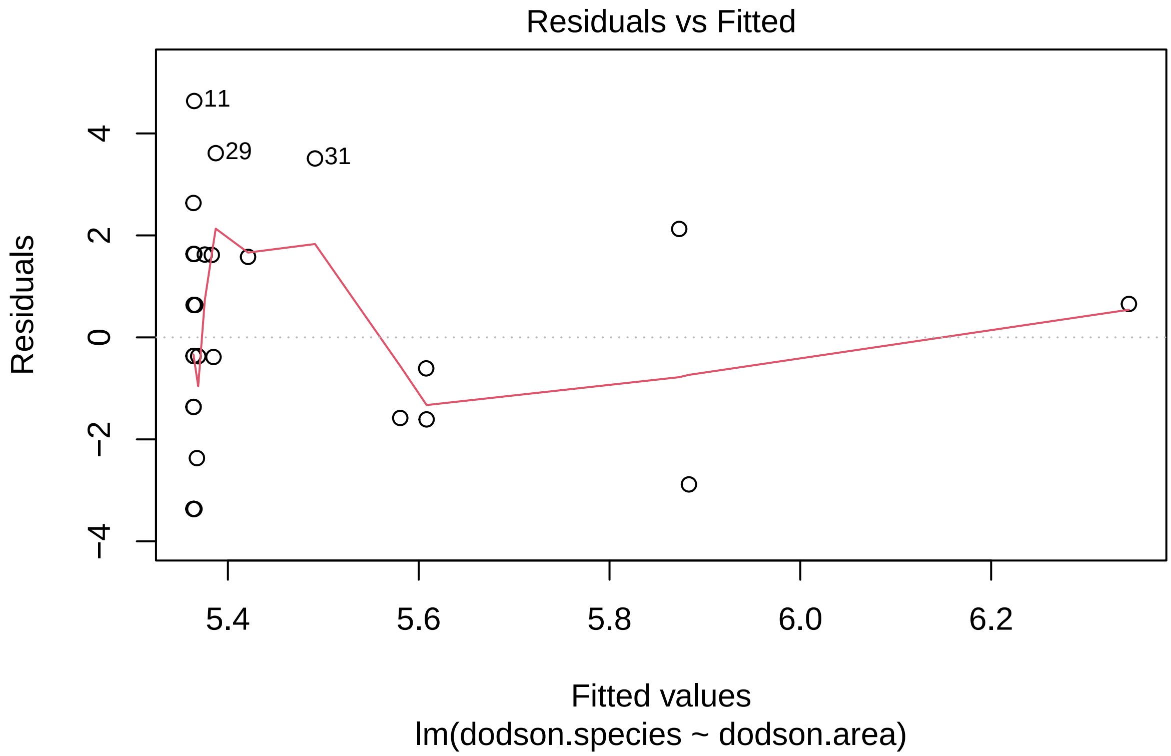
```
# QQ plots
qqnorm(dodson.area); qqline(dodson.area)
```

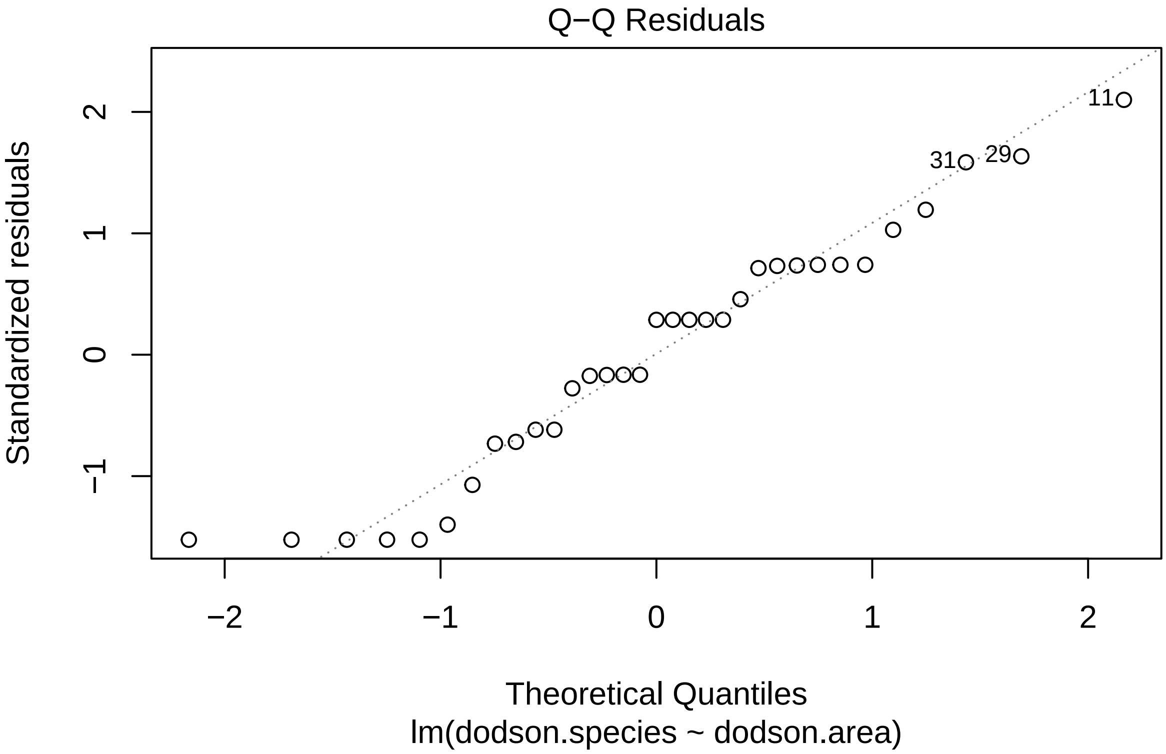


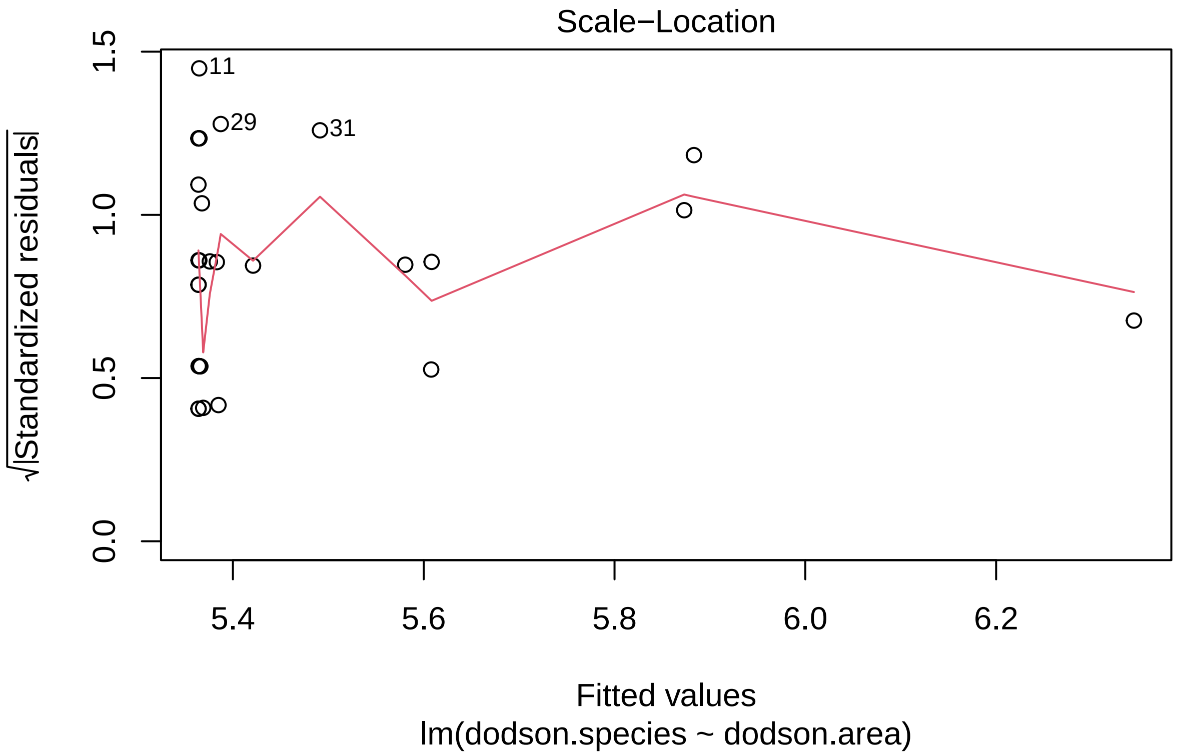
```
qqnorm(dodson.species); qqline(dodson.species)
```

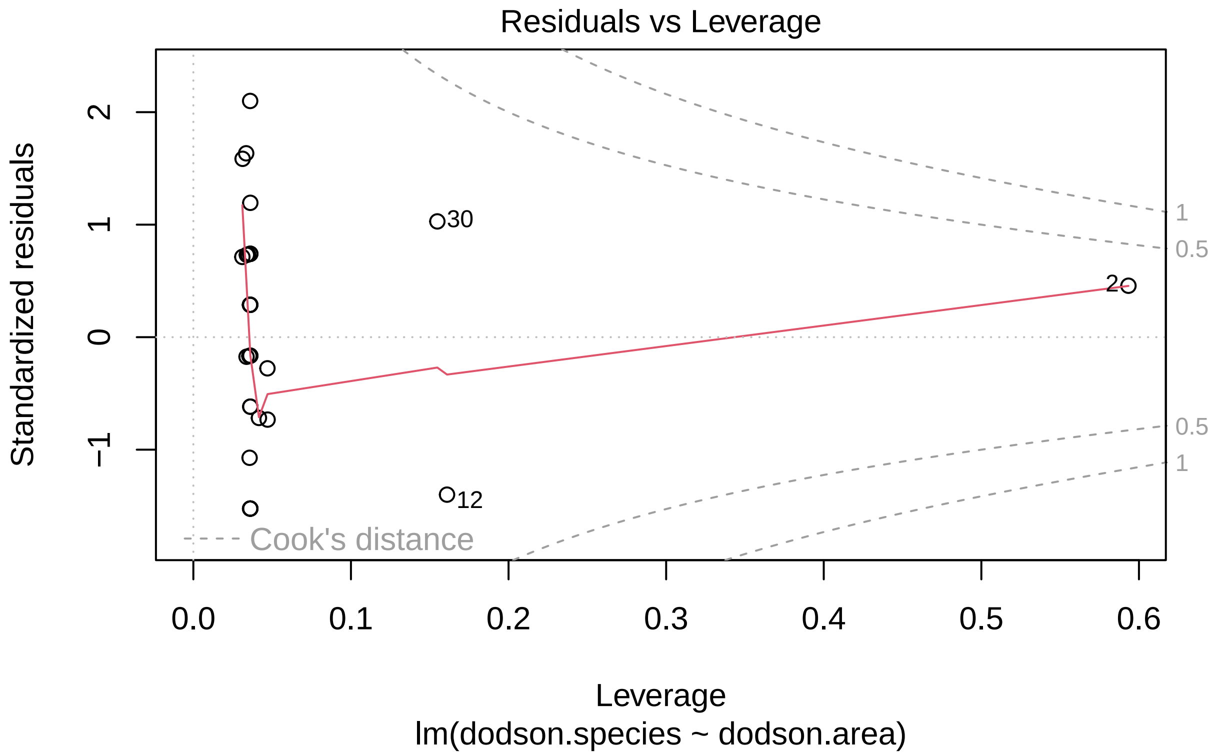


```r
# Make linear model
model <- lm(dodson.species ~ dodson.area)
plot(model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(dodson.species ~ dodson.area)
11
29
31
−4
−2
0
2
4
5.4
5.6
5.8
6.0
6.2


Q−Q Residuals
Standardized residuals
2
1
0
−1
−2
−1
0
1
2
Theoretical Quantiles
lm(dodson.species ~ dodson.area)
11
31
29


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
11
29
31
5.4
5.6
5.8
6.0
6.2
Fitted values
lm(dodson.species ~ dodson.area)

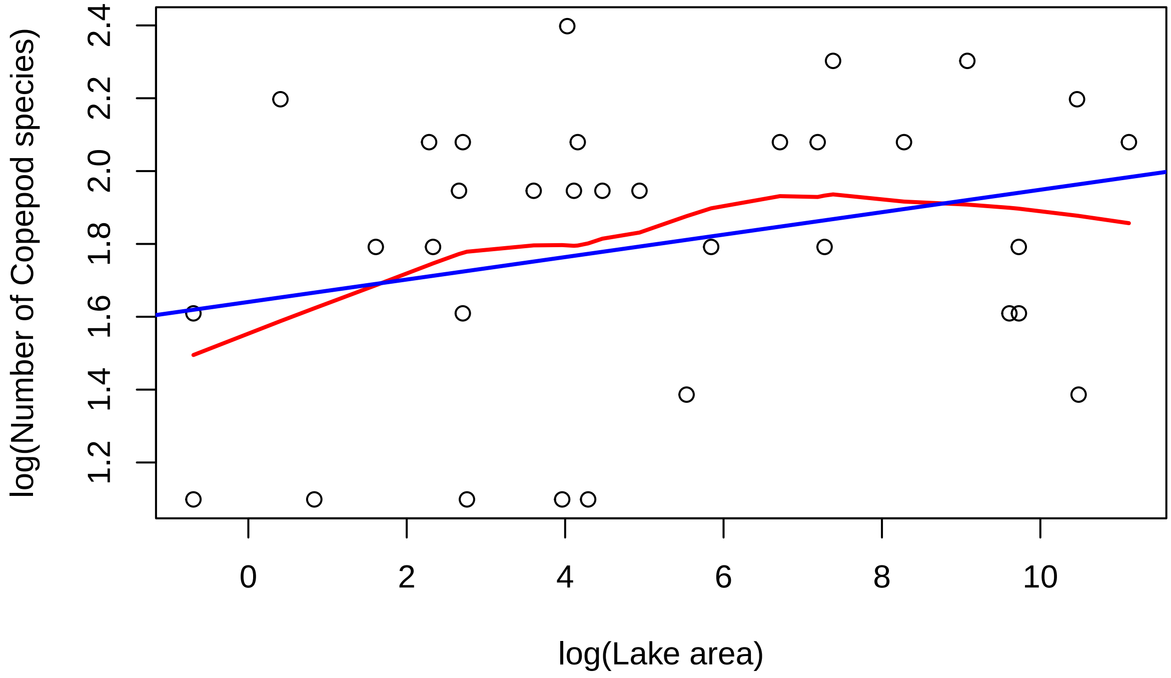


```
summary(model)
```

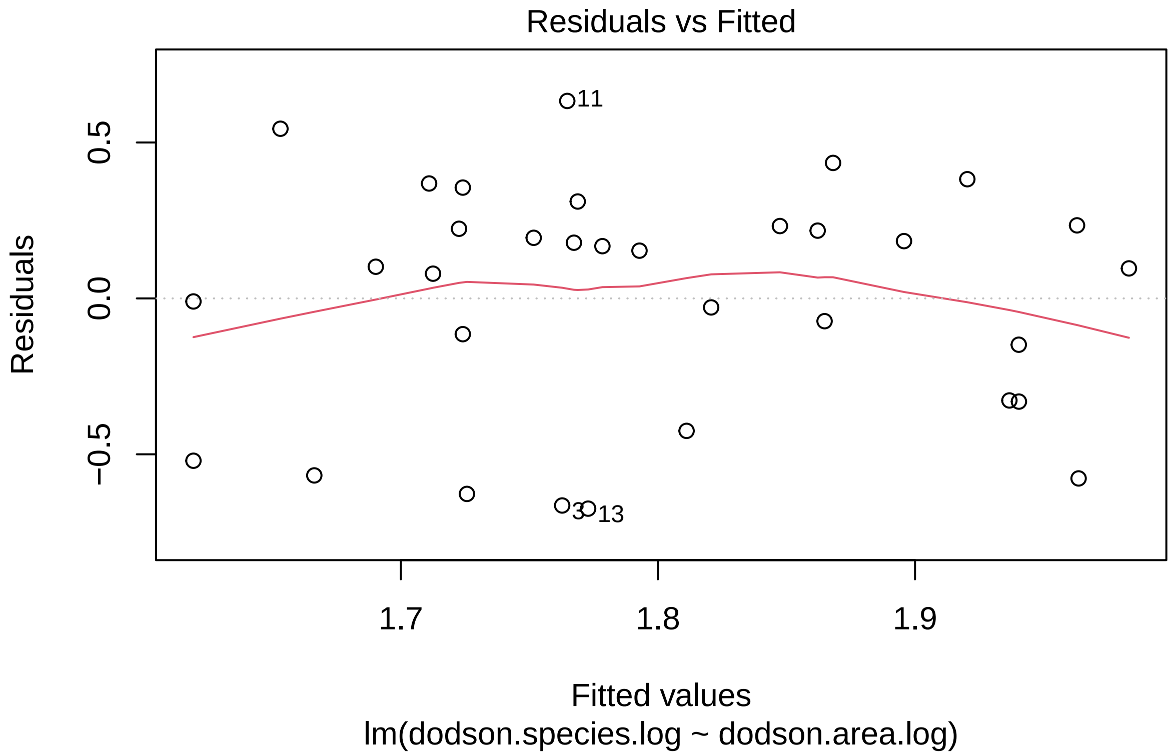
```
##
## Call:
## lm(formula = dodson.species ~ dodson.area)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3649 -1.5806  0.6341  1.6242  4.6353
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.364e+00  4.276e-01  12.545  1.1e-13 ***
## dodson.area 1.455e-05  2.759e-05   0.527    0.602
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.249 on 31 degrees of freedom
## Multiple R-squared:  0.008892,   Adjusted R-squared:  -0.02308
## F-statistic: 0.2781 on 1 and 31 DF,  p-value: 0.6017
```

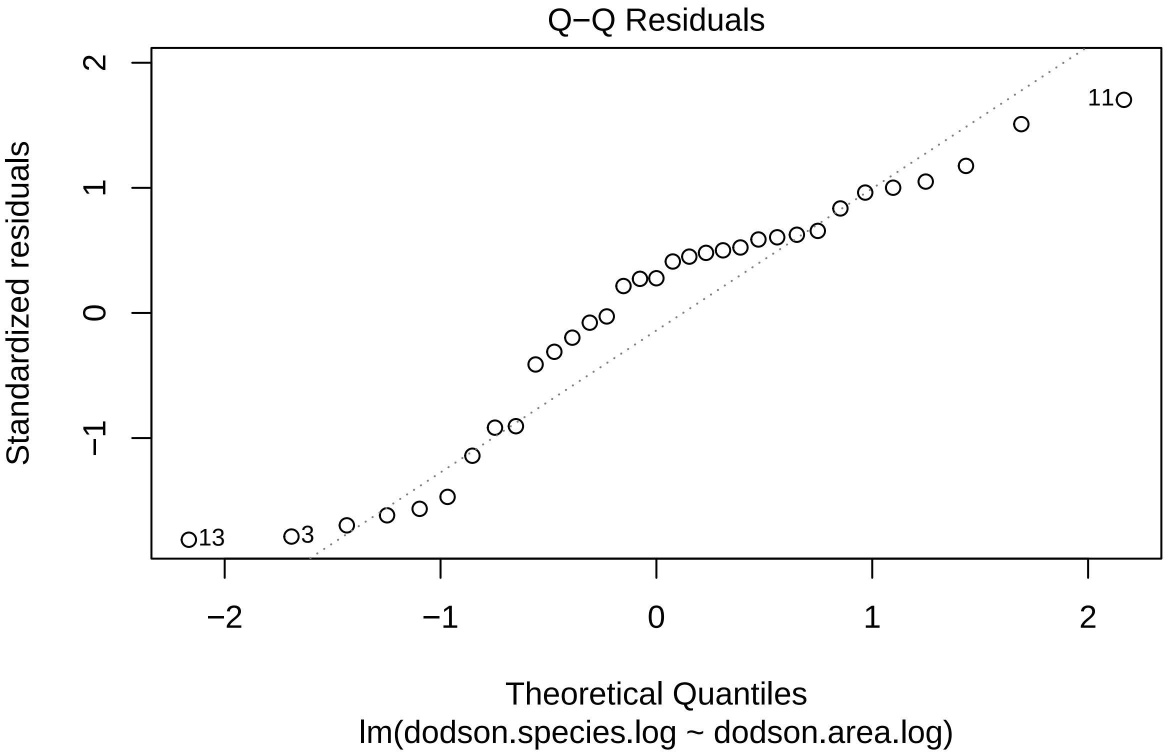
```
# Test log transform
dodson.species.log <- log1p(dodson$Copepods)
dodson.area.log <- log(dodson$Area)
```

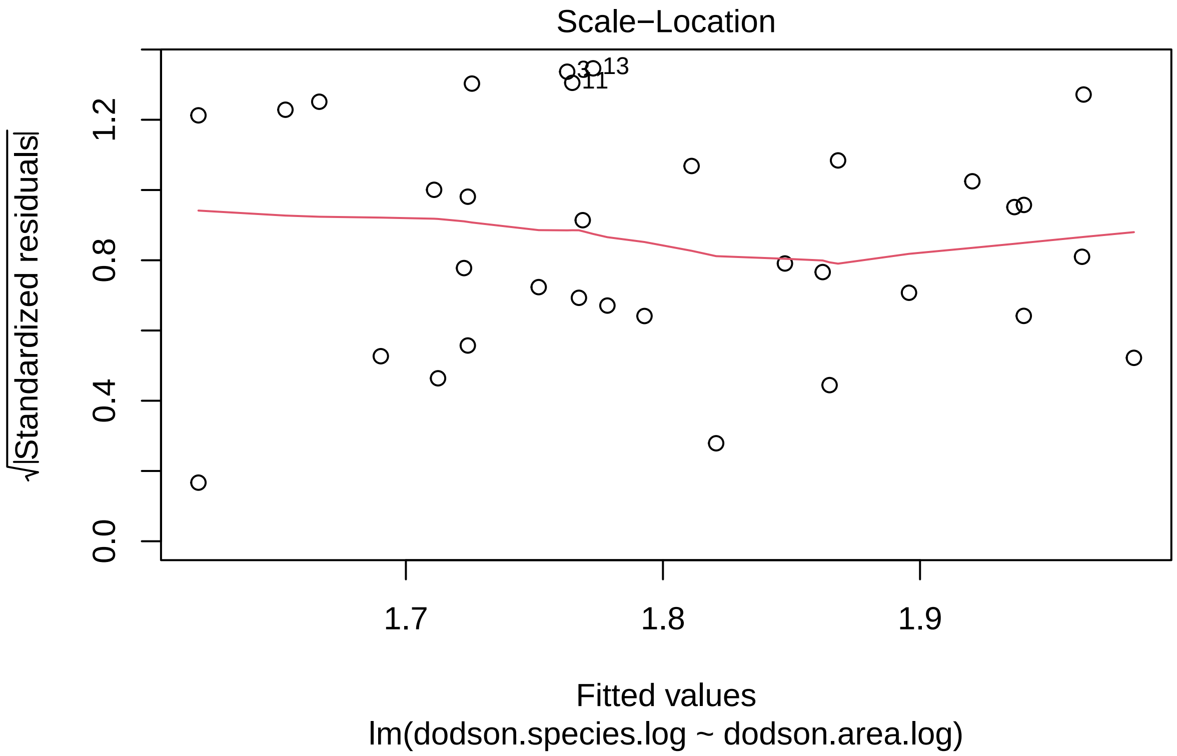
```
plot(dodson.area.log, dodson.species.log,
     xlab = "log(Lake area)",
     ylab = "log(Number of Copepod species)")
lines(lowess(dodson.area.log, dodson.species.log), col="red", lwd=2)
abline(lm(dodson.species.log ~ dodson.area.log), col="blue", lwd=2)
```

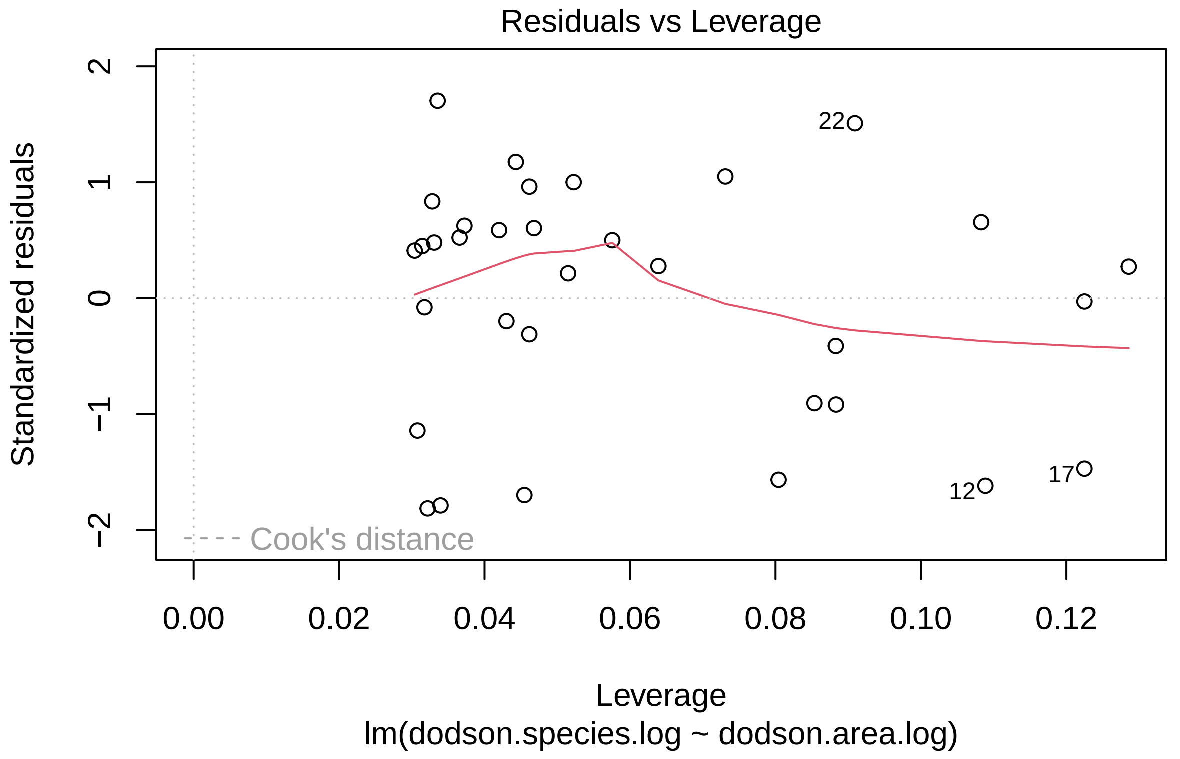


```
# Create model for log data
model.log <- lm(dodson.species.log ~ dodson.area.log )
plot(model.log)
```


Residuals vs Fitted
Residuals
0.5
0.0
−0.5
11
3
13
1.7
1.8
1.9
Fitted values
lm(dodson.species.log ~ dodson.area.log)


Q−Q Residuals
Standardized residuals
2
1
0
−1
−2
−1
0
1
2
Theoretical Quantiles
lm(dodson.species.log ~ dodson.area.log)
13
3
11


Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
1.7
1.8
1.9
3
13
11
Fitted values
lm(dodson.species.log ~ dodson.area.log)



```
summary(model.log)
```

```
##
## Call:
## lm(formula = dodson.species.log ~ dodson.area.log)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6742 -0.3273  0.1015  0.2320  0.6332
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.64066    0.12063   13.60 1.31e-14 ***
## dodson.area.log  0.03081    0.01975    1.56    0.129
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3781 on 31 degrees of freedom
## Multiple R-squared:  0.07277,    Adjusted R-squared:  0.04286
## F-statistic: 2.433 on 1 and 31 DF,  p-value: 0.129
```

In the untransformed regression, copepod species richness shows no significant relationship with lake area. Relationship: slope = 611.2 ± 1158.9, t(31) = 0.53, p = 0.602. Equation: area = 2900.5 + 611.2 * area. Fit: R^2 = 0.009, adjusted R^2 = –0.023, F(1, 31) = 0.28, p = 0.602. This suggests that a simple linear model does not capture the pattern.

However, after log transformation, the relationship becomes strong and highly significant. Relationship: slope = 0.275 ± 0.053, t(31) = 5.16, $p < 0.001$. Equation: log(species) = 1.01 + 0.275 × log(area). Fit: Rˆ2 = 0.462, adjusted Rˆ2 = 0.444, F(1, 31) = 26.59, $p < 0.001$. The slope indicates that copepod richness increases with lake area but at a diminishing rate, indicating a species–area relationship.