Supplement for Lecture 7: Outliers and Influential Points

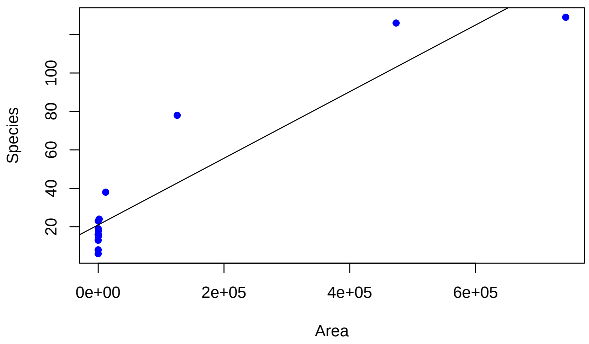
Load Data from Textbook

```
data("SpeciesArea") # Load Data

species = SpeciesArea[,-c(4,5)] # Shorten Name and Remove Last 2 Columns

rm(SpeciesArea) #Removes Old Object from Environment

head(species)
```

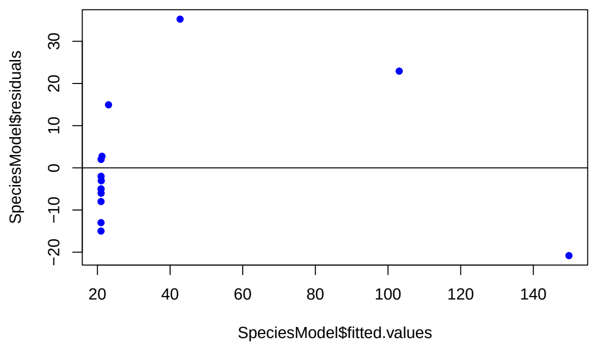
```
##       Name   Area Species
## 1   Borneo 743244     129
## 2  Sumatra 473607     126
## 3     Java 125628      78
## 4   Bangka  11964      38
## 5 Bunguran   1594      24
## 6   Banggi    450      18
```

Linear Model for Species vs Area

```
SpeciesModel=lm(Species~Area, data=species)

plot(Species ~ Area,data=species,pch=16,col="blue")
abline(SpeciesModel)
```

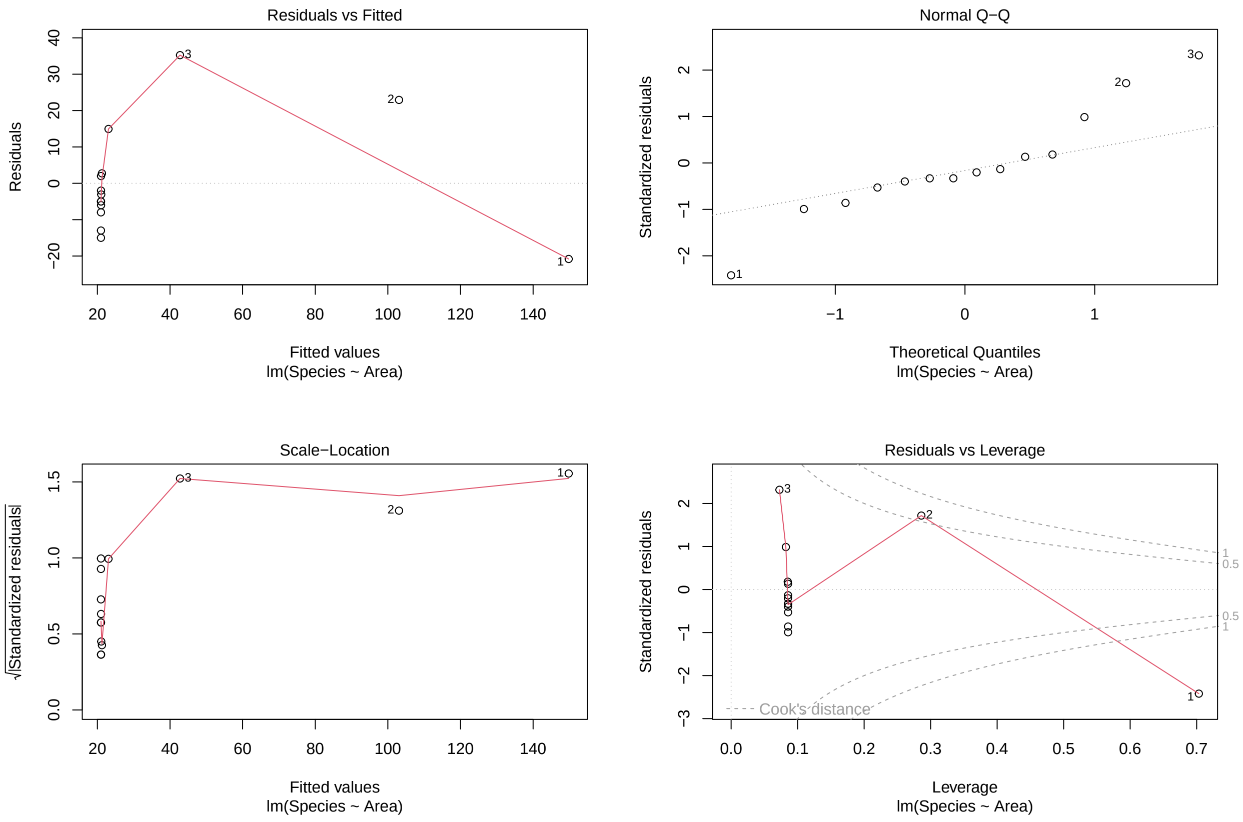


```
plot(SpeciesModel$residuals~SpeciesModel$fitted.values,pch=16,col="blue")
abline(0,0)
```



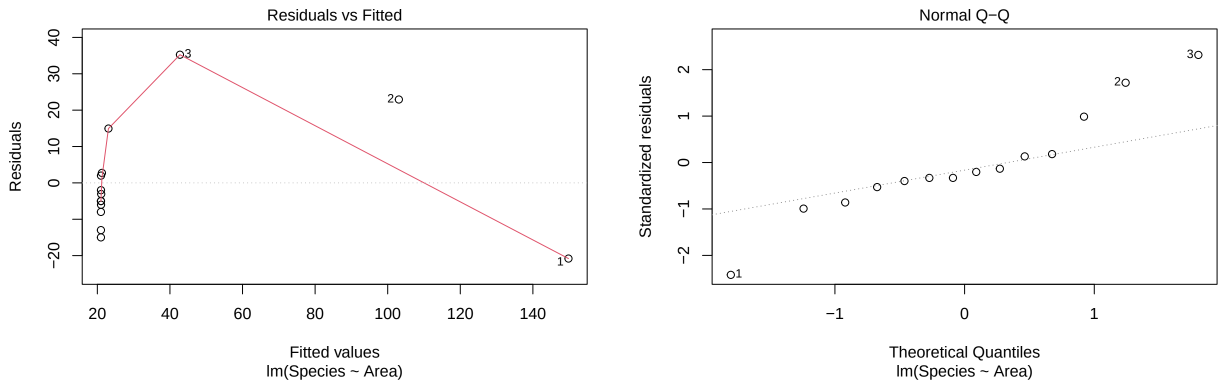
Default Plots from lm Object

```
plot(SpeciesModel) #Default Argument which=c(1,2,3,5)
```



```
#See ?plot.lm (6 Plots Available)

#Plot 1: Residuals vs Fitted (We Did This Already)
#Plot 2: Normal QQ (We Did This Already)
#Plot 3: Square Root of Absolute Standardized Residual vs Fitted (Check for Outliers)
#Plot 4: Not Part of Default
#Plot 5: Residuals vs Leverage (Check for Influence)
```

```r
#Plot 6: Not Part of Default

plot(SpeciesModel, which=c(1,2)) #Only Want First Two
```



Identifying Outliers

```r
#Alternate to SpeciesModel$residuals
residuals(SpeciesModel)
```

```
##          1          2          3          4          5          6          7 
## -20.807685  22.927548  35.241524  14.942506   2.739901  -3.061813  -6.017442 
##          8          9         10         11         12         13         14 
##  -2.006349   1.996424  -5.003402  -4.991789 -12.988150  -7.986070 -14.985203
```

```r
max(residuals(SpeciesModel)) #Largest
```

```
## [1] 35.24152
```

```r
min(residuals(SpeciesModel)) #Smallest
```

```
## [1] -20.80769
```

```r
max(abs(residuals(SpeciesModel))) #Farthest Away from 0
```

```
## [1] 35.24152
```

```r
which.max(abs(residuals(SpeciesModel))) #Returns Index of the Max
```

```
## 3 
## 3
```

```r
#Bad Idea: Remove Data Point with Largest Outlier
species.without.3 = species
species.without.3
```

```
##                  Name   Area Species
## 1              Borneo 743244     129
## 2             Sumatra 473607     126
## 3                Java 125628      78
## 4              Bangka  11964      38
## 5            Bunguran   1594      24
## 6              Banggi    450      18
## 7             Jemaja    194      15
## 8    Karimata Besar    130      19
## 9             Tioman    114      23
## 10           Siantan    113      16
## 11         Sirhassan     46      16
## 12            Redang     25       8
## 13        Penebangan     13      13
## 14 Perhentian Besar      8       6
```

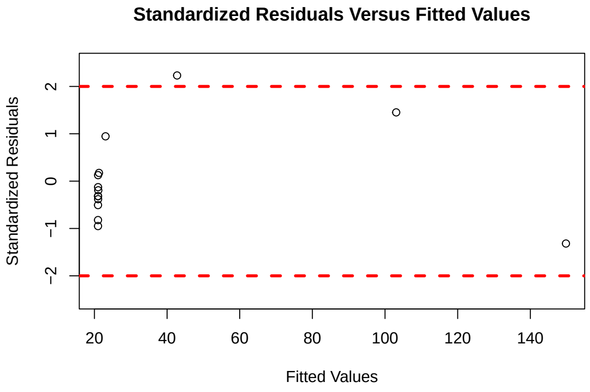
```
#Standardizing Residuals
SSE=sum(residuals(SpeciesModel)^2)
n=length(residuals(SpeciesModel))
std.error.regression = sqrt(SSE/(n-2))
std.error.regression #Expect to be Off By 15 Species On Average When Using Model to Predict
```

```
## [1] 15.79018
```

```
summary(SpeciesModel)
```

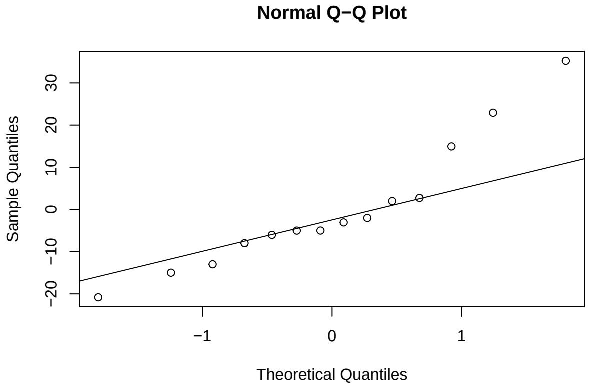
```
##
## Call:
## lm(formula = Species ~ Area, data = species)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.808  -7.494  -4.027   2.554  35.242
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.098e+01  4.621e+00   4.541 0.000677 ***
## Area        1.733e-04  1.942e-05   8.925 1.21e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.79 on 12 degrees of freedom
## Multiple R-squared:  0.8691, Adjusted R-squared:  0.8582
## F-statistic: 79.66 on 1 and 12 DF,  p-value: 1.206e-06
```

```
#Plot of Standardized Residuals
std.res=residuals(SpeciesModel)/std.error.regression
plot(y=std.res,x=fitted(SpeciesModel),
     main="Standardized Residuals Versus Fitted Values",
     xlab="Fitted Values", ylab="Standardized Residuals",
     ylim=c(-2.5,2.5))
abline(h=c(-2,2),
       col="red",
       lty=c(2,2),
       lwd=c(3,3))
```

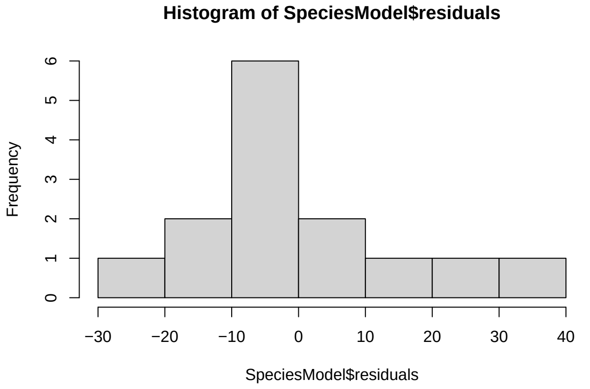


Assess Conditions of Residuals

```
qqnorm(SpeciesModel$residuals)
qqline(SpeciesModel$residuals)
```



```
hist(SpeciesModel$residuals)
```



Transformation

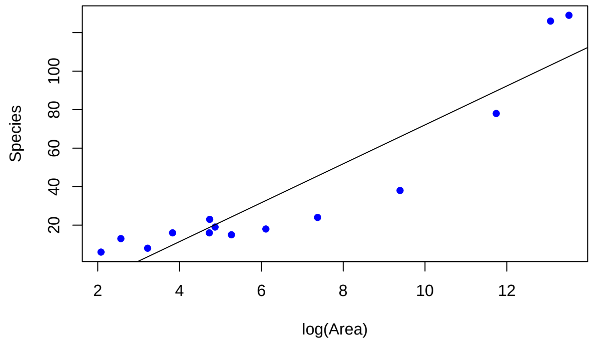
Three Possible Options Using Natural Logarithm Transformation Only

- $y = \beta_0 + \beta_1 log(x) + \epsilon$
- $log(y) = \beta_0 + \beta_1 x + \epsilon$
- $log(y) = \beta_0 + \beta_1 log(x) + \epsilon$

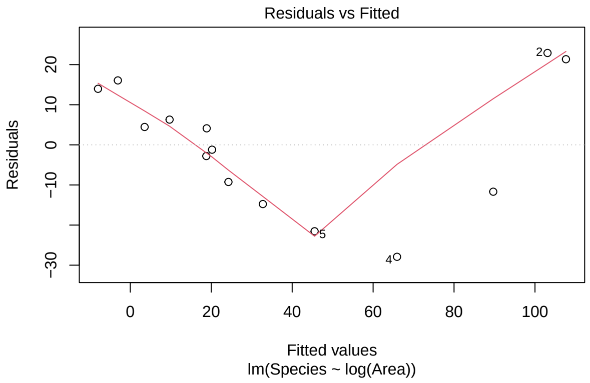
```
# Run the Code and Try All Three To Choose Best
TransMod=lm(Species~log(Area), data=species)
summary(TransMod)
```

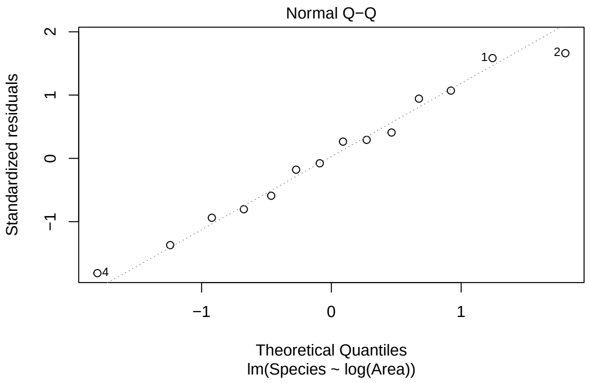
```
##
## Call:
## lm(formula = Species ~ log(Area), data = species)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.916 -11.075   1.455  12.052  22.905
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -28.989      8.923  -3.249  0.00697 **
## log(Area)     10.107      1.178   8.580 1.82e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.34 on 12 degrees of freedom
## Multiple R-squared:  0.8599, Adjusted R-squared:  0.8482
## F-statistic: 73.62 on 1 and 12 DF,  p-value: 1.822e-06
```

```
plot(Species ~ log(Area),data=species,pch=16,col="blue")
abline(TransMod)
```



```
#abline(v=c(5,12),col=c("red","red"),lwd=c(3,3),lty=c(3,3)) #Run for First Option

plot(TransMod,c(1,2))
```





Getting Fitted Values, Residuals, and Predictions

Using the model we fit to make predictions:

$$\widehat{\log(y)} = \hat{\beta}_0 + \hat{\beta}_1 \log(x)$$

Transforming predictions back to the units of Y = Number of Species

$$\hat{y} = e^{\widehat{\log(y)}} = e^{\hat{\beta}_0 + \hat{\beta}_1 log(x)} = e^{\hat{\beta}_0} x^{\hat{\beta}_1}$$

If we run `coef(TransMod)`, we find our estimates

$$\hat{y} = e^{1.6249 + 0.2355 \log(x)} = e^{1.6249} x^{0.2355}$$

```
#Recall Property of Logs
10^log10(5)
```

```
## [1] 5
```

```
exp(log(7))
```

```
## [1] 7
```

```
#Extract Coefficients
as.numeric(coef(TransMod))
```

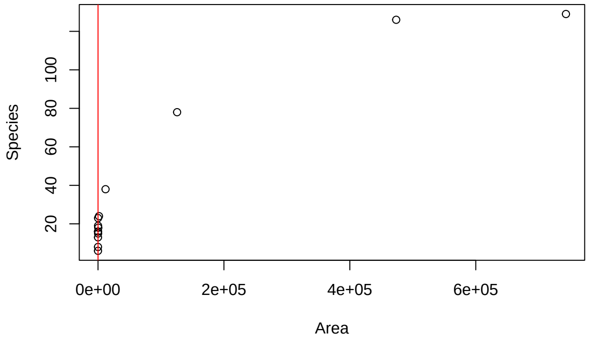
```
## [1] -28.98914  10.10736
```

```
beta0=as.numeric(coef(TransMod))[1]
beta1=as.numeric(coef(TransMod))[2]

#Raw Predictions of Log(Species)
species$logfit = fitted(TransMod)

#3 Methods for Predictions of Actual Species
species$fit1 = exp(species$logfit)
species$fit2= exp(beta0+beta1*log(species$Area))
species$fit3 = exp(beta0)*species$Area^beta1

#Obtain Residuals in Original Units of Species Variable to Calculate SE of Regression
```

```
species$res = species$Species - species$fit1
sqrt(sum(species$res^2)/(length(species$res)-2)) #Standard Error of Regression
```

```
## [1] 1.630266e+46
```

```
#Plot Fitted Model on Raw Untransformed Data
plot(Species~Area, data=species)

#Extrapolate for the Island Australia (7,700,000 km^2)
plot(Species~Area, data=species)
curve(exp(beta0+beta1*log(x)), add=TRUE, col="red")
```


Species
0e+00
2e+05
4e+05
6e+05
Area