Final Assignment

Chao Zhang & Ibrahim Kanj
Group 13

Exercise Galapagos

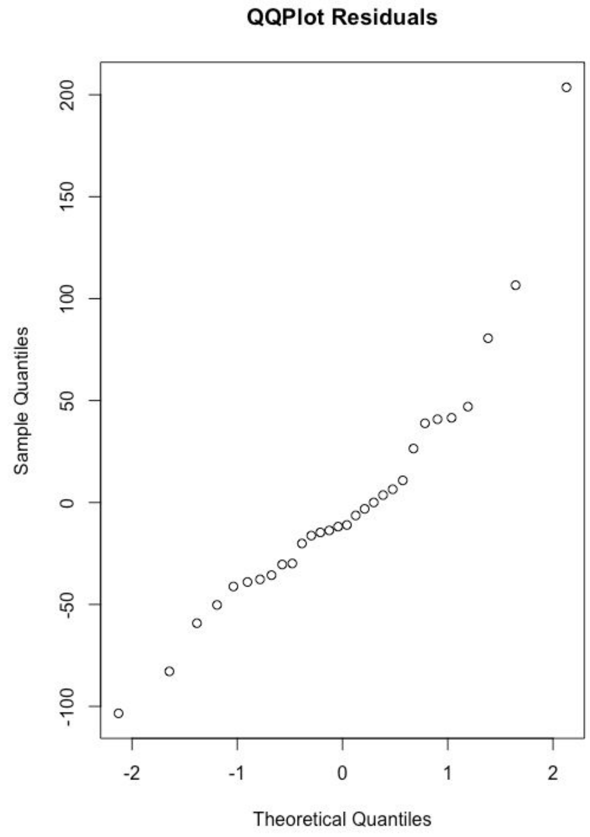
1)

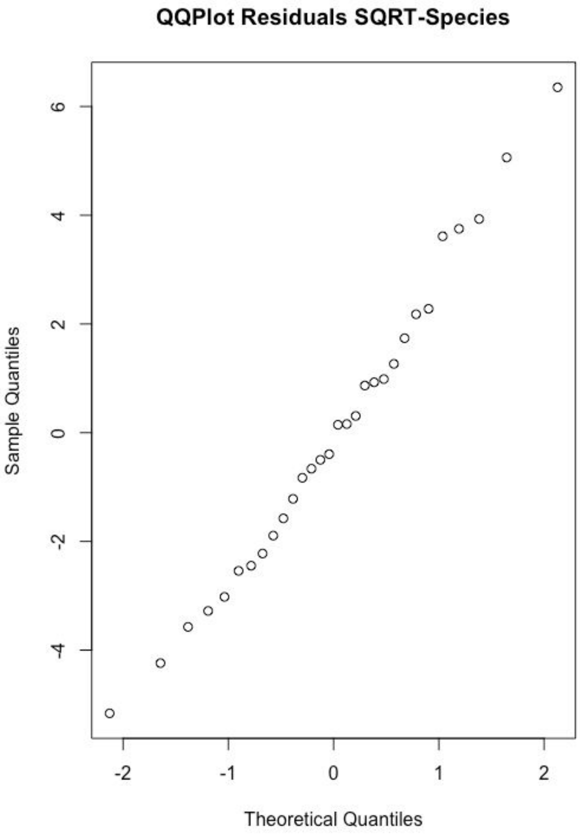
According to the step-down method by removing Nearest, Area and then Scruz we have a linear model Species = 1.43287 + 0.2767*Elevation - 0.06889*Adjacent

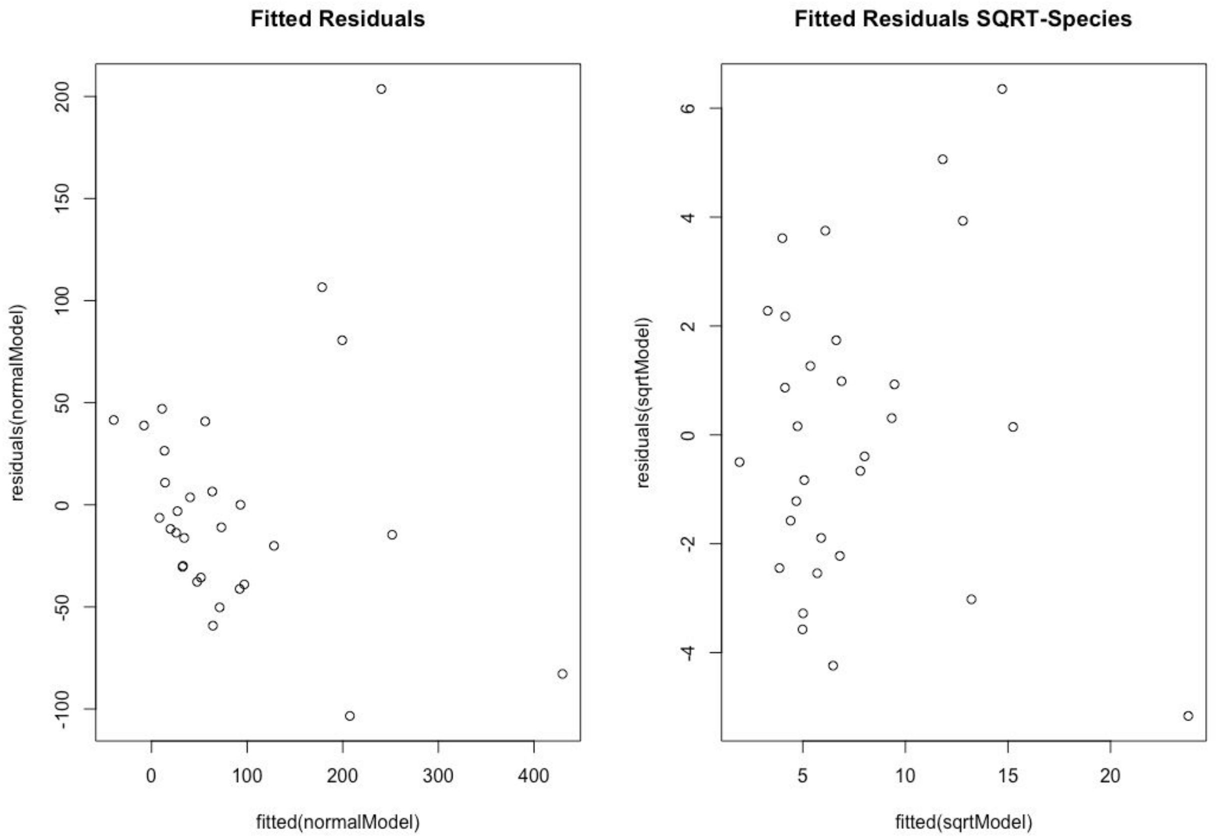
2)

#according to the step down method by removing Nearest, Scruz and then Area we have a linear model sqrt(Species) = 3.5378950 + 0.0129421*Elevation - 0.0028992*Adjacent

3)


QQPlot Residuals


QQPlot Residuals SQRT-Species

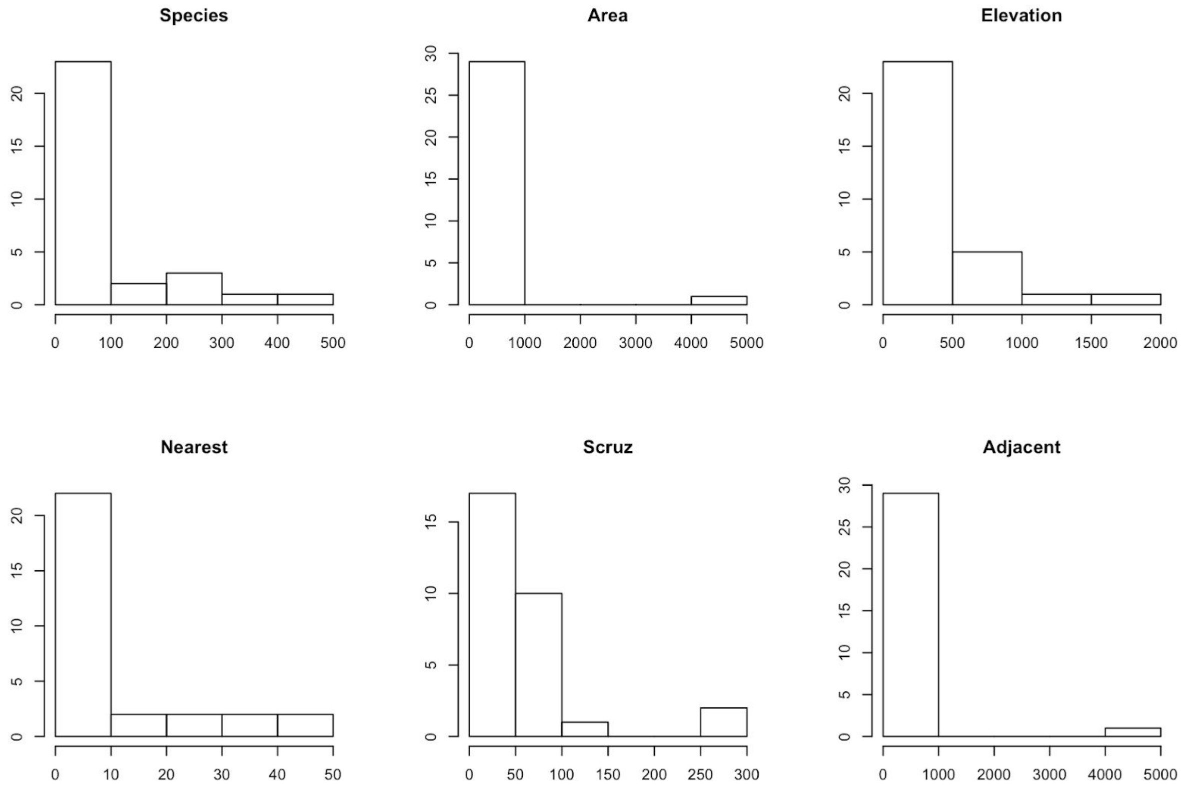


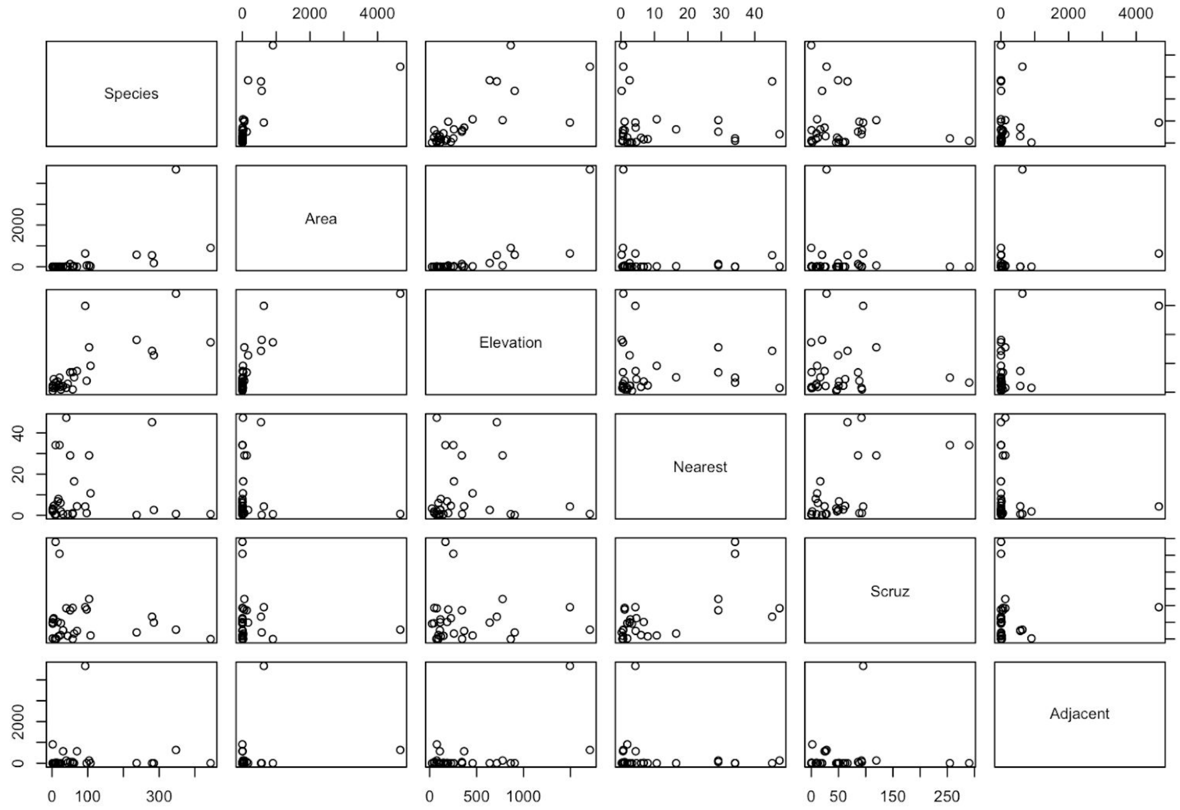
From the plots one could see less outliers in the fitted Residual plot for the model based on square root species. Also the QQplot for residuals based on square root of species is normally distributed while the other QQplot is not normally distributed.
Since there is not much of a difference between multiple R^2 ,thus the resulting linear model 2 is better.

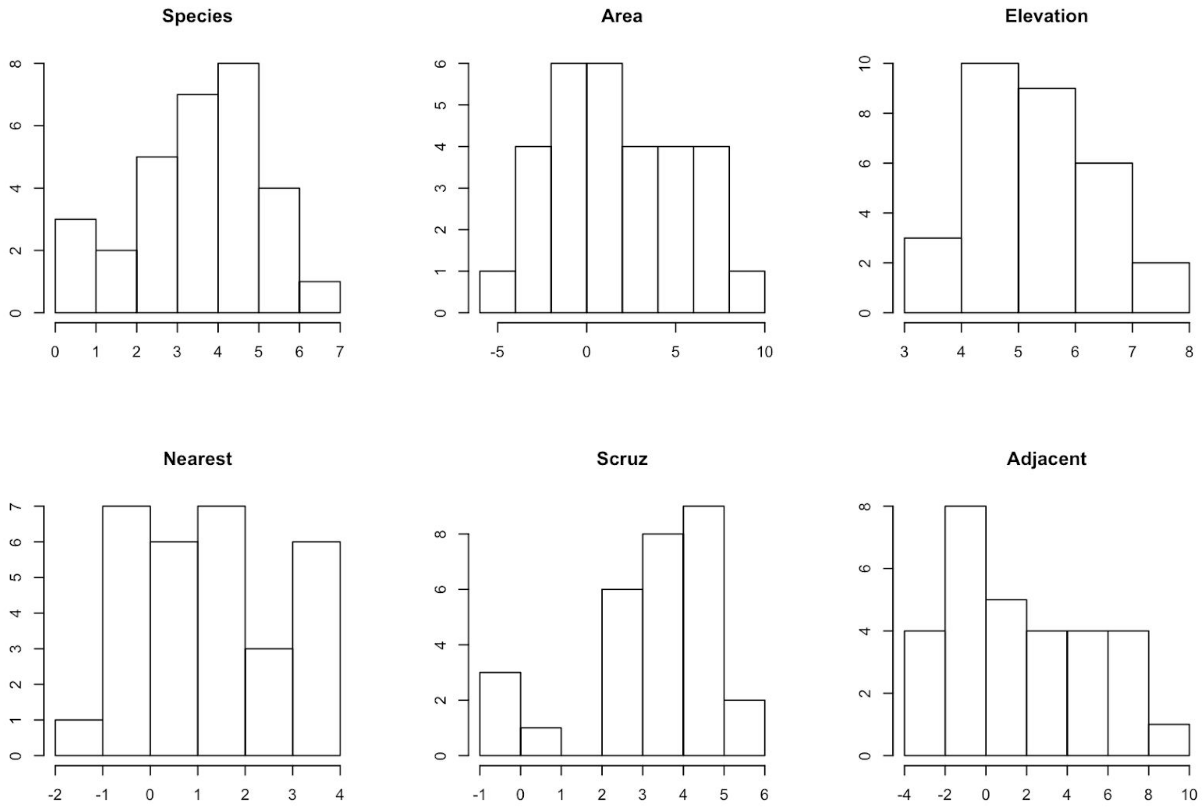
4)

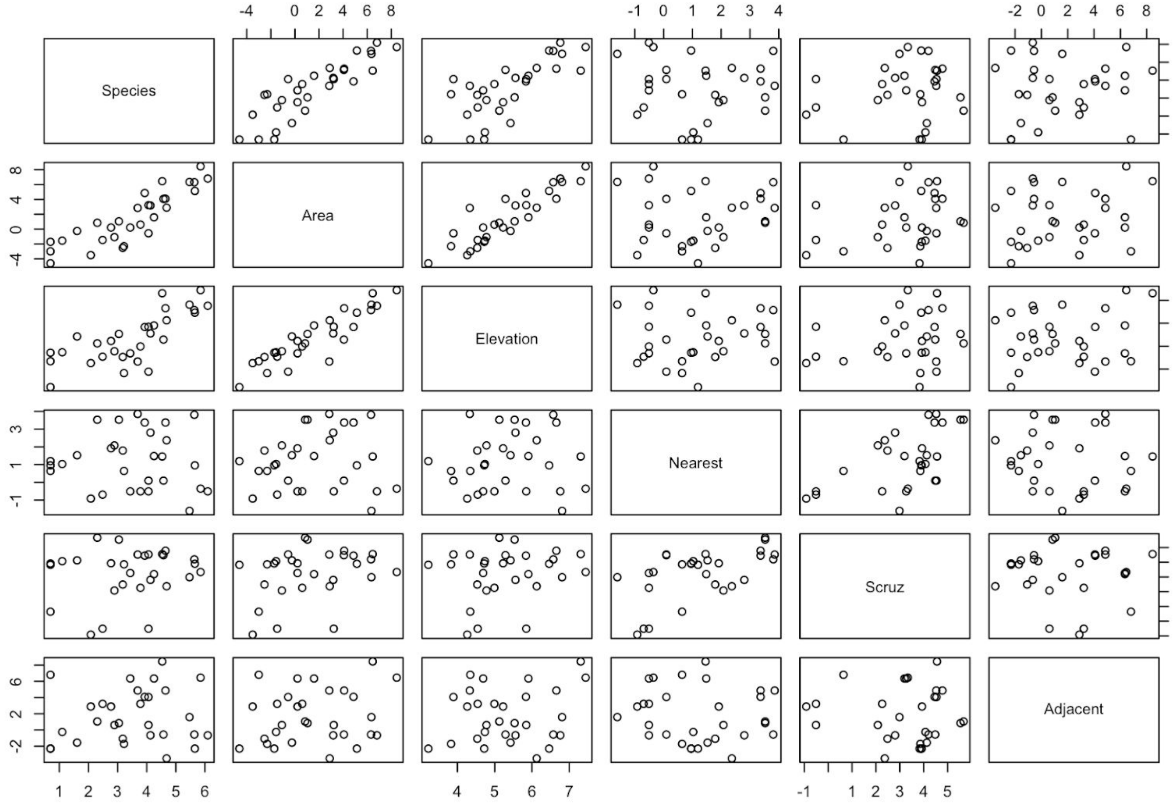
Island 16 - Isabela is an influence point (where cook's distance = 1.64 > 1)
If we remove this island, the model changes into 3.5606398 + 0.0084362* area + 0.0107337*Elevation - 0.0031392*Adjacent since by step-down method Area wouldn't be dropped.

5)


Species
0
5
10
15
20
0
100
200
300
400
500
Area
0
5
10
15
20
25
30
0
1000
2000
3000
4000
5000
Elevation
0
5
10
15
20
0
500
1000
1500
2000
Nearest
0
5
10
15
20
0
10
20
30
40
50
Scruz
0
5
10
15
0
50
100
150
200
250
300
Adjacent
0
5
10
15
20
25
30
0
1000
2000
3000
4000
5000


Species
Area
Elevation
Nearest
Scruz
Adjacent
0
2000
4000
0
10
20
30
40
0
2000
4000
0
200
400
0
2000
0
1000
0
20
40
0
100
250
0
2000
0
100
300
0
500
1000
0
50
150
250


Species
0
2
4
6
8
0 1 2 3 4 5 6 7
Area
0
1
2
3
4
5
6
-5 0 5 10
Elevation
0
2
4
6
8
10
3 4 5 6 7 8
Nearest
0
1
2
3
4
5
6
7
-2 -1 0 1 2 3 4
Scruz
0
2
4
6
8
-1 0 1 2 3 4 5 6
Adjacent
0
2
4
6
8
-4 -2 0 2 4 6 8 10


Species
Area
Elevation
Nearest
Scruz
Adjacent
-4 0 2 4 6 8
-1 0 1 2 3 4
-2 0 2 4 6 8
1 3 5
-4 0 4 8
4 5 6 7
-1 1 3
-1 1 3 5
-2 2 6
1 2 3 4 5 6
4 5 6 7
-1 1 2 3 4 5

Comparing the first two pictures and last pictures where the logarithm function was applied, We could see that the variables' plots look more normally distributed. Thus a better linear regression model could be achieved. Looking at the different pairs after applying the log function, we could see some collinearity between the variables and thus these variables would probably be dropped.

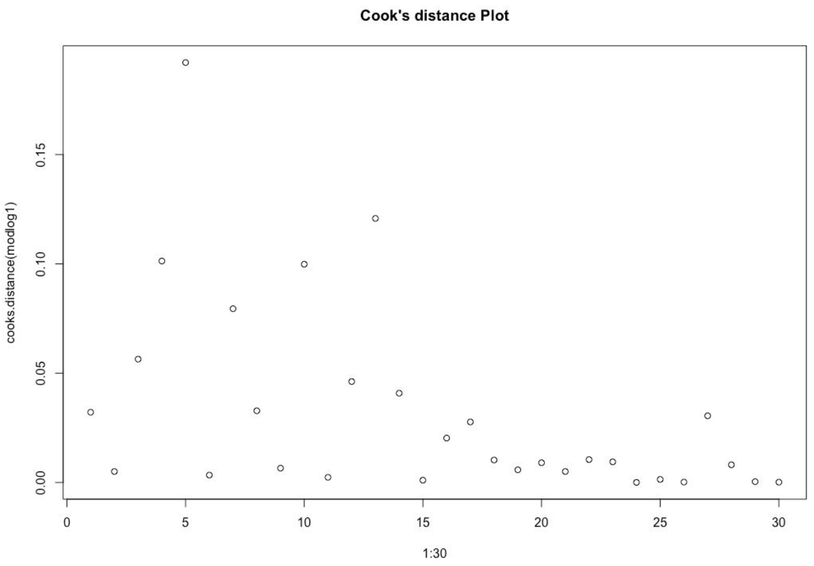
6)

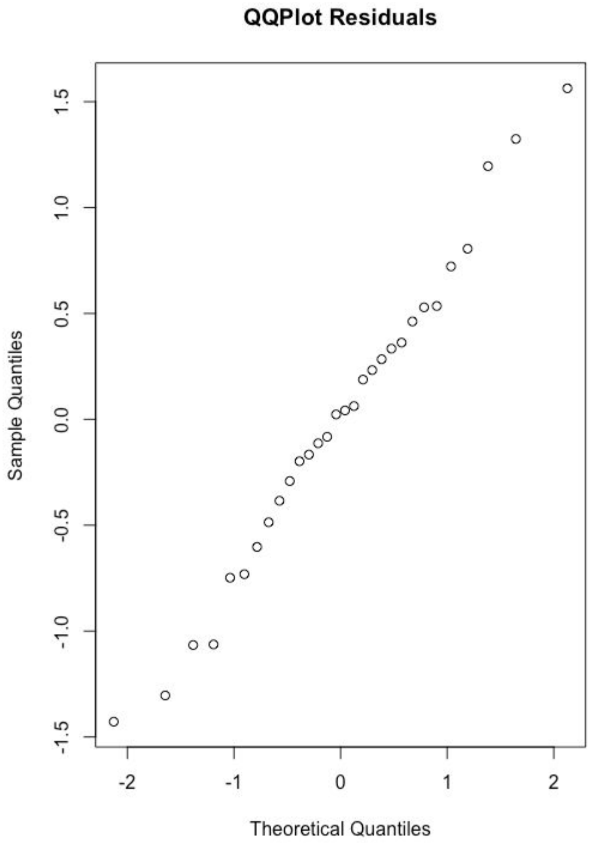
If we look at summary modlog1, we see that the p-value of Scruz + 1 is still > 0.05. Therefore this variable isn't significant. Modlog1 didn't fully complete the step down method and thus Scruz + 1 needs to be removed if the model need to be finished as by the step down method.

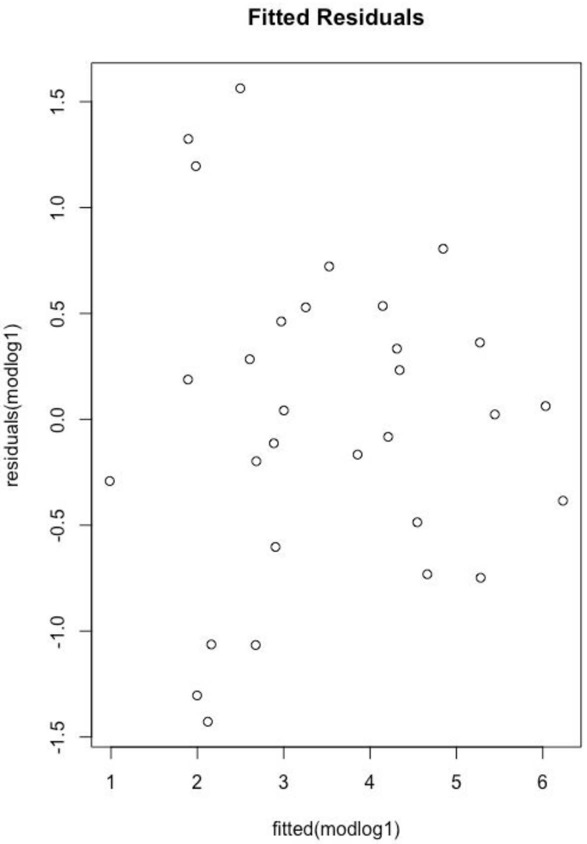
7)

Since there could be a lot of plant species then species is a large number. Similarly Area is a large number thus it makes sense to use a logarithm function in a way to reduce the numbers.

8)

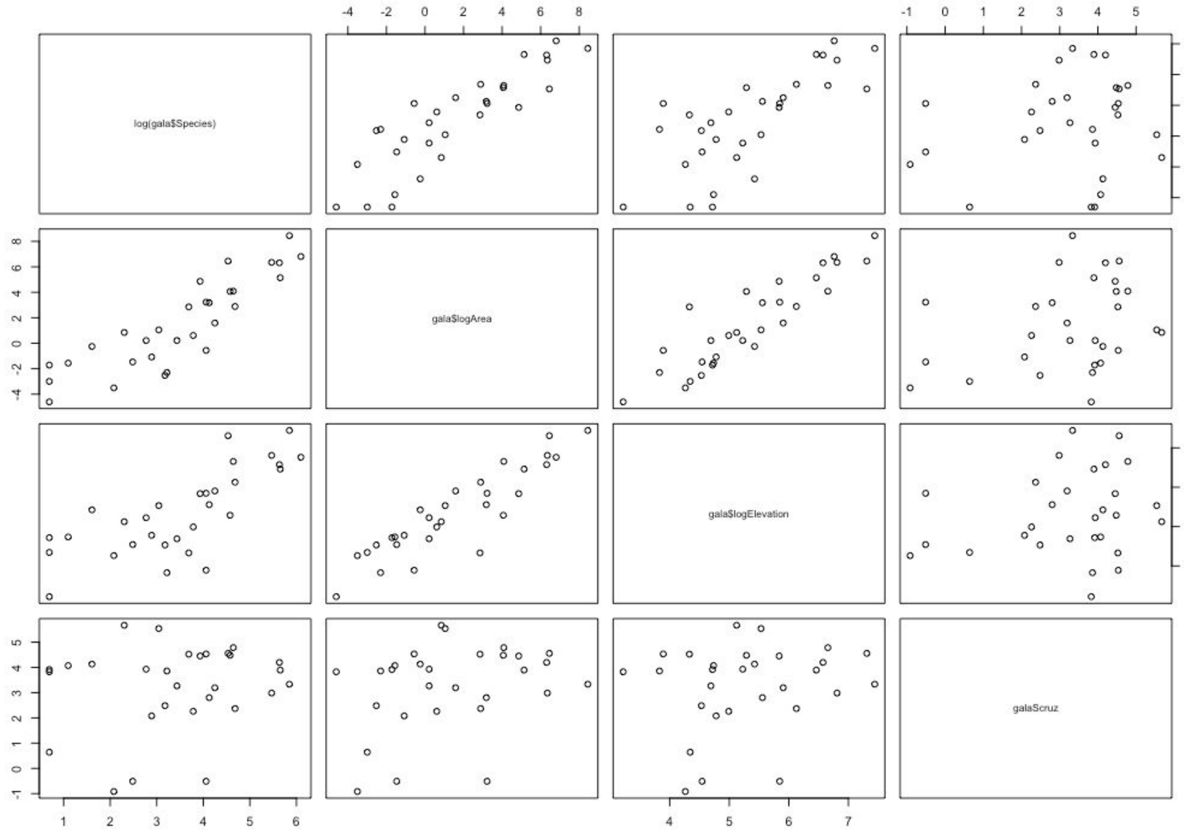






According to Cook's distance there are no influence points. The plot of the residuals shows that the data is normally distributed and from the residuals vs. fitted plot, we can see that there are no outliers and the range is small [-1.5,1.7].

9)



There is a high collinearity between logElevation and logArea (0.9 using numerical pairwise correlation) which means that this model would be better by dropping one of them. Showing interaction between logArea and logElevation would not make sense as they have the same effect on logSpecies.

10)

We prefer modlog1 model, although it has one extra explanatory variable Scruz + 1 but it actually has a better multiple R^2 than the other models and it has no influence points and the residuals are normally distributed as well as a normal fitted residual plot.

Appendix

Full R Code

```
remove(list = ls())
par(mfrow=c(1,1))

#Exercise 1)
gala = read.table("gala.txt",header=TRUE)

#1)
summary(lm(gala$Species~gala$Area+gala$Elevation+gala$Nearest+gala$Scruz+gala$Adjacent))
#removeNearest
summary(lm(gala$Species~gala$Area+gala$Elevation+gala$Scruz+gala$Adjacent))
#removeArea
summary(lm(gala$Species~gala$Elevation+gala$Scruz+gala$Adjacent))
#removeScruz
summary(lm(gala$Species~gala$Elevation+gala$Adjacent))
normalModel = lm(gala$Species~gala$Elevation+gala$Adjacent)
#according to the step down method by removing Nearest, Area and then Scruz we have a linear model response = 1.43287 + 0.2767*Elevation - 0.06889*Adjacent


#2)
summary(lm(sqrt(gala$Species)~gala$Area+gala$Elevation+gala$Nearest+gala$Scruz+gala$Adjacent))
#remove Nearest
summary(lm(sqrt(gala$Species)~gala$Area+gala$Elevation+gala$Scruz+gala$Adjacent))
#remove Scruz
summary(lm(sqrt(gala$Species)~gala$Area+gala$Elevation+gala$Adjacent))
#remove Area
summary(lm(sqrt(gala$Species)~gala$Elevation+gala$Adjacent))
sqrtModel = lm(sqrt(gala$Species)~gala$Elevation+gala$Adjacent)
#according to the step down method by removing Nearest, Scruz and then Area we have a linear model sqrt(response) =  3.5378950 + 0.0129421*Elevation - 0.0028992*Adjacent


#3)

par(mfrow=c(1,2))

qqnorm(residuals(normalModel),main = "QQPlot Residuals")
qqnorm(residuals(sqrtModel),main = "QQPlot Residuals SQRT-Species")

plot(fitted(normalModel),residuals(normalModel),main = "Fitted Residuals")
plot(fitted(sqrtModel),residuals(sqrtModel),main = "Fitted Residuals SQRT-Species")
#from the plots one could see less outliers in the fitted Residual plot for the model based on square root species
#and also the qqplot for residuals based on square root of species is normal while the other qqplot is not normally distributed
# since there is not much difference between R squared then resulting linear model 2 is better
```

#4)

```
round(cooks.distance(sqrtModel),2)
#island 16 - Isabela is an influence point (1.64)
gala <- gala[-c(16), ] #removing island 16
# the model changes to 3.5606398 + 0.0084362* area + 0.0107337*Elevation - 0.0031392*Adjacent since by step-down Area wouldn't be dropped
```

#5)

```
par(mfrow=c(2,3))
for (i in 1:6) hist(gala[,i],main=colnames(gala)[i],xlab="",ylab="")
pairs(gala)
for (i in 1:6) hist(log(gala[,i]),main=colnames(gala)[i],xlab="",ylab="")
pairs(log(gala))
```

#6)

```
modlog=lm(log(Species)~log(Area)+log(Elevation)+log(Nearest)+log(Scruz+1)
        +log(Adjacent),data=gala)

modlog1=step(modlog)
summary(modlog1)
# if we look at summary modlog1, we see that the p value of Scruz + 1 is still > 0.05 so modlog1 didn't fully
#complete the step down method and thus Scruz + 1 needs to be removed
```

#7)

#8)

```
par(mfrow=c(1,2))
round(cooks.distance(modlog1),2) # no influence points
plot(1:30,cooks.distance(modlog1),main = "Cook's distance Plot")
qqnorm(residuals(modlog1),main = "QQPlot Residuals")
```

```
plot(fitted(modlog1),residuals(modlog1),main = "Fitted Residuals")


#according to cook's distance there are no influence points. the plot of the residuals shows that
the data is normally distributed
# and from the residuals vs. fitted plot we can see that there are no outliers and the range is
small [-1.5,1.7]
```

#9)

```
gala$logElevation=log(gala$Elevation)
gala$logArea=log(gala$Area)

galaScruz = log(gala$Scruz)
galaScruz[25] = NA
modlog2 = lm(log(gala$Species)~gala$logArea+gala$logElevation+galaScruz)

pairs(log(gala$Species)~gala$logArea+gala$logElevation+galaScruz)
gala$logScruz = galaScruz
round(cor(gala[,7:9]),2)

# there is a high collinearity between logElevation and logArea which means that this model
would be better
# by dropping one of them. showing interaction between logArea and logElevation would not
make sense as they
#have the same effect
```

#10)

```
# We prefer modlog1 model, although it has one extra explanatory variable Scruz + 1 .. but it
actually has a better
# R squared and it has no influence points and normally distributed residuals as well as normal
fitted
# residual plot.
```