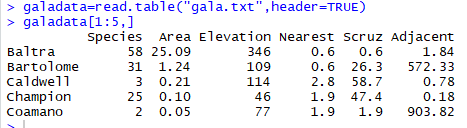
**Final Assignment\_Group11**

**Name:** Jiamian Liu **VU student number:** 2632301

**Name:** Xiaoyu Yang **VU student number:** 2640948

**Name:** Fangzheng Lyu **VU student number:** 2644757

We first read the “gala.txt” data via the following code and the output is shown in the same figure. We read the gala.txt to ‘galadata’ instead of ‘gala’ to avoid name overlapping.



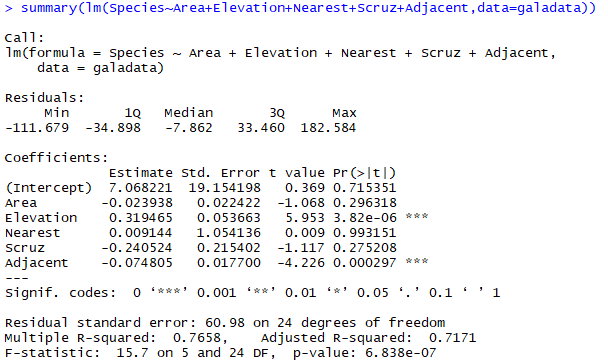
**Figure 1: read table and output**

From Fig.1 we could know that the header of this table is Species, Area, Elevation, Nearest, Scruz, Adjacent.

**Task 1**

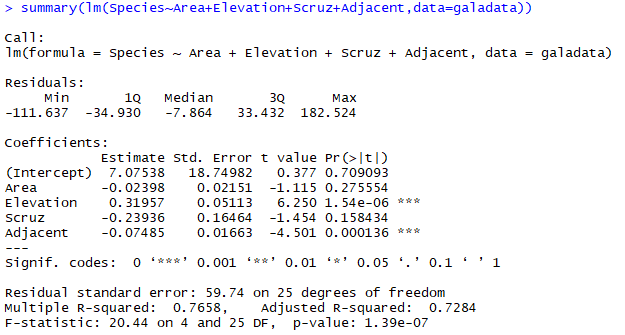
In this task, we use the step-down method to solve the problem for the response variable Species.

First code and output are shown in Fig.2.



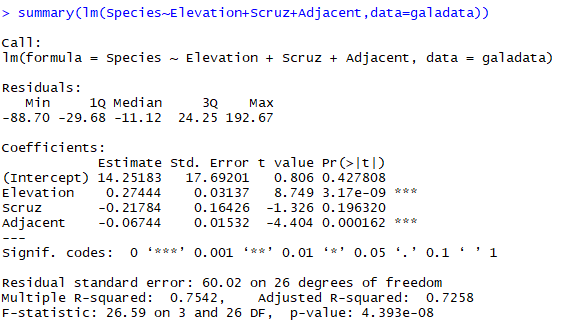
**Figure 2: First step of step-down method**

We could know that the p-value of “Nearest” is highest (0.993151) and is larger than 0.05, so we remove it. The next step code and output are shown in Fig.3.



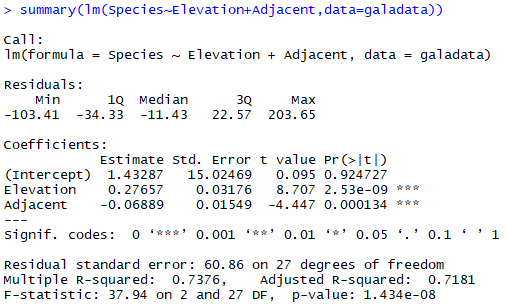
**Figure 3: Second step of step-down method**

We could know that the p-value of “Area” is highest (0.275554) and is larger than 0.05, so we remove it. The next step code and output are shown in Fig.4.



**Figure 4: Third step of step-down method**

We could know that the p-value of “Scruz” is highest (0.196320) and is larger than 0.05, so we remove it. The next step code and output are shown in Fig.5.



**Figure 5: Fourth step of step-down method**

We could know that all the p-value is smaller than 0.05, So all explanatory variables in the model are significant.

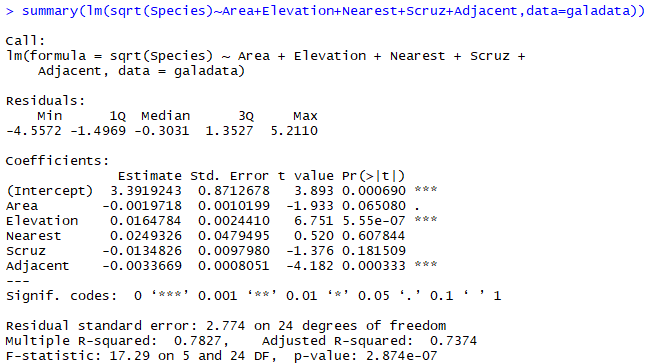
The resulting of the step-up method is:

Species = 1.43287 + 0.27657\*Elevation – 0.06889\*Adjacent + error

**Task 2**

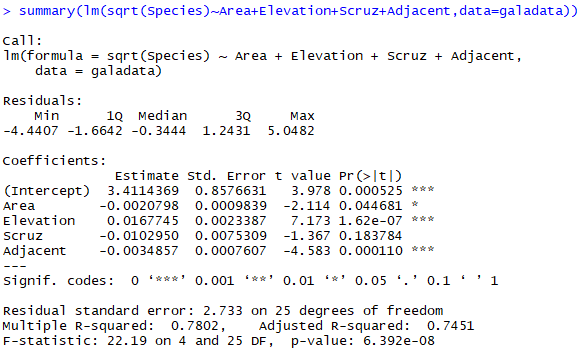
For the transformed response variable sqrt(Species) we do the similar steps.

First code and output are shown in Fig.6.



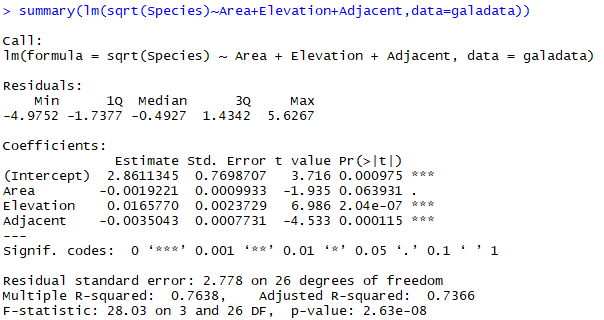
**Figure 6: First step of step-down method**

We could know that the p-value of “Nearest” is highest (0.607844) and is larger than 0.05, so we remove it. The next step code and output are shown in Fig.7.



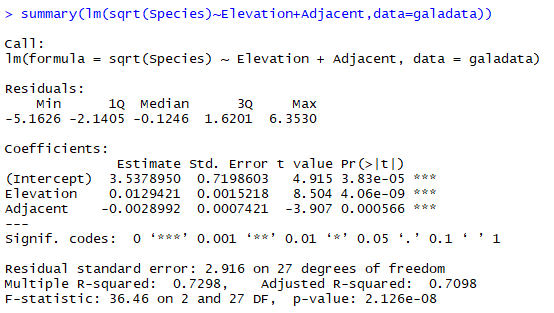
**Figure 7: Second step of step-down method**

We could know that the p-value of “Scruz” is highest (0.183784) and is larger than 0.05, so we remove it. The next step code and output are shown in Fig.8.



**Figure 8: Third step of step-down method**

We could know that the p-value of “Area” is highest (0.063931) and is larger than 0.05, so we remove it. The next step code and output are shown in Fig.9.



**Figure 9: Third step of step-down method**

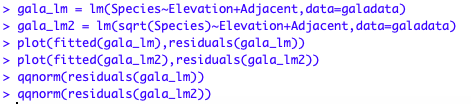
We could know that all the p-value is smaller than 0.05, So all explanatory variables in the model are significant.

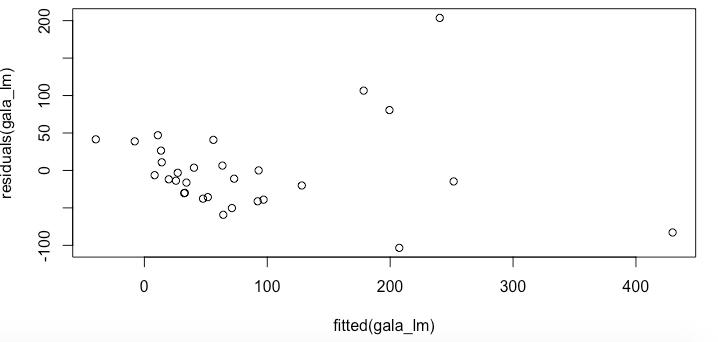
The resulting of the step-up method is:

sqrt(Species) = 3.5378950 + 0.0129421\*Elevation – 0.0028992\*Adjacent + error

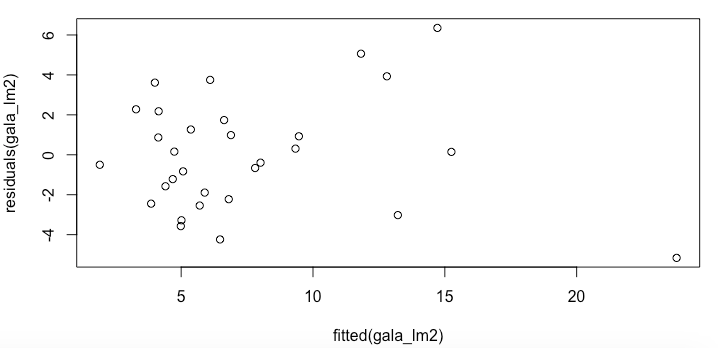
**Task 3**

The codes making plots are shown below:

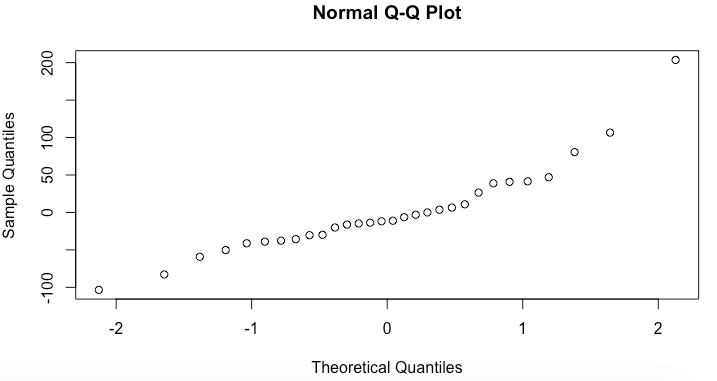
****

****

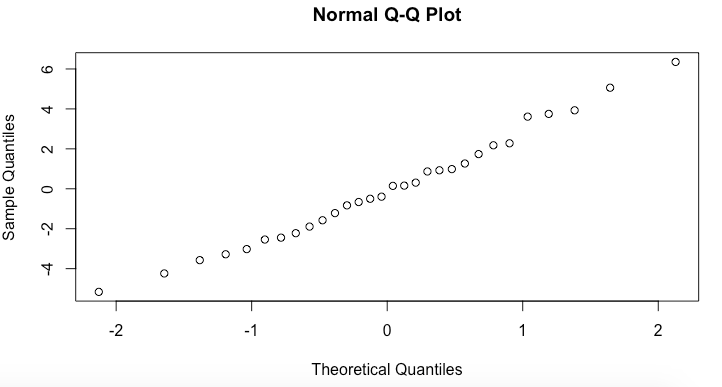
**Figure 10: fitted(gala\_lm)**



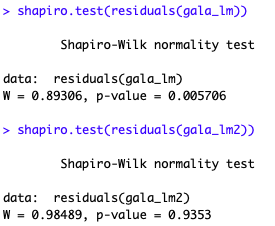
**Figure 11: fitted(gala\_lm2)**



**Figure 12: qqnorm(gala\_lm)**



**Figure 13: qqnorm(gala\_lm2)**



**Figure 14: Shapiro-test**

The QQ-plot of residuals and fitted plot are shown as Figure 10 to 13.

From the QQ-plot of “gala\_lm” we can see it have some curves and the second QQ-plot for sqrt model is approximating a straight line.

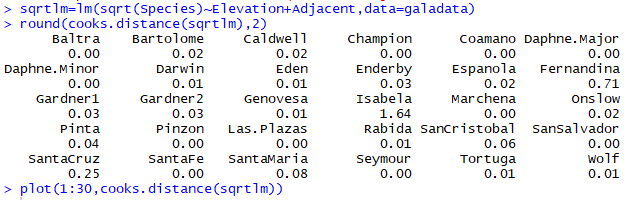
From the Shapiro.test we can see the p-value of residuals(gala\_lm) is smaller than 0.05 and p-value of residuals(gala\_lm2) is greater than 0.05. So the first sample is not normal while the second sqrt sample can be considered as normal.

So we believe the second resulting model using sqrt is better.

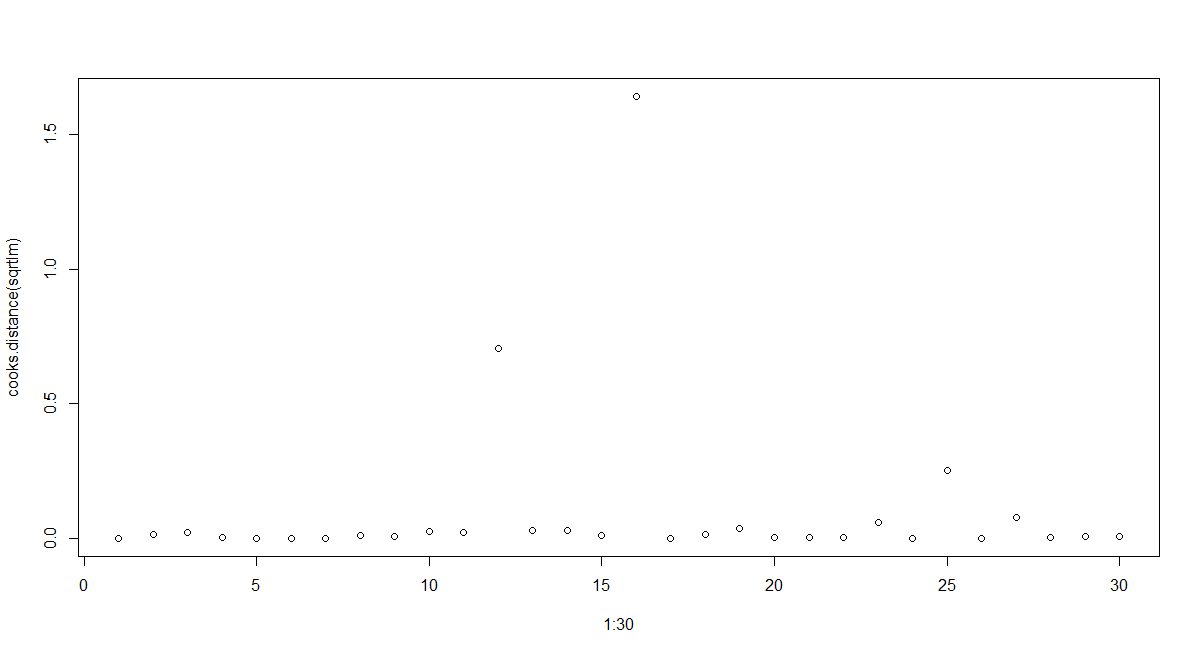
**Task 4**

We have two explanatory variables in our formula, which is “Elevation” and “Adjacent”. So we perform the Cook’s distance respectively.

The code for the “Elevation” and output is shown in Fig.15 and Fig.16.



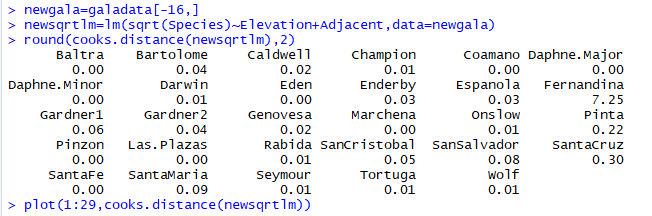
**Figure 15: Code and output for Cook’s distance**



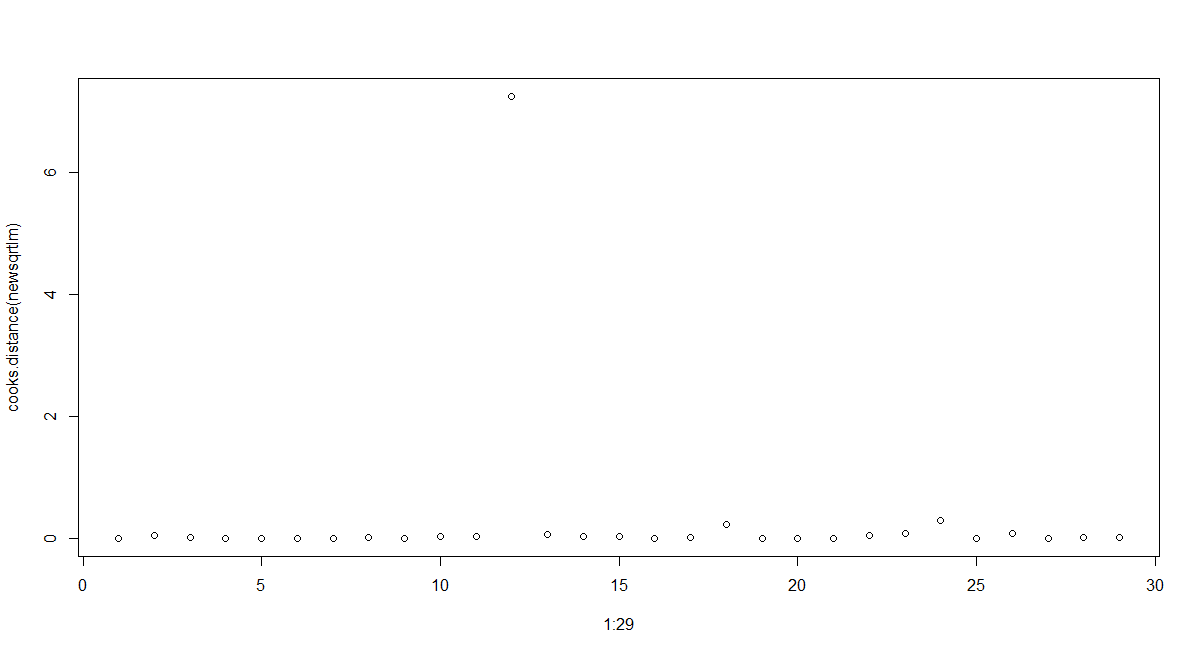
**Figure 16: Output for Cook’s distance**

From the output we could know that the value of “Isabela” is 1.64 larger than 1, so it is the influence point in this model.

We then remove the “Isabela”, from the data we could it is the 16th point. The code and output are shown in Fig.17 and Fig.18.



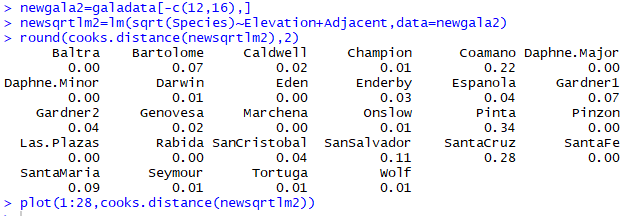
**Figure 17: Code and output for Cook’s distance 2**



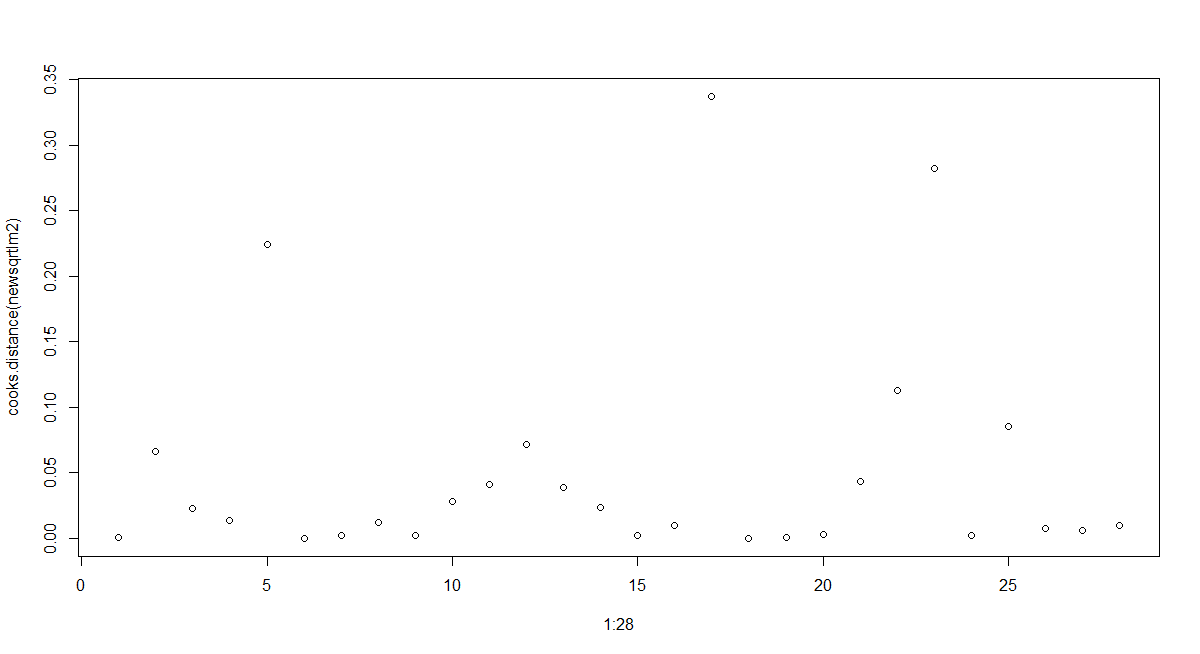
**Figure 18: Output for Cook’s distance 2**

From the output we could know that the value of “Fernandina” is 7.25 larger than 1, so it is still influence point.

We then remove the “Fernandina”, from the data we could it is the 12th point. The code and output are shown in Fig.19 and Fig.20.



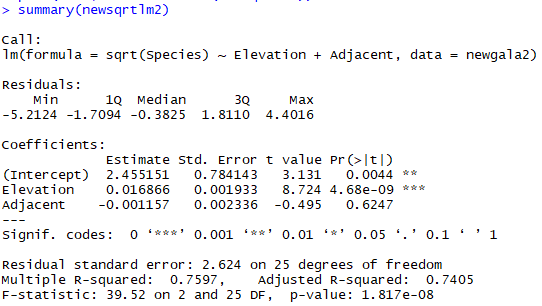
**Figure 19: Code and output for Cook’s distance 3**



**Figure 20: Output for Cook’s distance 3**

From the output we could know that all the output is smaller than 1 so there is no influence points now.

For the model, the code and output are shown in Fig21.



**Figure 21: New model output**

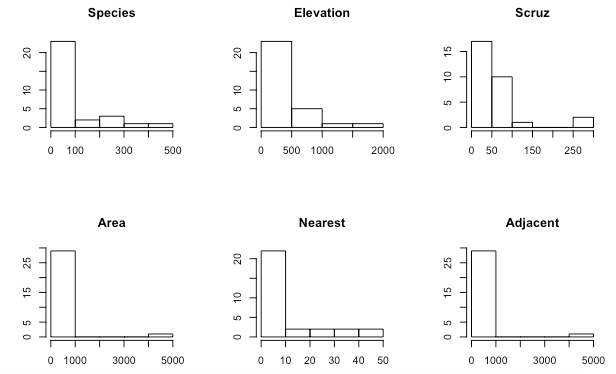
We could know that the new model is:

sqrt(Species) = 2.455151 + 0.016866\*Elevation – 0.001157\*Adjacent + error

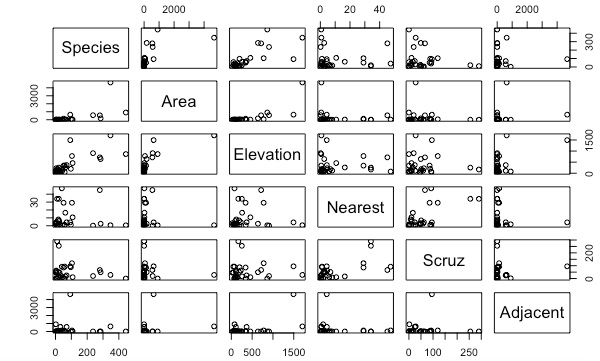
Which changed to some extent after removing the influence points.

**Task 5**

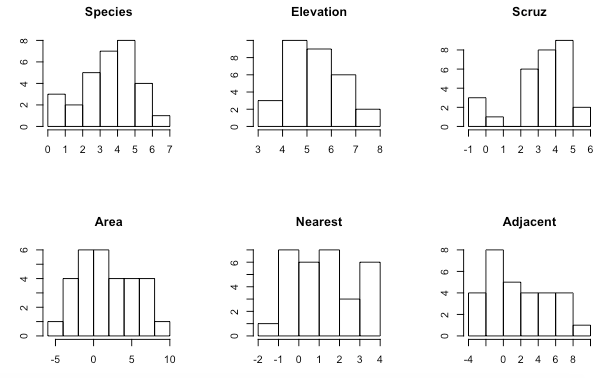
Using the given codes, we can make 4 diagrams shown as below:

****

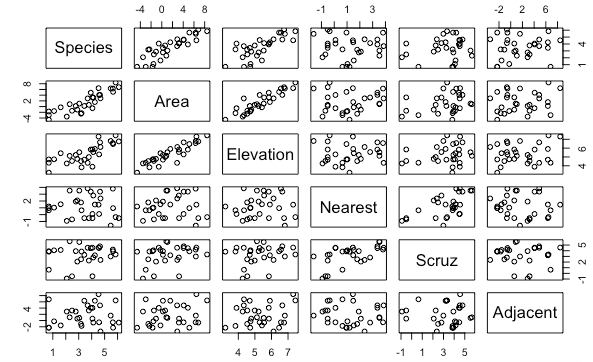
**Figure 22 histogram for each variable**

****

**Figure 23: Scatter plot for each variable**

****

**Figure 24 histogram for each variable**

****

**Figure 25: Scatter plot for each variable**

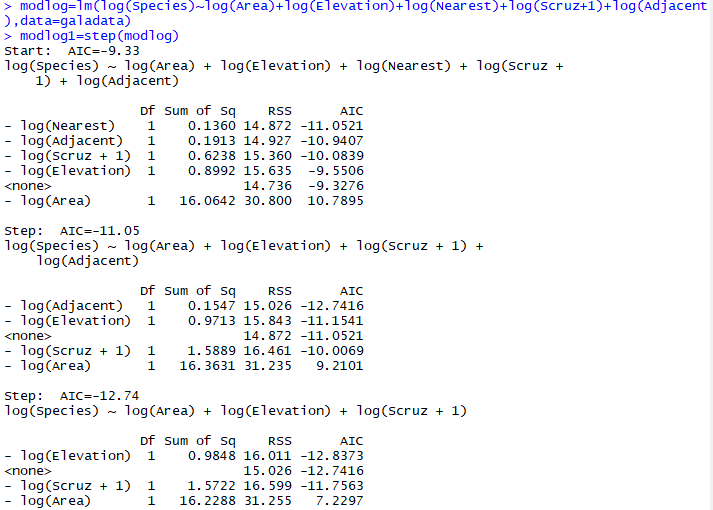
For the histograms, we can see the former ones without log show a Poisson distribution. After applying logarithm, the distribution shown by the histograms is close to the normal distribution.

And for the pairs, in the first diagram most of the data stack together, which make it difficult to observe and give a conclusion. But in the second diagram, we can find that the data get much scattered. And we can also find that there are some correlation between sqrt(Species)~sqrt(Area), sqrt(Species)~sqrt(Elevation), sqrt(Area)~sqrt(Elevation).

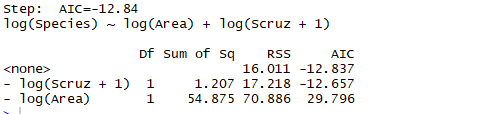
In general we can get a normal distribution, have a clearer view to the data and can find some relationship between the data.

**Task 6**

According to the instruction, the code and output are shown in Fig.26 and Fig.27.

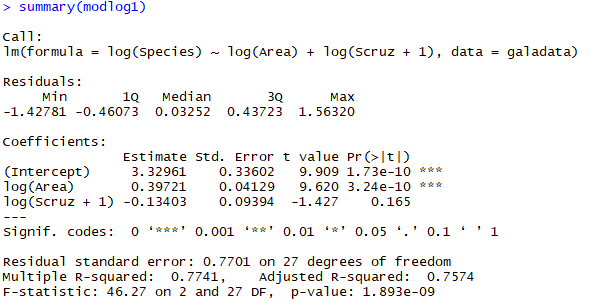


**Figure 26: Code and output 1**



**Figure 27: Code and output 2 (continue)**

From the AIC-criterion method, we know that there are finally 2 explanatory values, namely log(Scruz+1) and log(Area). We then summary the new modlog1, the code and output are shown in Fig.28.



**Figure 28: Summary for the model based on AIC-criterion**

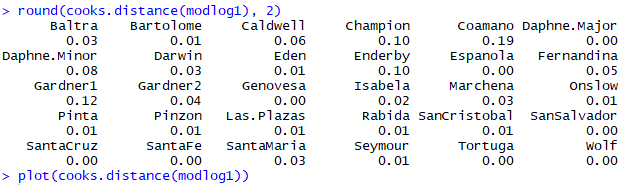
From the output we could know that the p-value of log(Scruz+1) is 0.165 larger than 0.05 and log(Area) is much smaller than 0.05. Hence, we know that only the log(Area) is significant.

**Task 7**

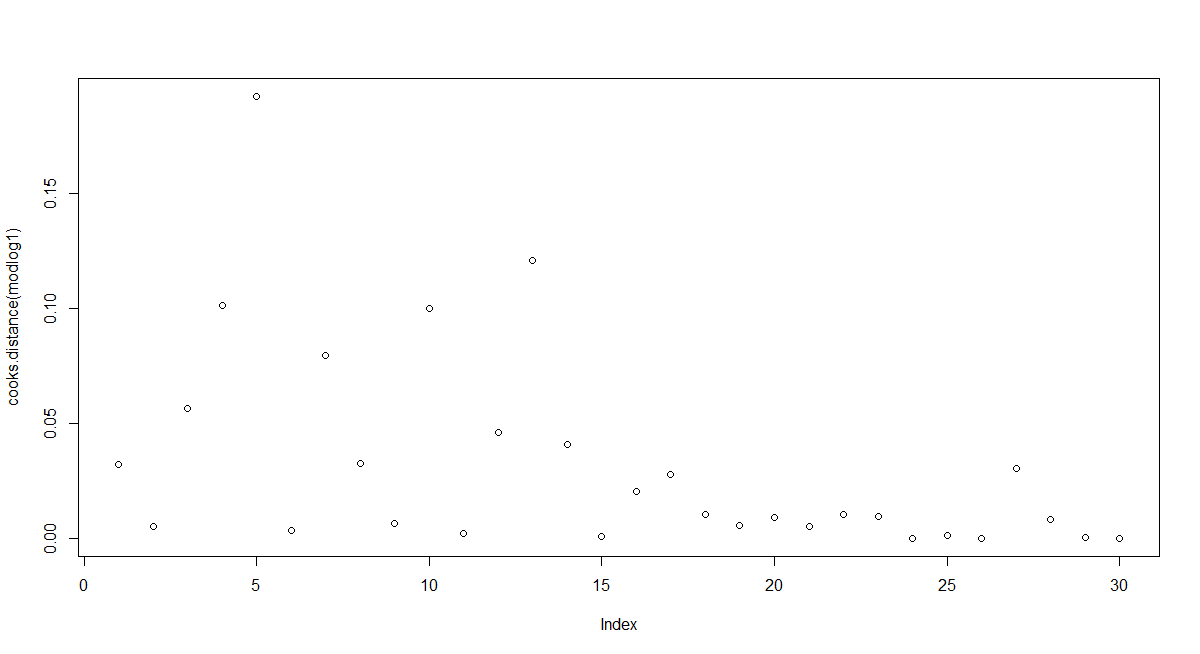
In task 5 we found the log(galadata) is from a normal distribution and some collinearities between log(Species) and log(Area). But we cannot find any collinearities between the variables without logarithm. Also we know the variable Species is follow a Poisson distribution, so it’s reasonable to use logarithm to model a linear function between log(Species) and log(Area).

**Task 8**

From Task 6 we could get the result of modlog1. We first plot Cook’s distance. The code and result are shown in Fig.29 and Fig.30.



**Figure 29: Code and output for Cook’s distance of modlog1**



**Figure 30: Output for Cook’s distance**

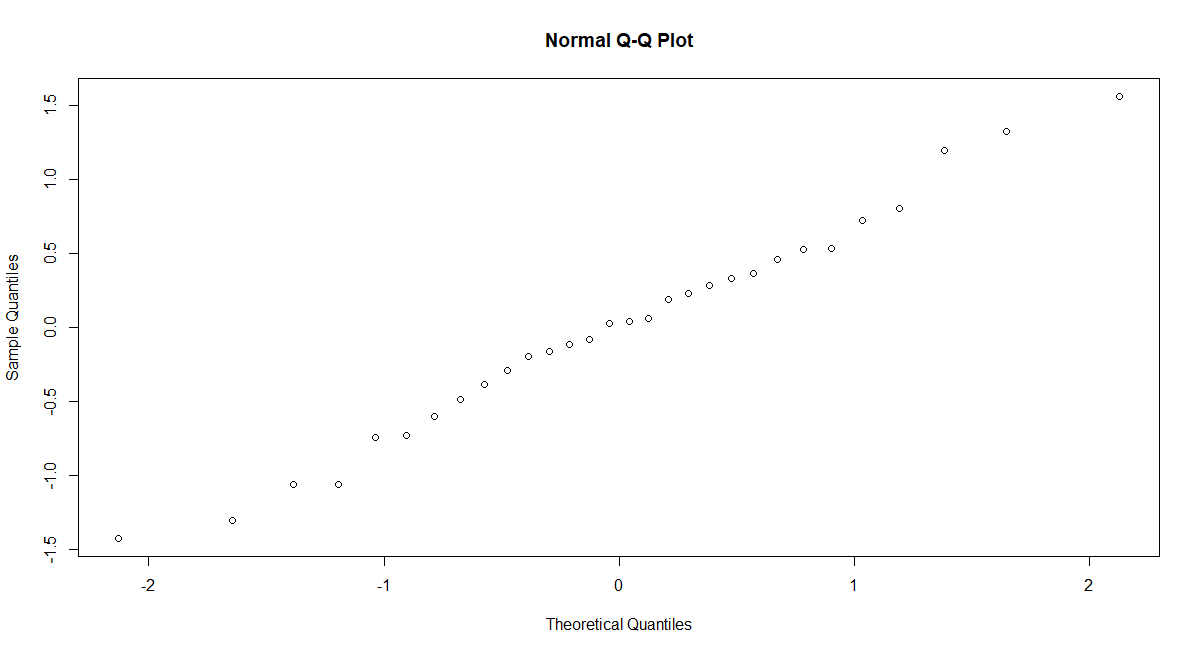
From the output we could know that all values of cook distance are smaller than 1, there is no influence point in this model.

We then draw the Q-Q plot of residuals and a plot of residuals versus fitted values via the following code as shown in Fig.31.

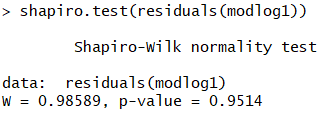


**Figure 31: Code for the rest of task 8**

The outputs are shown in Fig.32, Fig.33 and Fig.34 respectively.

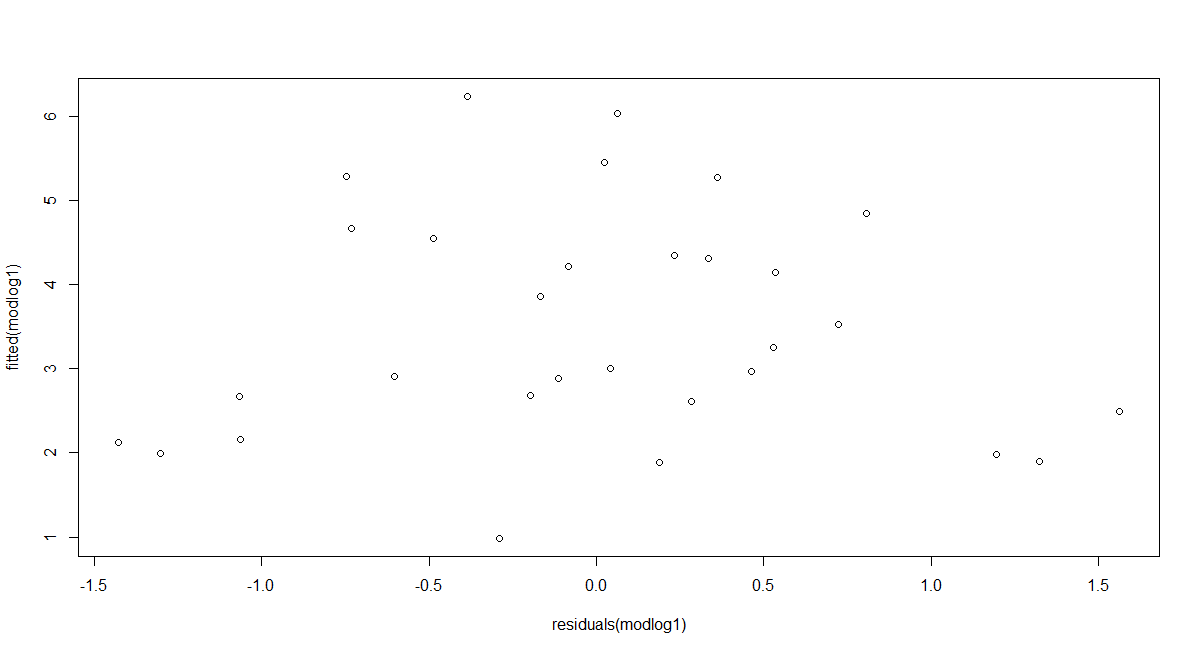


**Figure 32: Q-Q plot of residuals**



**Figure 33: Q-Q result of Shapiro test**

As the Q-Q plot and Shapiro test show, the distribution of residuals is normal distribution.



**Figure 34: Q-Q plot of fitted values**

From the plot of residuals versus fitted values, we can see there is scattered and no obvious structure, so there is no significant relation between the residuals and fitted values.

**Task 9**

After using the given codes to add columns to the gala\_data, we use pairs to make a plot as figure 35.

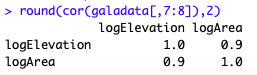




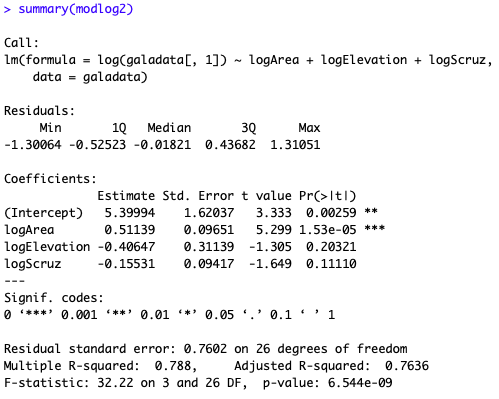
**Figure 35 Scatter plot for each variable**

From the diagram we can see that there might have possible collinearities between “logElevation” and “logArea”.

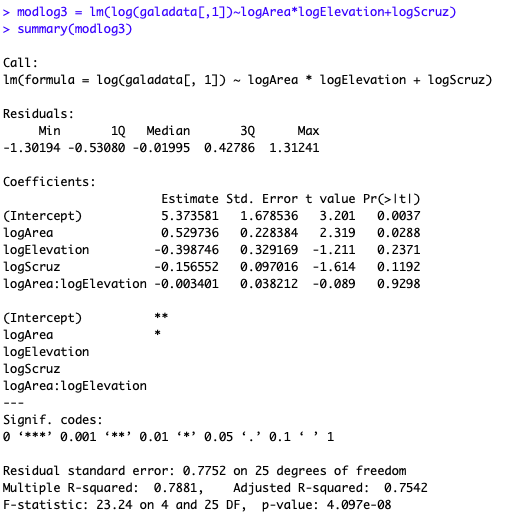
To check whether it is useful to include interactions between “logArea” and “logElevation”, we check the linear correlations between these two variables, the result is shown as figure 36.



**Figure 36 round for galadata**



**Figure 37 Summary**



**Figure 38 Summary**

We can find that both of variables have 0.9 correlations coefficients to the other in this pair. And also after using modlog3, we can see there’s no significant improvement of R-square So, it’s not useful to include interactions between “logElevation” and “logArea”.

**Task 10**

In task 3 we have compared two models and the model using sqrt(Species) is better than the model using Species.

For modlog1, there is scattered and no obvious structure, which means there’s no significant relation between residuals and fitted values. Therefore, we also discard this model.

And for modlog2, it contains two variables “logElevation” and “logArea” which have collinearities. Since it’s not useful to fit them in the same model, modlog2 is not the best choice.

To conclude, we prefer the model using sqrt(Species). Because it contains less variables and most of the variables in the model are relevant and easy to observe.