Exam

Tanjina Islam, Miguel Morales Expósito and Carlos Perales Liñan, group 12

21 March 2018

Exercise 1

First of all, we load the data from the data source.

```
gala_data = read.table("gala.txt", header = TRUE)
```

Task 1

```
gala_glm = glm(Species~Area+Elevation+Nearest+Scruz+Adjacent, family= poisson, data = gala_data
summary(gala_glm)
```

```
##
## Call:
## glm(formula = Species ~ Area + Elevation + Nearest + Scruz +
##     Adjacent, family = poisson, data = gala_data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -8.275  -4.497  -0.944   1.917  10.185
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.15e+00   5.17e-02   60.96  < 2e-16 ***
## Area        -5.80e-04   2.63e-05  -22.07  < 2e-16 ***
## Elevation    3.54e-03   8.74e-05   40.51  < 2e-16 ***
## Nearest      8.83e-03   1.82e-03    4.85  1.3e-06 ***
## Scruz       -5.71e-03   6.26e-04   -9.13  < 2e-16 ***
## Adjacent    -6.63e-04   2.93e-05  -22.61  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3510.73  on 29  degrees of freedom
## Residual deviance:  716.85  on 24  degrees of freedom
## AIC: 889.7
##
## Number of Fisher Scoring iterations: 5
```

As we can see, all the p-values are smaller than 0.05 now. Therefore, all the variables are significant for our model.

The resulting model of the step-down method is:

Species = 3.15 - 5.80e-04 * Area + 3.54e-03 * Elevation + 8.83e-03 * Nearest - 5.71e-03 * Scruz - 6.63e-04 * Adjacent + error

Task 2

```
galaglm=glm(sqrt(Species)~Area+Elevation+Nearest+Scruz+Adjacent,family=poisson,data=gala_data)
summary(galaglm)
```

```
##
## Call:
## glm(formula = sqrt(Species) ~ Area + Elevation + Nearest + Scruz +
##     Adjacent, family = poisson, data = gala_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7199  -0.8139   0.0244   0.5669   2.0002
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.39e+00   1.38e-01   10.04  < 2e-16 ***
## Area        -3.21e-04   9.80e-05   -3.28   0.0011 **
## Elevation    1.97e-03   2.80e-04    7.05  1.7e-12 ***
## Nearest      4.90e-03   5.96e-03    0.82   0.4118
## Scruz       -2.17e-03   1.55e-03   -1.40   0.1620
## Adjacent    -3.62e-04   9.32e-05   -3.89   0.0001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 106.111  on 29  degrees of freedom
## Residual deviance:  28.899  on 24  degrees of freedom
## AIC: Inf
##
## Number of Fisher Scoring iterations: 4
```

The variable Nearest has the highest p-value with a p-value of 0.411. Since this p-value is bigger than 0.05, we discard it for our model and continue to the next iteration.

```
galaglm=glm(sqrt(Species)~Area+Elevation+Scruz+Adjacent,family=poisson,data=gala_data)
summary(galaglm)
```

```
##
## Call:
## glm(formula = sqrt(Species) ~ Area + Elevation + Scruz + Adjacent,
##     family = poisson, data = gala_data)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7804  -0.8311  -0.0067   0.6591   1.8166
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.39e+00   1.38e-01   10.06  < 2e-16 ***
## Area        -3.39e-04   9.52e-05   -3.56  0.00037 ***
## Elevation    2.02e-03   2.73e-04    7.40  1.3e-13 ***
## Scruz       -1.41e-03   1.22e-03   -1.16  0.24665
## Adjacent    -3.85e-04   8.85e-05   -4.36  1.3e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 106.111  on 29  degrees of freedom
## Residual deviance:  29.555  on 25  degrees of freedom
## AIC: Inf
##
## Number of Fisher Scoring iterations: 4
```

The variable Scruz has the highest p-value with a p-value of 0.2466. Since this p-value is bigger than 0.05, we discard it for our model and continue to the next iteration.

```r
galaglm=glm(sqrt(Species)~Area+Elevation+Adjacent,family=poisson,data=gala_data)
summary(galaglm)
```

```
##
## Call:
## glm(formula = sqrt(Species) ~ Area + Elevation + Adjacent, family = poisson,
##     data = gala_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7682  -0.8578   0.0422   0.7971   1.6931
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.31e+00   1.24e-01   10.61  < 2e-16 ***
## Area        -3.26e-04   9.42e-05   -3.46  0.00053 ***
## Elevation    2.02e-03   2.73e-04    7.39  1.5e-13 ***
## Adjacent    -3.99e-04   8.80e-05   -4.53  5.8e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
```

```
##     Null deviance: 106.111  on 29  degrees of freedom
## Residual deviance:  30.983  on 26  degrees of freedom
## AIC: Inf
##
## Number of Fisher Scoring iterations: 4
```

As we can see, all the p-values are smaller than 0.05, thus meaning that all the variables are significant for our model.

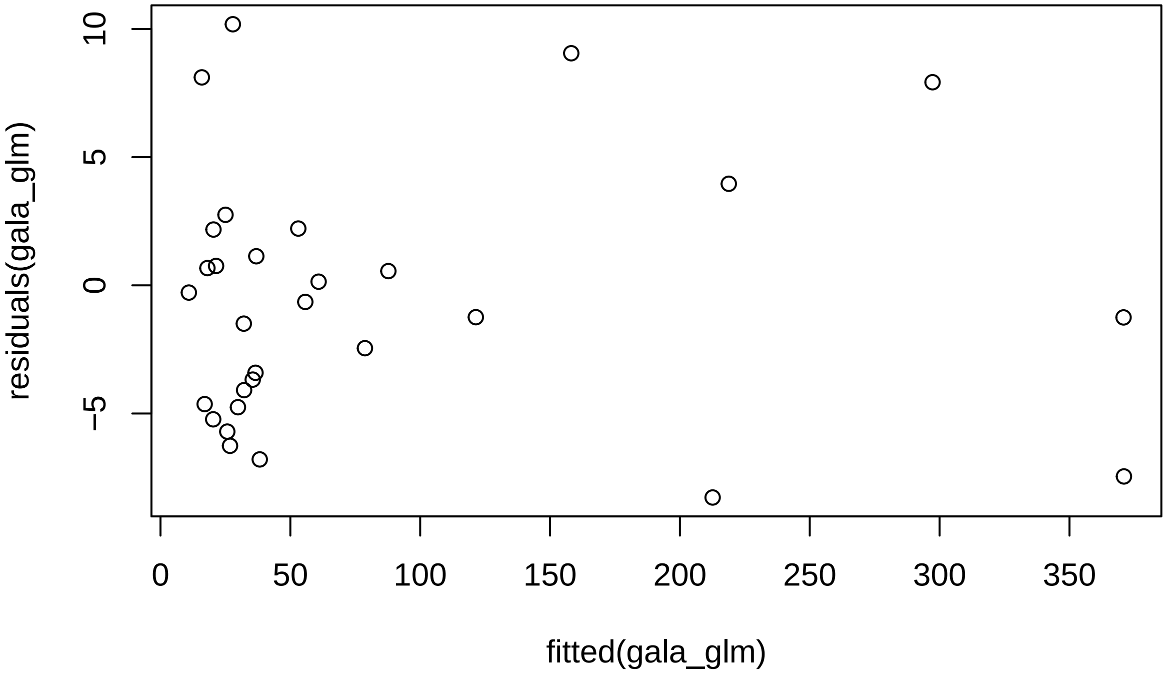
The resulting model of the step-down method is:

sqrt(Species) = 1.314 - 3.262e-04*Area + 2.018e-03*Elevation - 3.987e-04*Adjacent + error

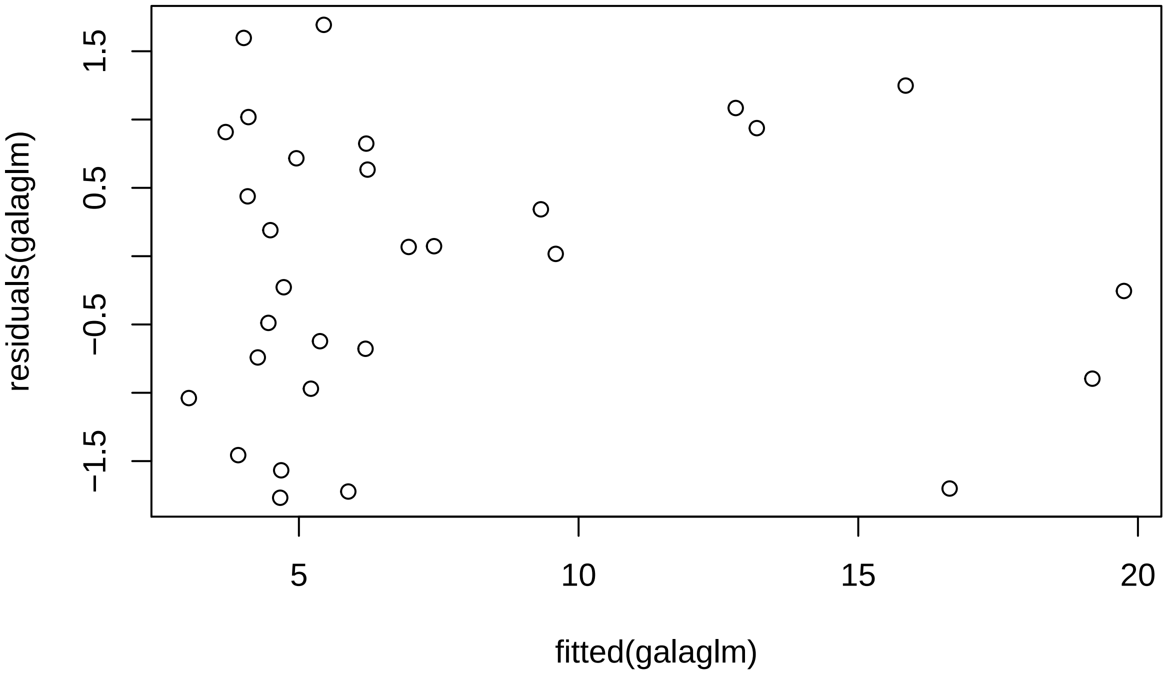
Task 3

Plots: Fitted vs Residuals

```
plot(fitted(gala_glm), residuals(gala_glm))
```



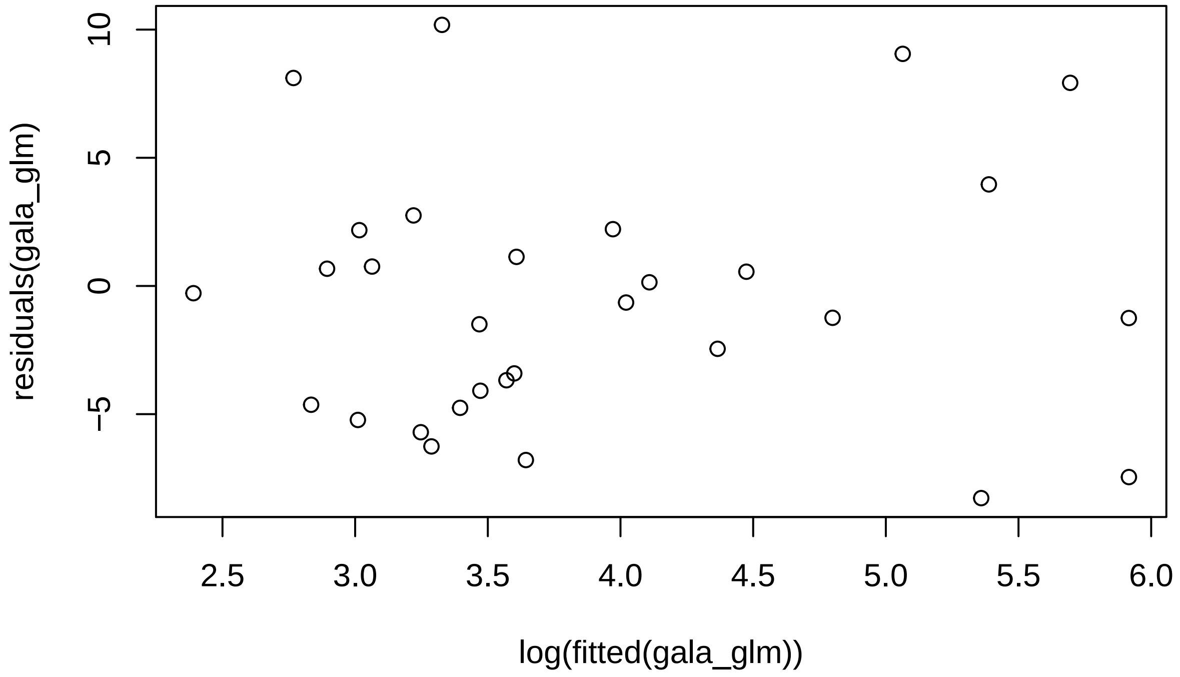
```
plot(fitted(galaglm), residuals(galaglm))
```



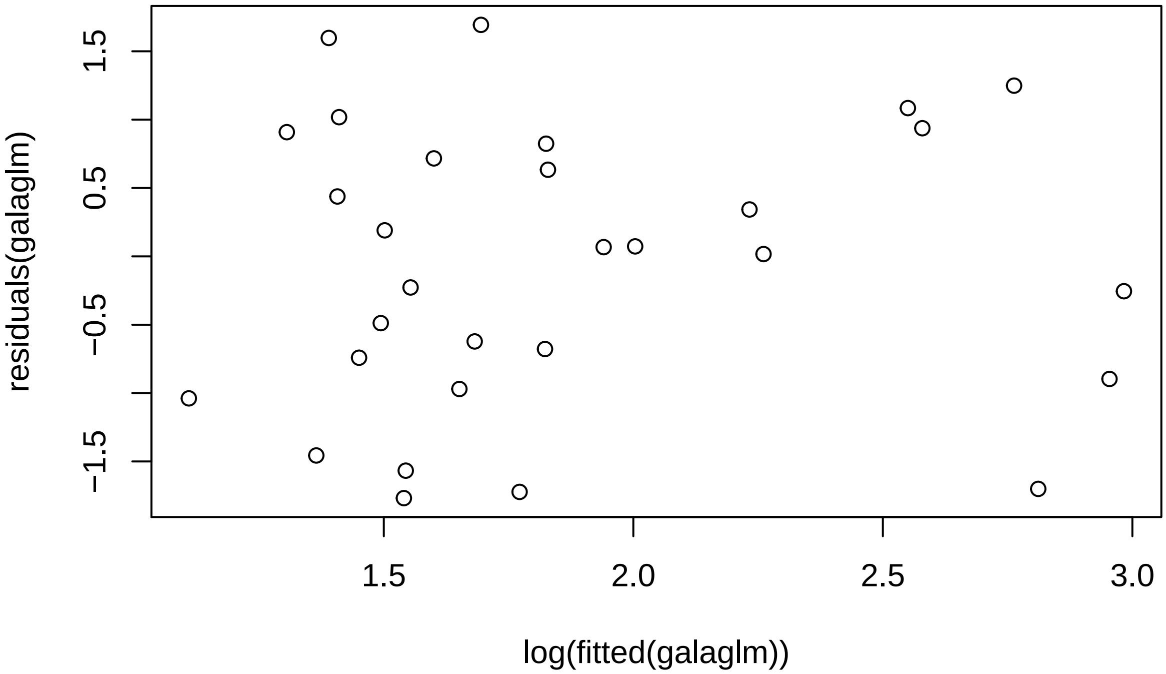
Both plots show no specific pattern. They're scatter plots, not visually good as it would suppose to be in linear regression model.

Plots: Logarithmic-Fitted values vs Residuals

```
plot(log(fitted(gala_glm)), residuals(gala_glm))
```



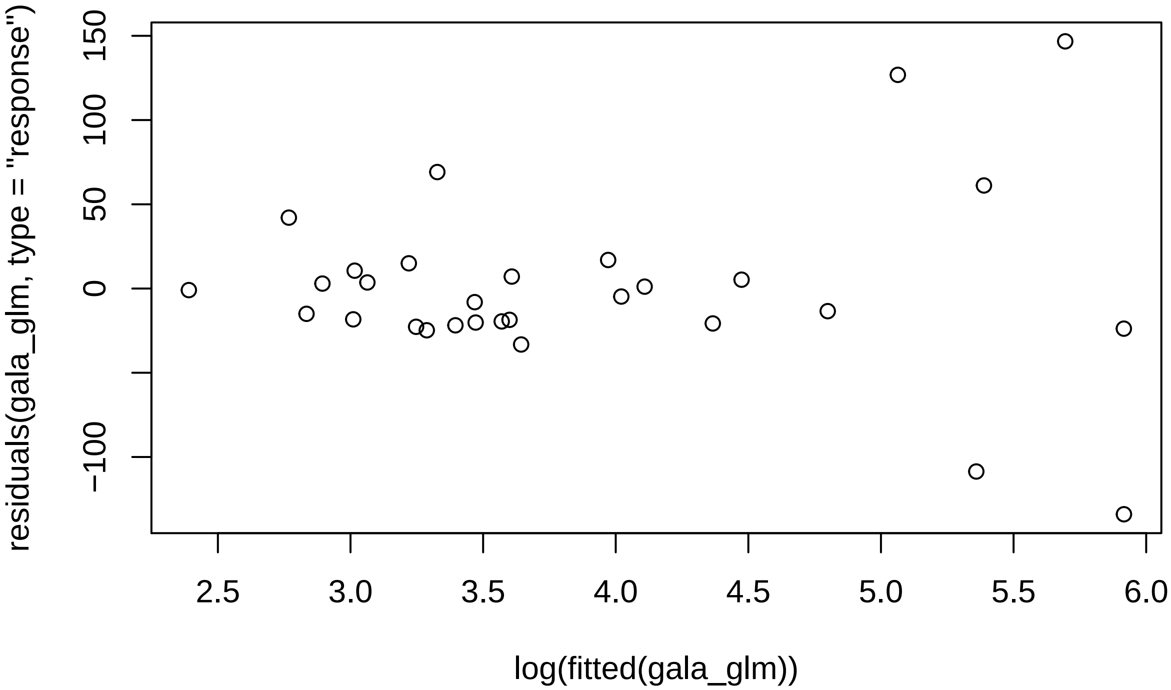
```
plot(log(fitted(galaglm)), residuals(galaglm))
```



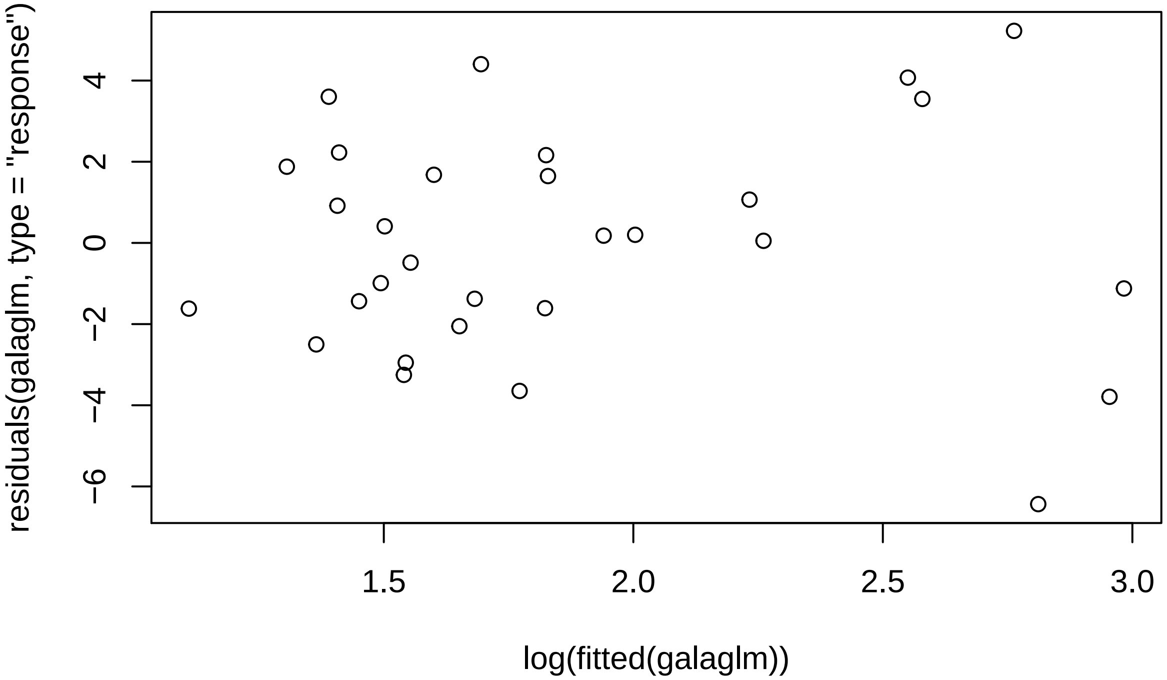
We took logarithm to make the x-values fitted by a linear function in the plots. Both plots seem OK yet still no specific pattern. In addition to this, they also look scattered.

Plots: Logarithmic-Fitted values vs Residuals with type = Response

```
plot(log(fitted(gala_glm)), residuals(gala_glm, type = "response"))
```



```r
plot(log(fitted(galaglm)), residuals(galaglm, type = "response"))
```



With the first model's plot, we can see that the response residuals clearly increase with log of the fitted values as expected from Poisson model. However, we can't quite say the same from the second model.

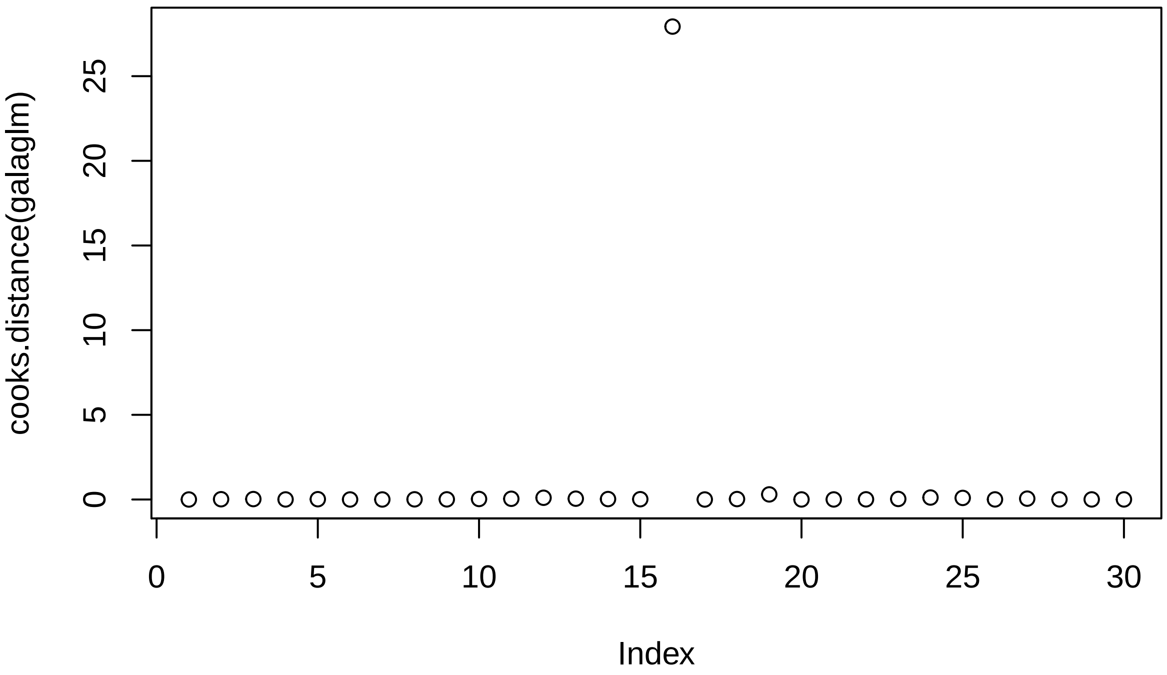
In conclussion, both models show very similar characteristics when looking at the plots. Considering the last plot, we conclude that the fisrt model looks better than the sqrt model. Therefore, we believe this model is better despite the similarities between them on the other aspects.

Task 4

```
round(cooks.distance(galaglm), 2)
```

```
##       Baltra    Bartolome     Caldwell     Champion      Coamano
##         0.00         0.02         0.02         0.00         0.02
## Daphne.Major Daphne.Minor       Darwin         Eden      Enderby
##         0.00         0.00         0.01         0.01         0.03
##     Espanola   Fernandina     Gardner1     Gardner2     Genovesa
##         0.04         0.10         0.05         0.03         0.02
##      Isabela     Marchena       Onslow        Pinta       Pinzon
##        27.93         0.00         0.02         0.30         0.00
##   Las.Plazas       Rabida SanCristobal  SanSalvador    SantaCruz
##         0.00         0.01         0.03         0.12         0.09
##      SantaFe   SantaMaria      Seymour      Tortuga         Wolf
##            0.00          0.05          0.01          0.00          0.00
```

```r
plot(cooks.distance(galaglm))
```



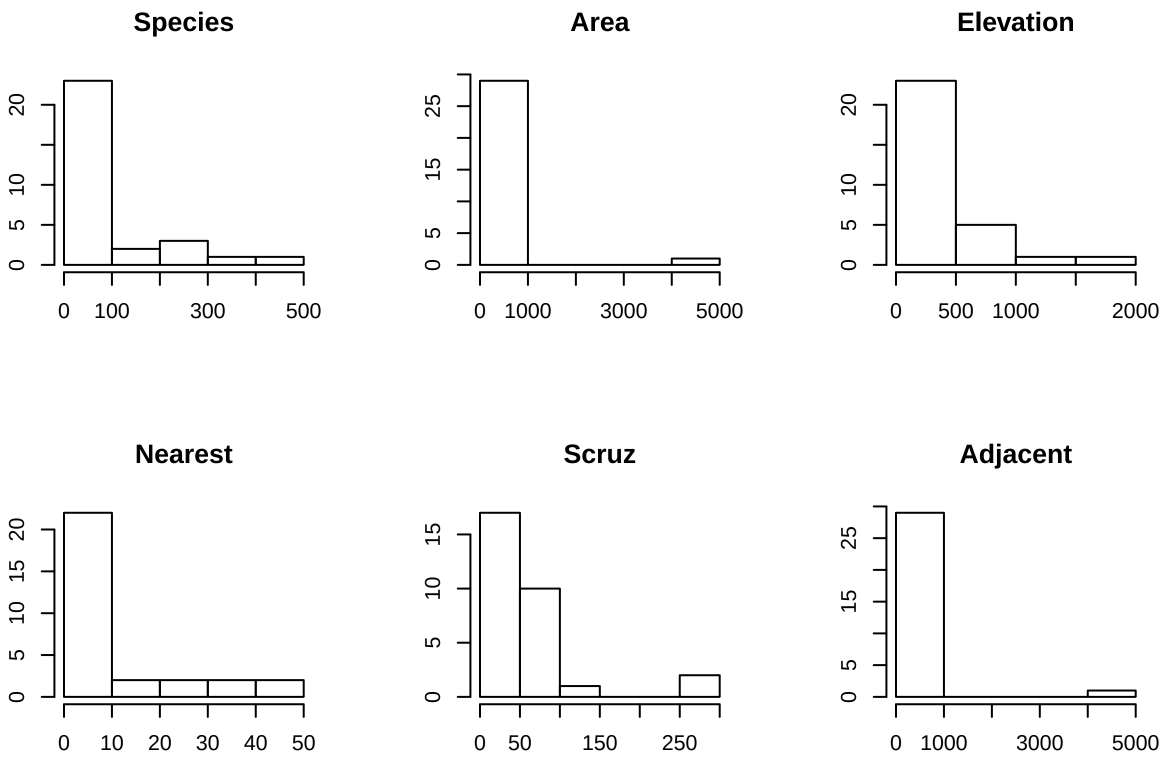
Here we clearly have encountered an influence point: the Cook's distance is 27.93 that refers to "Isabela" as the potential point.

If this Island is removed, then we will not get any influential points from Cook's distance except "Isabela", since there is no data point with a Cook's distance that is close to or larger than 1 rather than Isablea = 27.93.

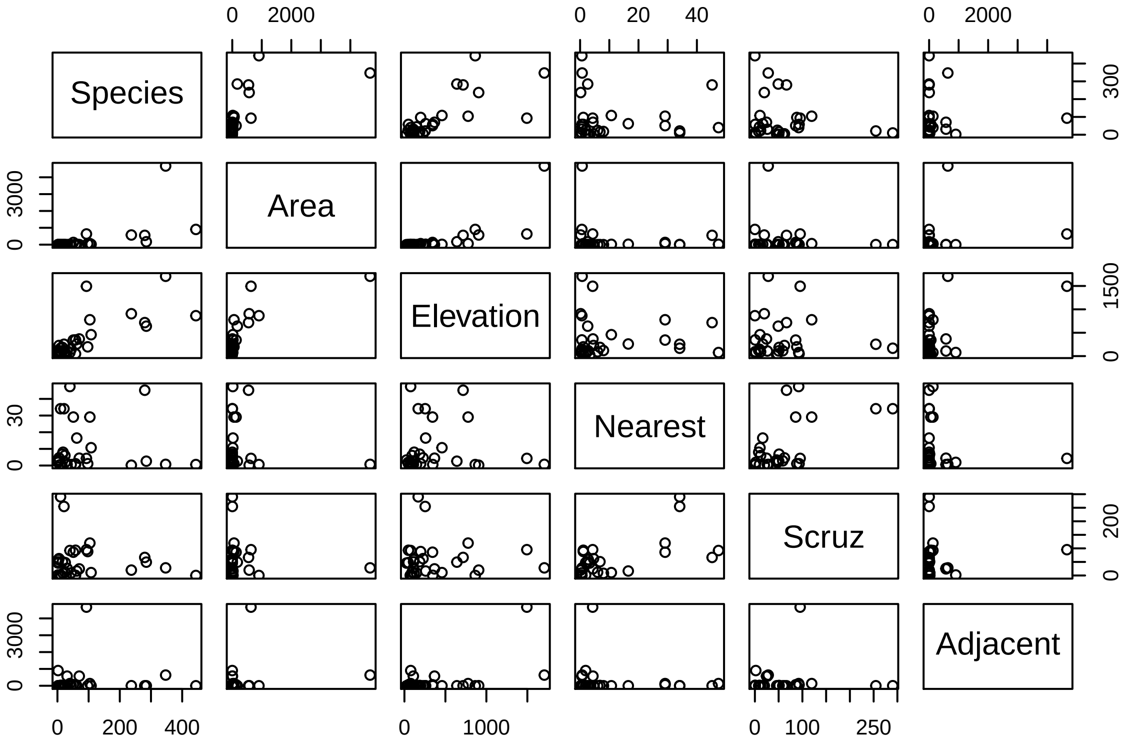
Therefore, we can conclude that there will be no influence point if we remove it from the model.

Task 5

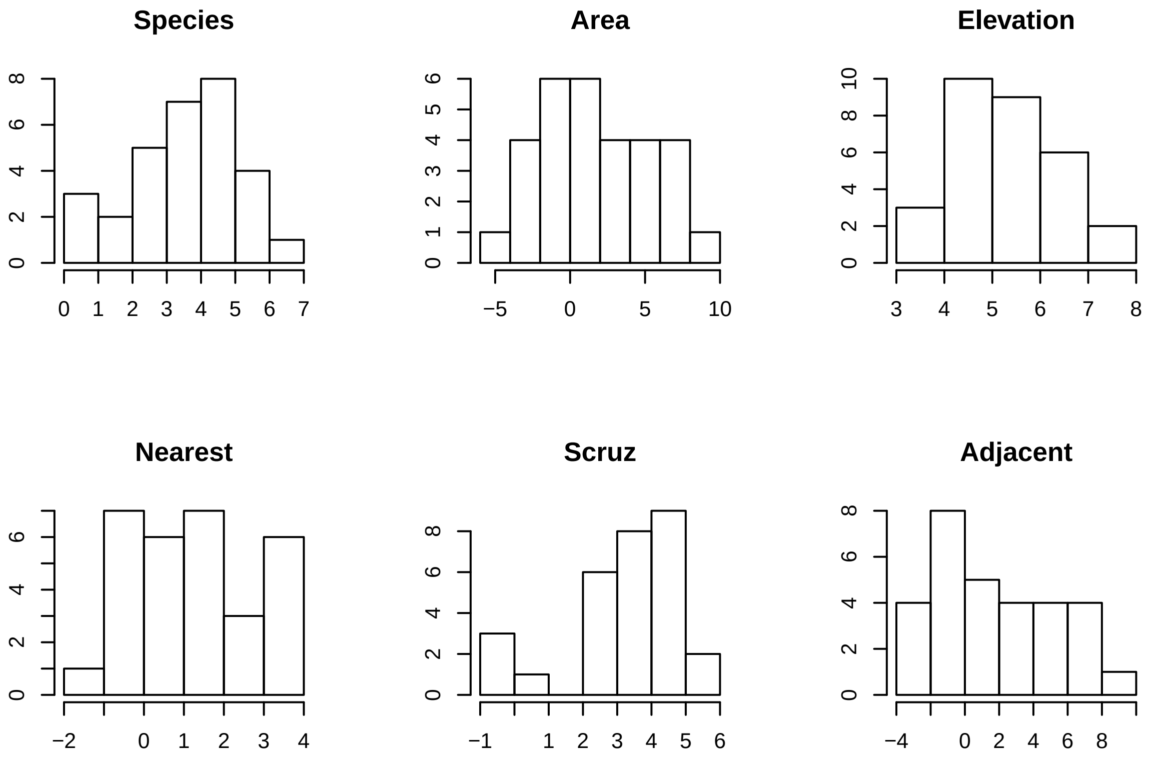
```r
par(mfrow=c(2,3))
for (i in 1:6) hist(gala_data[,i],main=colnames(gala_data)[i],xlab="",ylab="")
```



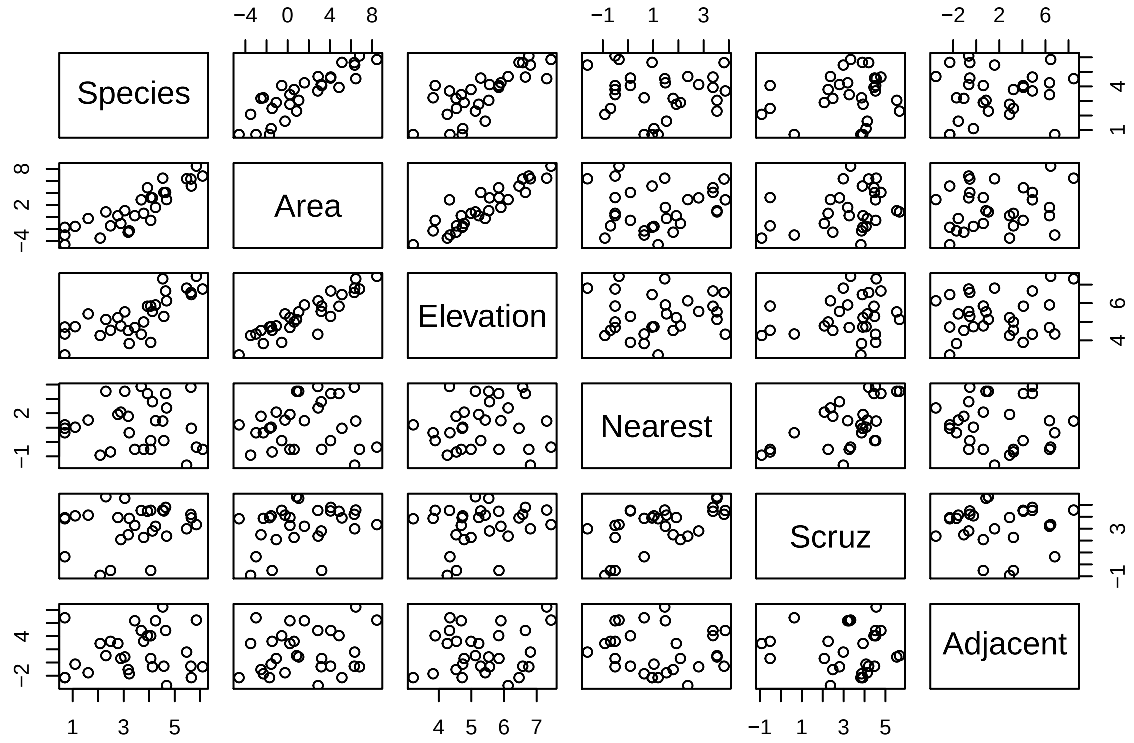
```r
pairs(gala_data)
```



```r
for (i in 1:6) hist(log(gala_data[,i]),main=colnames(gala_data)[i],xlab="",ylab="")
```



```r
par(mfrow=c(1,1))
pairs(log(gala_data))
```



If we look into the histograms, for the first one (without log), we see that the histograms show the pattern of a poisson distribution as expected. On the other hand, when we applied logarithm in the second version, we see that they are approximately equal to a normal distribution.

If we look into the pairs then for the first one we see that, the data doesn't look significant to make any conclusion. However, when we applied the lograthim and cheked the pairs then we can see the data get more scattered and we can find some colinearity between Species~Area, Species~Elevation and Area~Elevation and there might be some colinearity between Nearest~Scruz, but it's not obvious.

In summary, by applying the this we tried to normalize the distribution and we get more relevant information.

Task 6

```
modlog=lm(log(Species)~log(Area)+log(Elevation)+log(Nearest)+log(Scruz+1)+log(Adjacent),data=ga
modlog1=step(modlog)
```

```
summary(modlog1)
```

```
##
## Call:
## lm(formula = log(Species) ~ log(Area) + log(Scruz + 1), data = gala_data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4278 -0.4607  0.0325  0.4372  1.5632
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.3296     0.3360    9.91  1.7e-10 ***
## log(Area)        0.3972     0.0413    9.62  3.2e-10 ***
## log(Scruz + 1)  -0.1340     0.0939   -1.43     0.17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.77 on 27 degrees of freedom
## Multiple R-squared:  0.774,  Adjusted R-squared:  0.757
## F-statistic: 46.3 on 2 and 27 DF,  p-value: 1.89e-09
```

If we look into the summary of modlog1, we can see that log(Scruz + 1) has a p-value = $0.17 > 0.05$. So maybe this variable is not significant but we are not sure whether AIC follows the exact same criteria in selecting a model. The same way step-up only follows the R-squared values, we think that the p-values might not be relevant to determine the final fitness of AIC.

Task 7

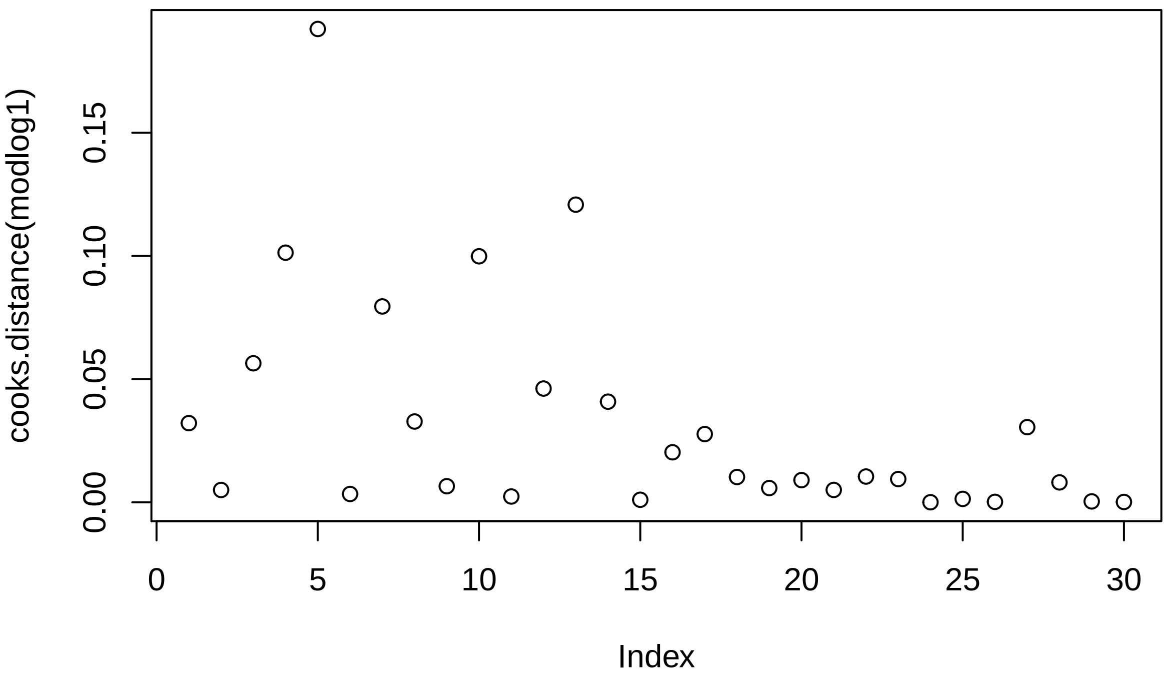
When we apply the logarithm, we normalized the data and we found collinearity between Spices and Area. In addition to this, if we consider them without logarithm we do not really see any colinearity between them. By the definition of Species, we get that we are counting the number of specifes in an island which follows Poissons regression model. Since Poisson uses logarithm to get the linear function, it makes more sense to apply logarithm to model log(Species) as a linear function of log(Area) than Species as a linear function of Area.

Task 8

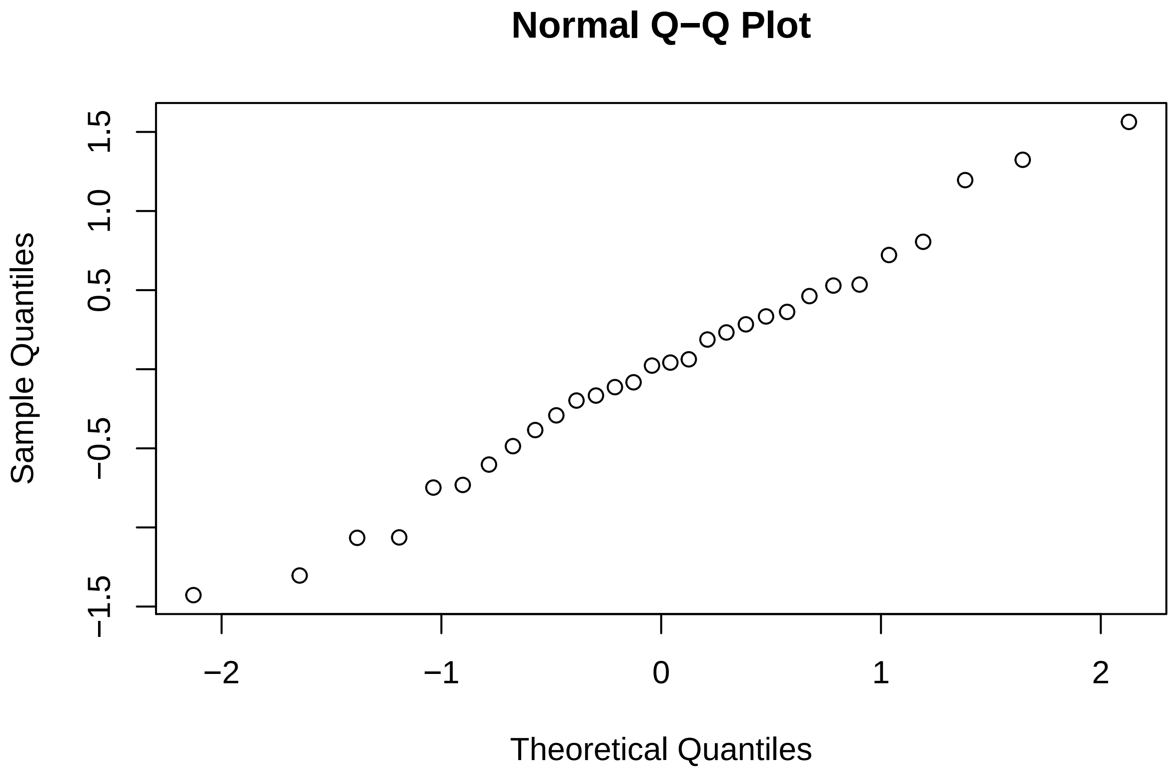
```r
round(cooks.distance(modlog1), 2)
```

```
##       Baltra    Bartolome     Caldwell     Champion      Coamano
##         0.03         0.01         0.06         0.10         0.19
## Daphne.Major Daphne.Minor       Darwin         Eden      Enderby
##         0.00         0.08         0.03         0.01         0.10
##     Espanola   Fernandina     Gardner1     Gardner2     Genovesa
##         0.00         0.05         0.12         0.04         0.00
##      Isabela     Marchena       Onslow        Pinta       Pinzon
##         0.02         0.03         0.01         0.01         0.01
##   Las.Plazas       Rabida SanCristobal  SanSalvador    SantaCruz
##         0.01         0.01         0.01         0.00         0.00
##      SantaFe   SantaMaria      Seymour      Tortuga         Wolf
##         0.00         0.03         0.01         0.00         0.00
```

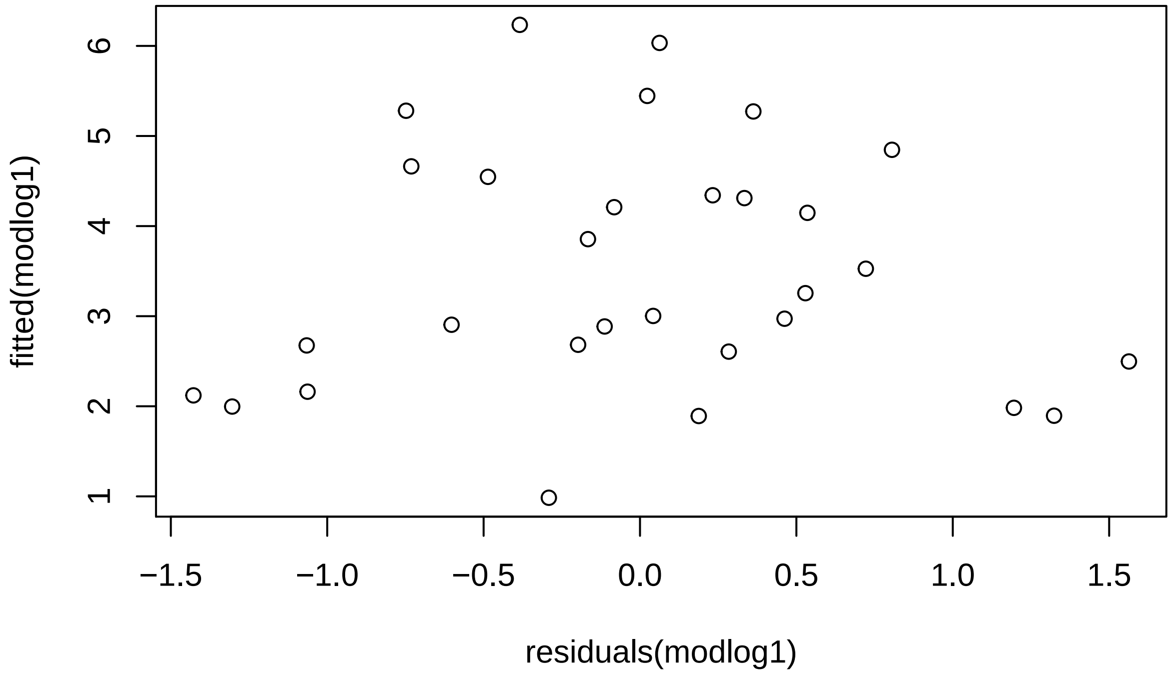
```r
plot(cooks.distance(modlog1))
```



```r
qqnorm(residuals(modlog1))
```



```
plot(residuals(modlog1), fitted(modlog1))
```



Since we can see that there are no data point with a Cook's distance close to or larger than 1, we can conclude that there is no influence point. We can see that the QQ-plot shows normality and the residuals plot looks scattered but doesn't follow any specific structure.

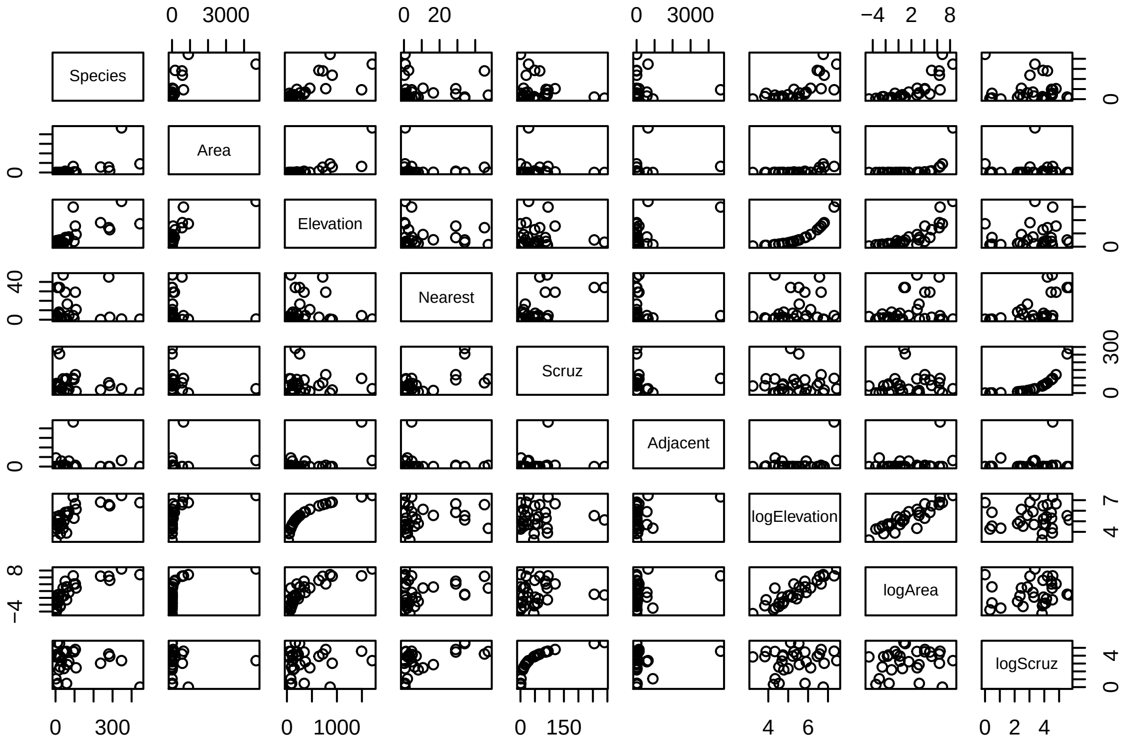
Task 9

To investigate collinearity problems, we first have to check the pairs with the scatter plot.

```
gala_data$logElevation = log(gala_data$Elevation)
gala_data$logArea = log(gala_data$Area)
gala_data$logScruz = log(gala_data$Scruz + 1)

modlog2 = lm(log(Species)~logArea+logElevation+logScruz,data = gala_data)

pairs(gala_data)
```



Here we can see that the variables “logElevation” and “logArea” look somewhat collinear. We will now check the linear correlation between both.

```r
round(cor(gala_data[,7:8]),2)
```

```
##              logElevation logArea
## logElevation          1.0     0.9
## logArea               0.9     1.0
```

As suspected, both variables are collinear with [0.9]. It’s not useful to include interaction between these two variables in the same model.

Task10

We can discard modlog2 because it contains two variables that are collinear (logElevation and logArea), so it’s not useful to fit them in the same model.

For modlog1, it seems that one of the variables that we are considering is not so significant according to the p-value. The model we considered in exercise1 has too many variables.

So finally, we think we should consider the model with sqrt(Species) because it has less explanatory variables, all the variables seem relevant and intuitively this model makes sense.