Homework 3

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(magrittr)
```

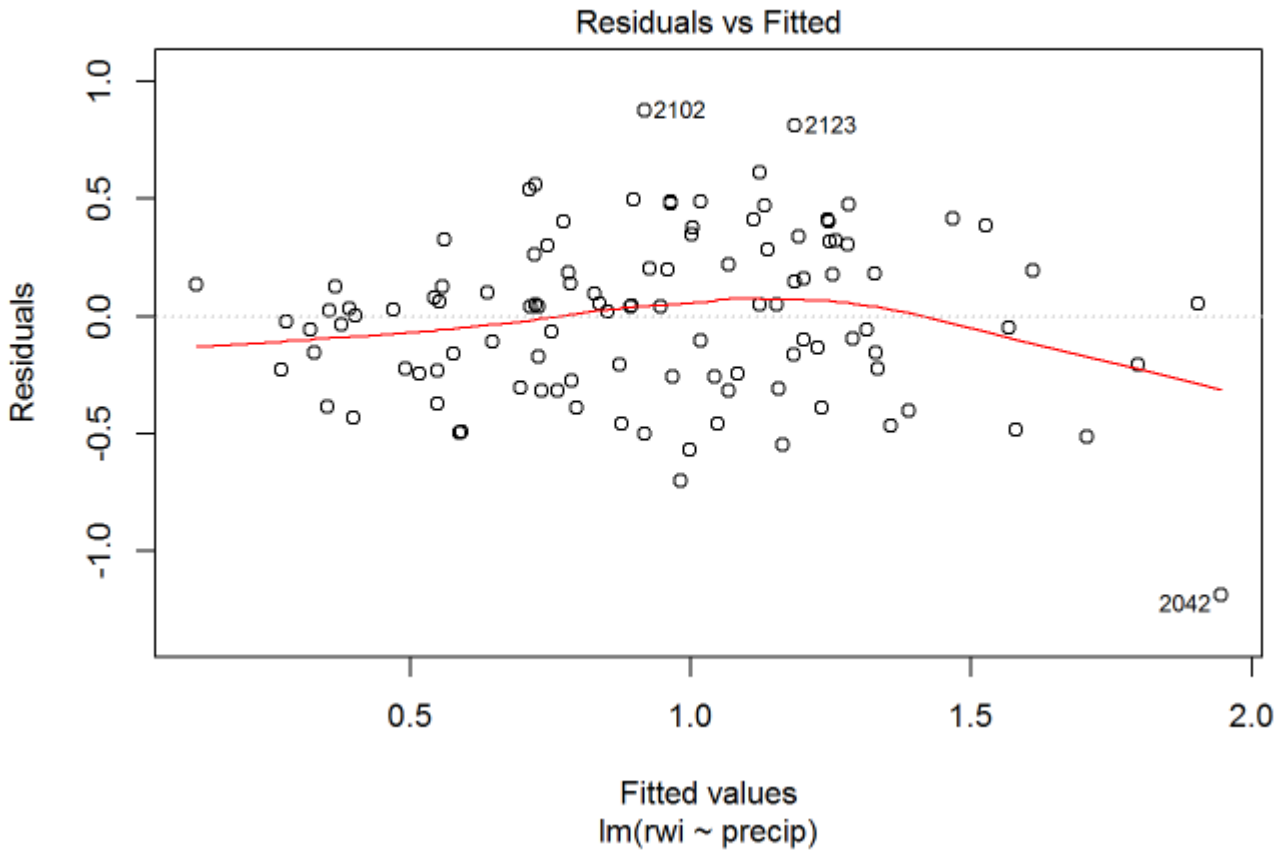
```
##
## Attaching package: 'magrittr'
##
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
precip <- read.csv('precip.csv')
rwi <- read.csv('rwi.csv')
precip.df <- precip %>% gather(key=month, value=precip, -year) %>% arrange(year, month)
precip.df <- precip.df %>% mutate(water_year = lead(year, 6))
annual_precip <- precip.df %>% group_by(water_year) %>%
  summarize(precip=sum(precip)) %>%
  filter(water_year>1895)
rwi.precip.df <- left_join(rwi, annual_precip, by=c('year'='water_year'))
model <- lm(rwi~precip, data=rwi.precip.df)
summary(model)
```

```
##
## Call:
## lm(formula = rwi ~ precip, data = rwi.precip.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.18958 -0.24341  0.03348  0.21877  0.87653
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.660621   0.146582  -4.507 1.69e-05 ***
## precip       0.108460   0.009682  11.202  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3477 on 107 degrees of freedom
##   (2032 observations deleted due to missingness)
## Multiple R-squared:  0.5397, Adjusted R-squared:  0.5354
## F-statistic: 125.5 on 1 and 107 DF,  p-value: < 2.2e-16
```

1. For the first three diagnostic plots, describe wich assumptions they help to visually evaluate.

```
plot(model, which=1)
```

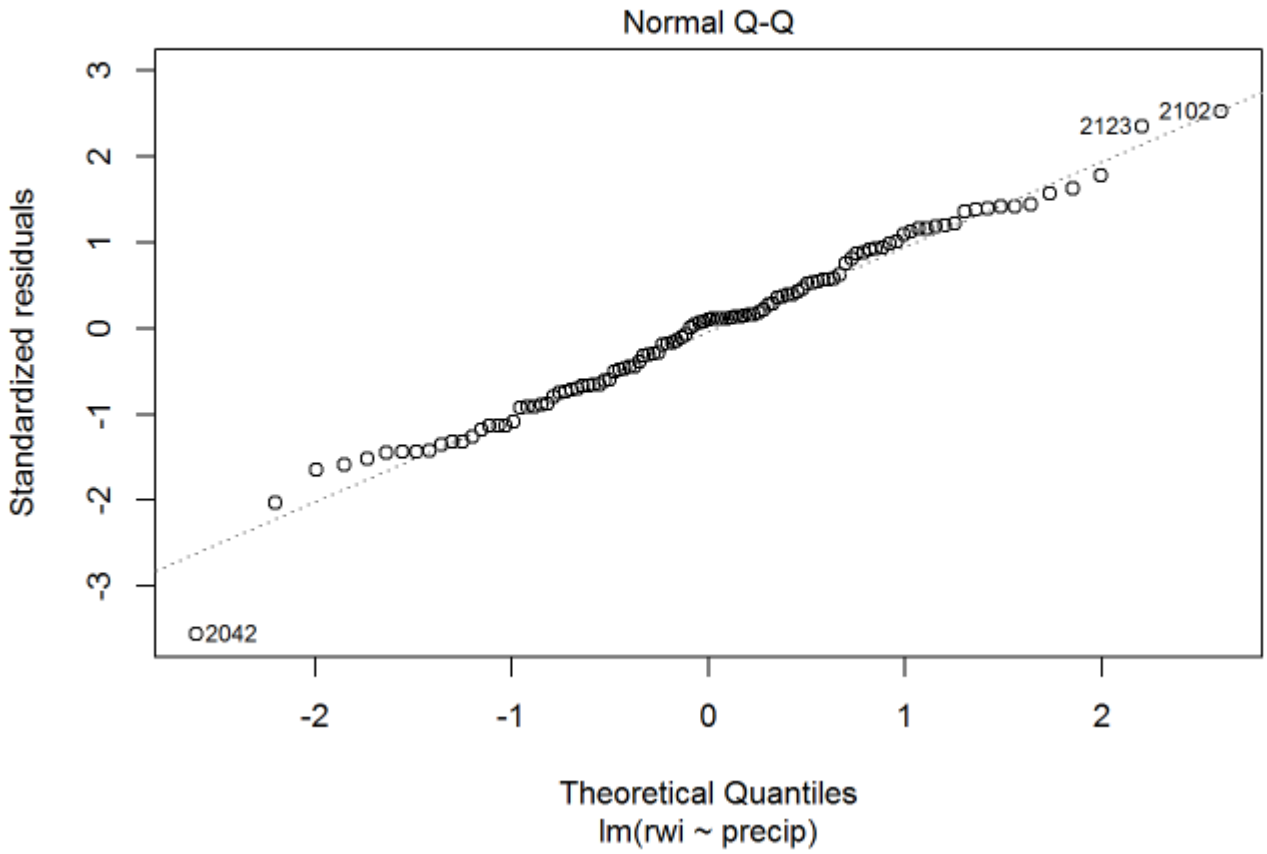


From the first diagnstic plot of the model where corresponding residuals are plotted against each fitted X value for every observation, we may see that the there is a positive correlation in the variables of the original regression model. The plot of residual points also show that there is a residuals could be predicted from one another, as in the case of time-series data in general, indicating to a amount of heteroscedasticity in the model. Ideally the points should be randomli distributed around the horizontal line, but in this case they follow a trend in the way they are distributed around the line.

The graph plots a best fit line among the residuals which is non-linear (shown in red), pointing that perhaps linear model is not the best solution for the data. The graph also shows, that there is a diminishing return of rainfall on tree ring growth after it reaches 20 inches a year. These observation points then work as outliers, which affects the effectibility of fit of a linear model to this data.

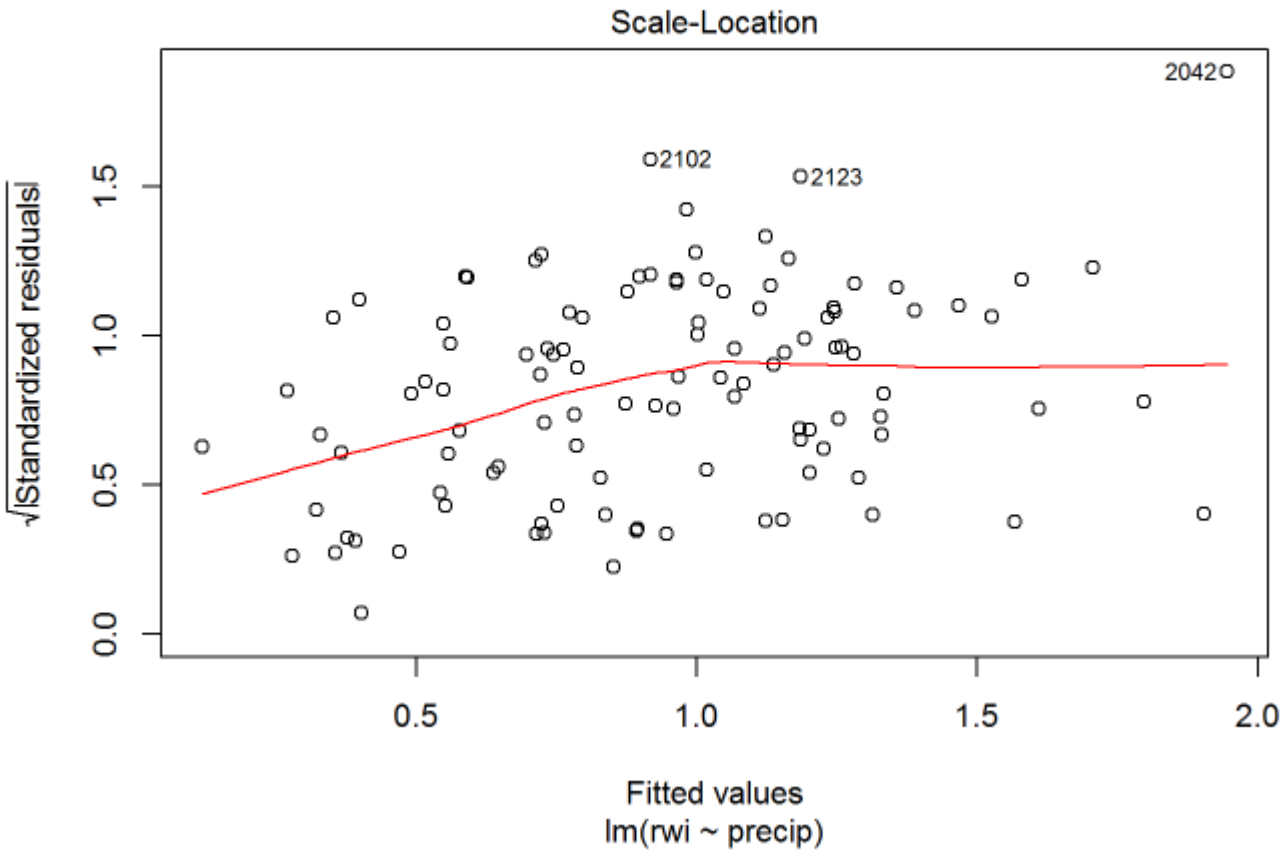
This helps to identify whether the residues are random and non-correlated, which they are not in this case.

```
plot(model, which=2)
```



The next diagnostic plot, the normal quantile - quantile plot, tries to fit a normal curve to the residues. The standardized residues are plotted aginst the theoritical quantiles, a normal curve is then tried to fit the data.To be a normal distribution, the data points should all have fallen on the line. But in case of our model, it may be seen from the graph that there is some variation from that norm. Also to be noted, that there is a pattern in the way the data points deviate from the line. There is a marked predictability among the residues. This helps to understand the resdues are not normally distrubuted.

```
plot(model, which=3)
```



The third diagnostic plot, the Scale-Location plot, shows the sqare root of the standardized errors plotted against the corresponding fitted values for every observation. We may see a marked pattern in the way the data points are distributed. This again points to the strong correlation among the data points with respect to time.

This again help to identify whether the reidues are random and non-corrlated. In this case, they are found to be correlated and non-random.

2. These data are time series data. In general, data are correlated in time, and this would be a problem for simple linear regression. Plot the regression residuals vs time, and visually assess whether there is evidence of temporal (serial) correlation.

```
model <- lm(rwi~precip, data=rwi.precip.df, na.action=na.exclude)
rwi.precip.df $ residual <- residuals(model)
tail(rwi.precip.df)
```

```
##      year    rwi samp.depth precip   residual
## 2136 1999  1.157         60  14.93  0.1983181
## 2137 2000  0.591         60  15.75 -0.4566188
## 2138 2001  0.913         60  15.48 -0.1053347
## 2139 2002 -0.033         59   9.77 -0.4320300
## 2140 2003  0.710         59  15.02 -0.2584432
## 2141 2004  0.419         53  12.87 -0.3162550
```

```
ggplot(data=rwi.precip.df, aes(x=year, y=residual)) + geom_point(shape=1) + geom_smooth(method=lm,
se=FALSE) + scale_x_continuous(limits=c(1890,2020)) + labs(y="Regression Residuals", x="Year", titl
e="Regression Residuals Vs. Time")
```

```
## Warning: Removed 2032 rows containing missing values (stat_smooth).
```

```
## Warning: Removed 2032 rows containing missing values (geom_point).
```

From the plot where regresion residuals have been plotted against the corresponding year, it may be seen that the points are distributed around the line in such a fashion that there is a correlation among the position of the points with respect to time. Generally if a point is above the line, the points preceeding or follwing that point is also above the line. This also holds true for the points falling below the line. This happens in a cycle, some points serially are above the line, then the next few dips below the line and then again the next set of points comes back up above the line. If there was no correlation, then the points would be scattered around the line just randomly and we would not have been able to find any such distribution pattern. This highlights a correlation among the residuals which is also known as

heteroscedasticity.

Thus, from the graph the temporal correlation among the data points (tree ring growth over the years) could be noted.

3. In the data folder is the dataset GalapagosData.txt. The species data represents the number of species recorded from each of the Galapagos islands. A fundamental 'law' of island biogeography is that species diversity tends to follow a power law relationship with island area, i.e.

$$\text{species} = \alpha \times \text{area}^{\beta}$$

This is not linear, but it suggests that the following regression might make sense:

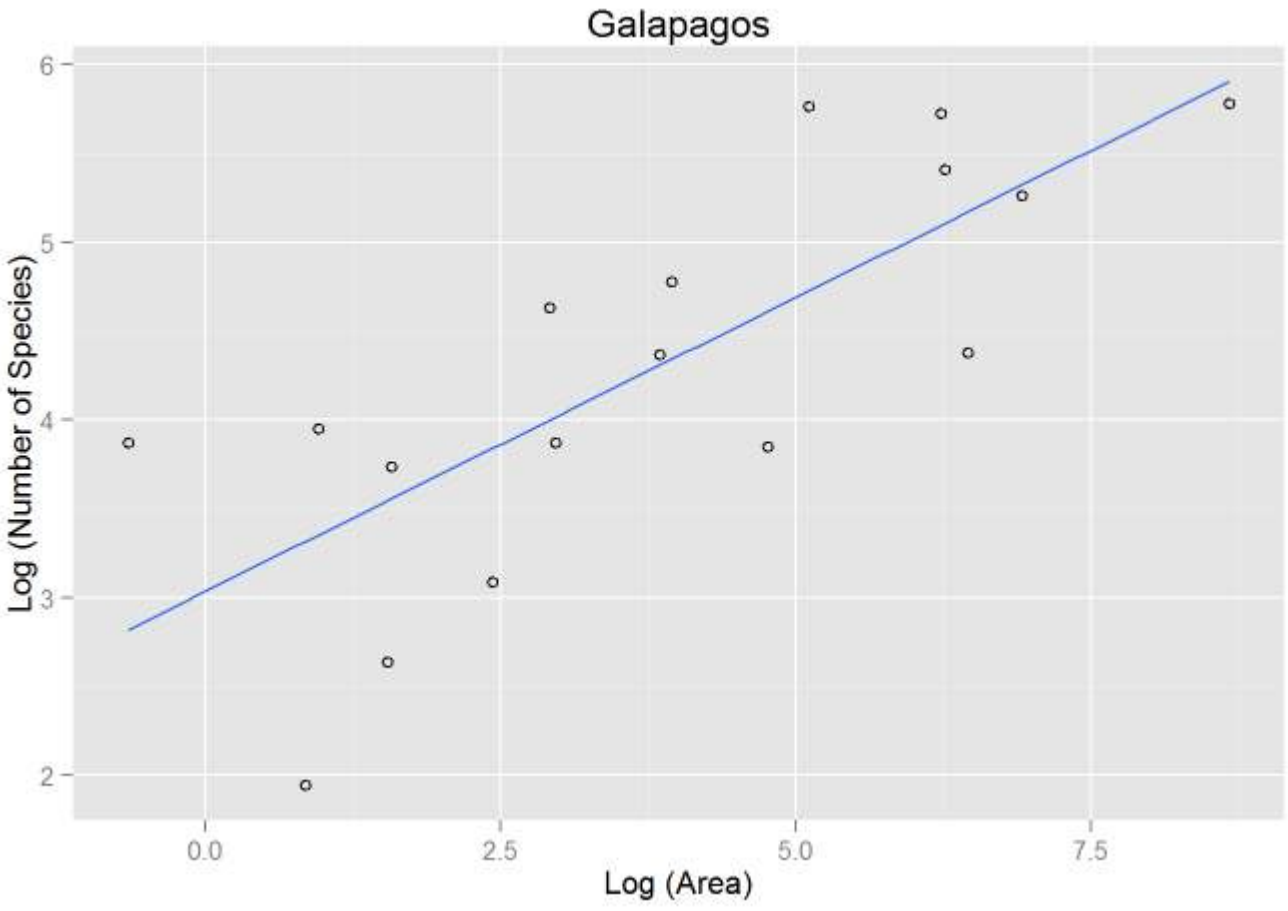
$$\log(\text{species}) = a + \beta \times log(\text{area})$$

a is not quite α, rather $a = \log(\alpha)$.
Fit this regression, and present a brief write-up that a) describes the results in words, and b) summarizes your conclusions from diagnostic model checking.

```
raw.galapagos <- read.table('GalapagosData.txt',sep='')
galapagos <- raw.galapagos
write.csv(galapagos,file='galapagos.csv', row.names=FALSE)
galapagosdata <- read.csv ('galapagos.csv')
galapagos.df <- galapagosdata %>% mutate(log_area = log(Area)) %>% mutate(log_species = log(Nspecie
s))
model1 <- lm(log_species~log_area, data=galapagos.df)
summary(model1)
```

```
##
## Call:
## lm(formula = log_species ~ log_area, data = galapagos.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.37282 -0.75233  0.06034  0.59768  1.04971
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.03895    0.32728   9.286 1.31e-07 ***
## log_area     0.33059    0.07194   4.595  0.00035 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7378 on 15 degrees of freedom
## Multiple R-squared:  0.5847, Adjusted R-squared:  0.557
## F-statistic: 21.12 on 1 and 15 DF,  p-value: 0.0003501
```

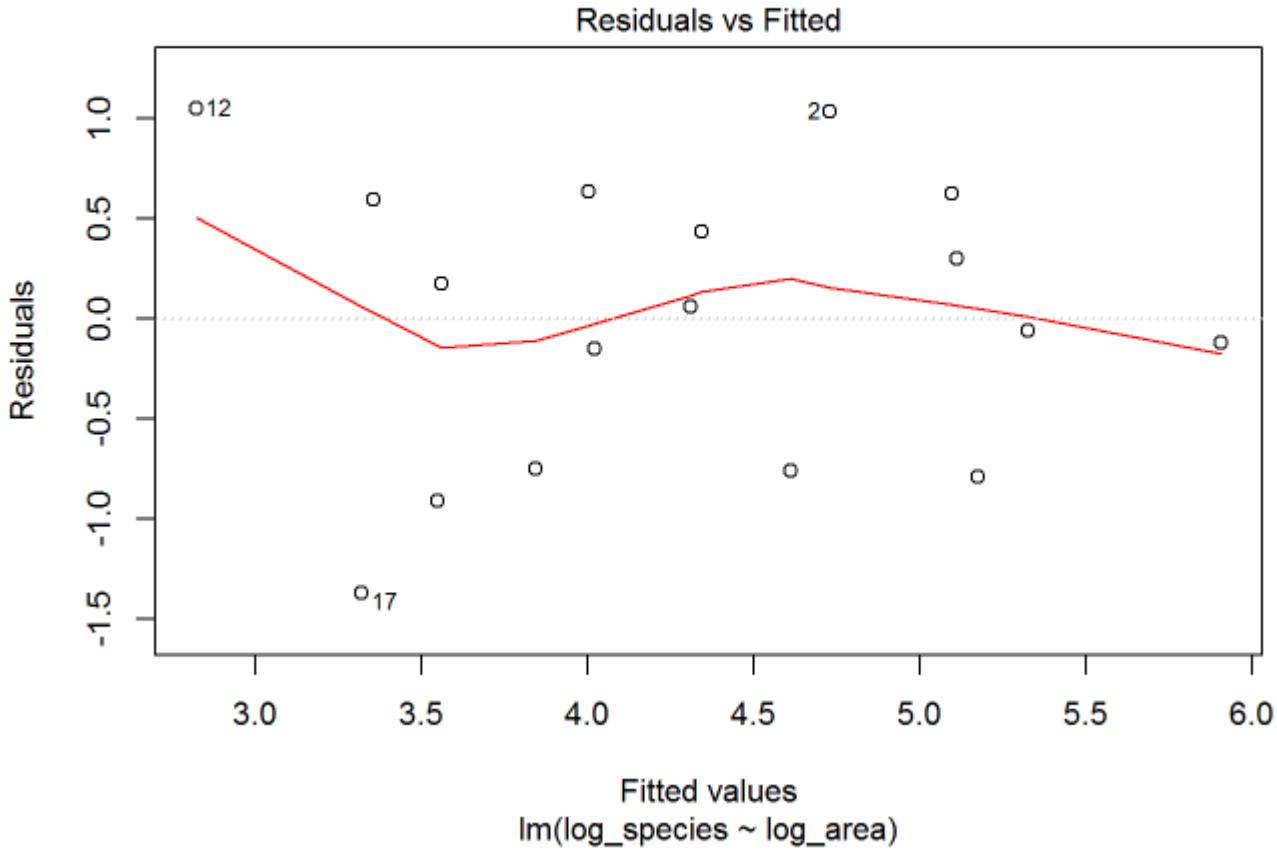
```
ggplot(data=galapagos.df, aes(x=log_area, y=log_species)) + geom_point(shape=1) + geom_smooth(metho
d=lm, se=FALSE) + labs(y="Log (Number of Species)", x="Log (Area)", title="Galapagos")
```



From the linear regression model, between log(species) and log(area) to highlight the fundamental law of Island Biogeography, we note the following summary:

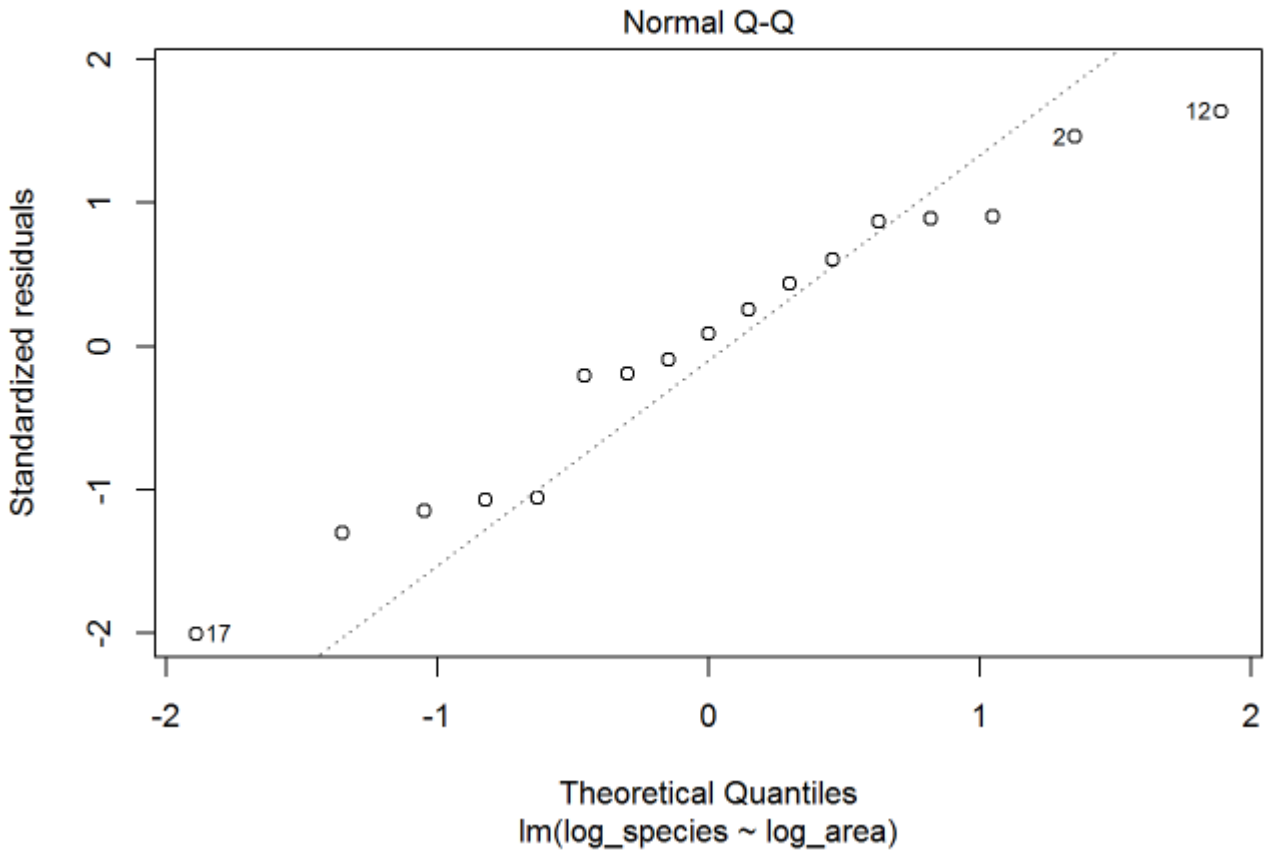
1. The maxium and the minimum difference between actual value and the predicted value of y variable (log of number of species) are 1.04971 and -1.37282 which are the two maximas of prediction error of the individual observations.
2. The model has an Intercept of 3.03895 on its Y axis and for every change of 1 unit along the X axis the corresponding change on Y is 3.03895.
3. The Residual Standard Error or the standard deviation of the residuals is 0.7378. This means most (68.27%) of the data points will occurs within +/- 0.07378 of the mean of Log(Species), and about 95.45% will fall within +/- 1.4756 of the mean.
4. The goodness of the fit has been expressed through the R Squared value of 0.5847 meaning about 58% of the variability in the log(species) could be explained by the predctor, log(area). The Adjusted R squared value, which gives the explanatory power of the model which is adjusted for the number of predictors, is 0.557. This means the model explains about 56% of the variability in Y variable, Log(species) in this case. The adjusted R squared measure is more robust to measurement accuracy by chance factors and may thus be considered as a better measue of goodness of fit.

```
plot(model1, which=1)
```



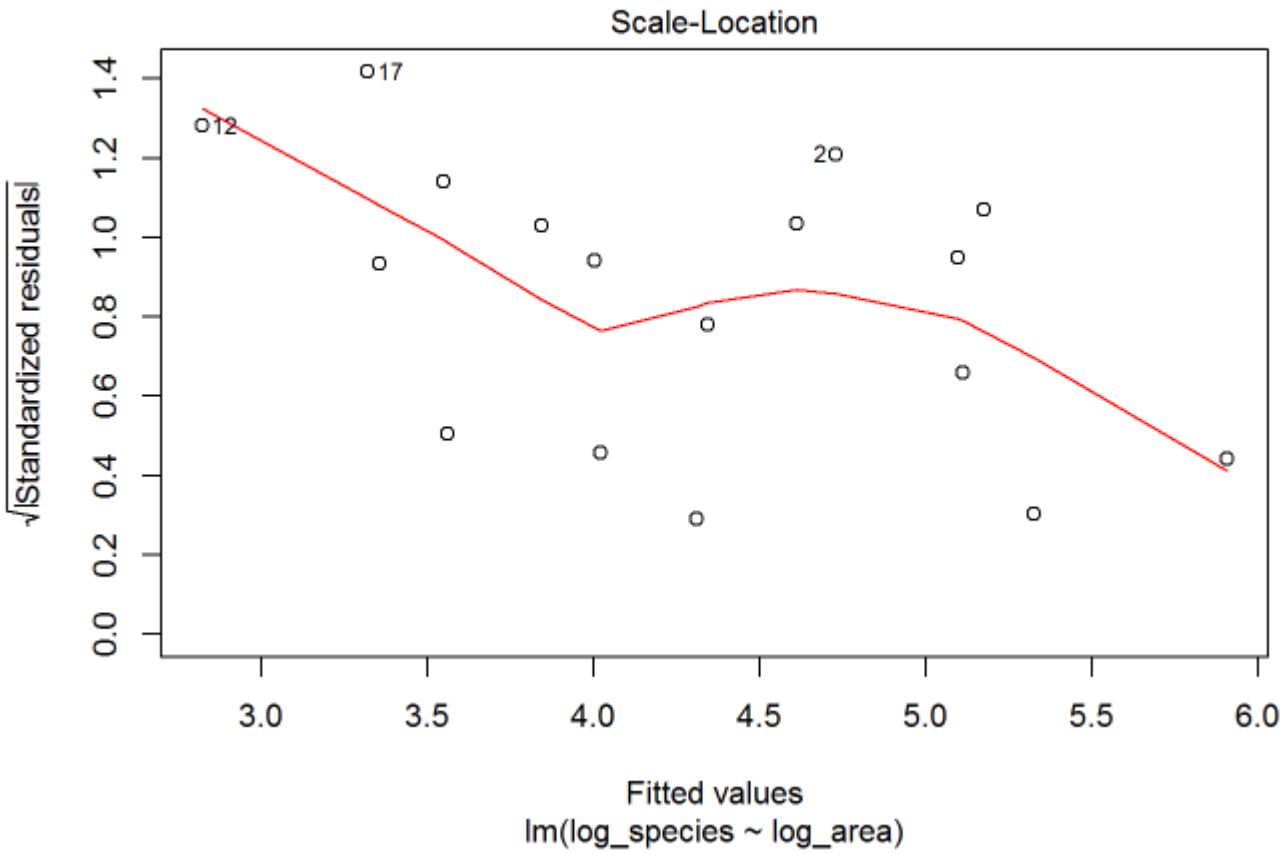
The first diagnostic plot of the model, Residual Vs. Fitted plot shows that the even thogh the points are somewhat randomly distributed about the horizontal line, the presence of ouliers at the lower end affects the fit of the model.These points come from two instances where high number of species are seen in a small island whereas on the other island a very low number of species were observed.

```
plot(model1, which=2)
```



The normal quantile quantile plot highlights that the residuals are perhaps normally distributed and there is probably a bimodal distribution. This has been highlighted by the distinct break in the points along the line.

```
plot(model1, which=3)
```



In the Scale-Location plot, there is no noted pattern in the way the points are distributed.

Thus, from the model and the diagnostic tests, we may say thst the fundumental law of Biogegraphy is holding true for the given dataset.