Lesson: Introduction to Linear Regression in R

Introduction

The aim of linear regression is to model a continuous variable Y as a mathematical function of one or many variable(s), so that we can use this regression model to predict the Y when only the X is known. This mathematical equation can be generalized as follows:

$$Y = \beta 1 + \beta 2X + \epsilon$$

where, β1 is the intercept and β2 is the slope. Collectively, they are called regression coefficients. ϵ is the error term, the part of Y the regression model is unable to explain.

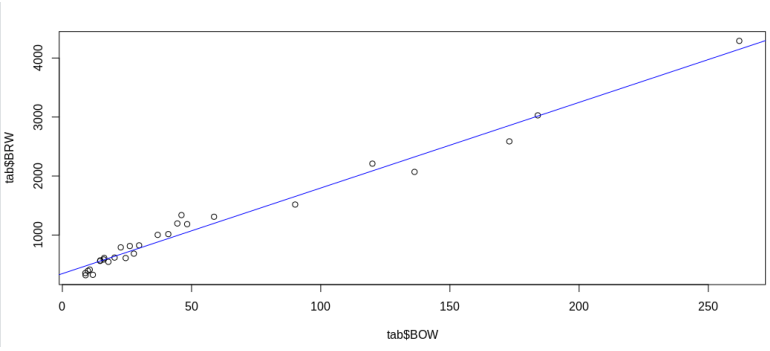
However the regression only makes sense if the variables that you started with are on the same scale . The regression is only useful if there is a relation or a correlation between 2 variables and there is some statistical procedure to find that.

In this lesson, we will learn how to use a linear regression in R and to conclude some information about a dataset.

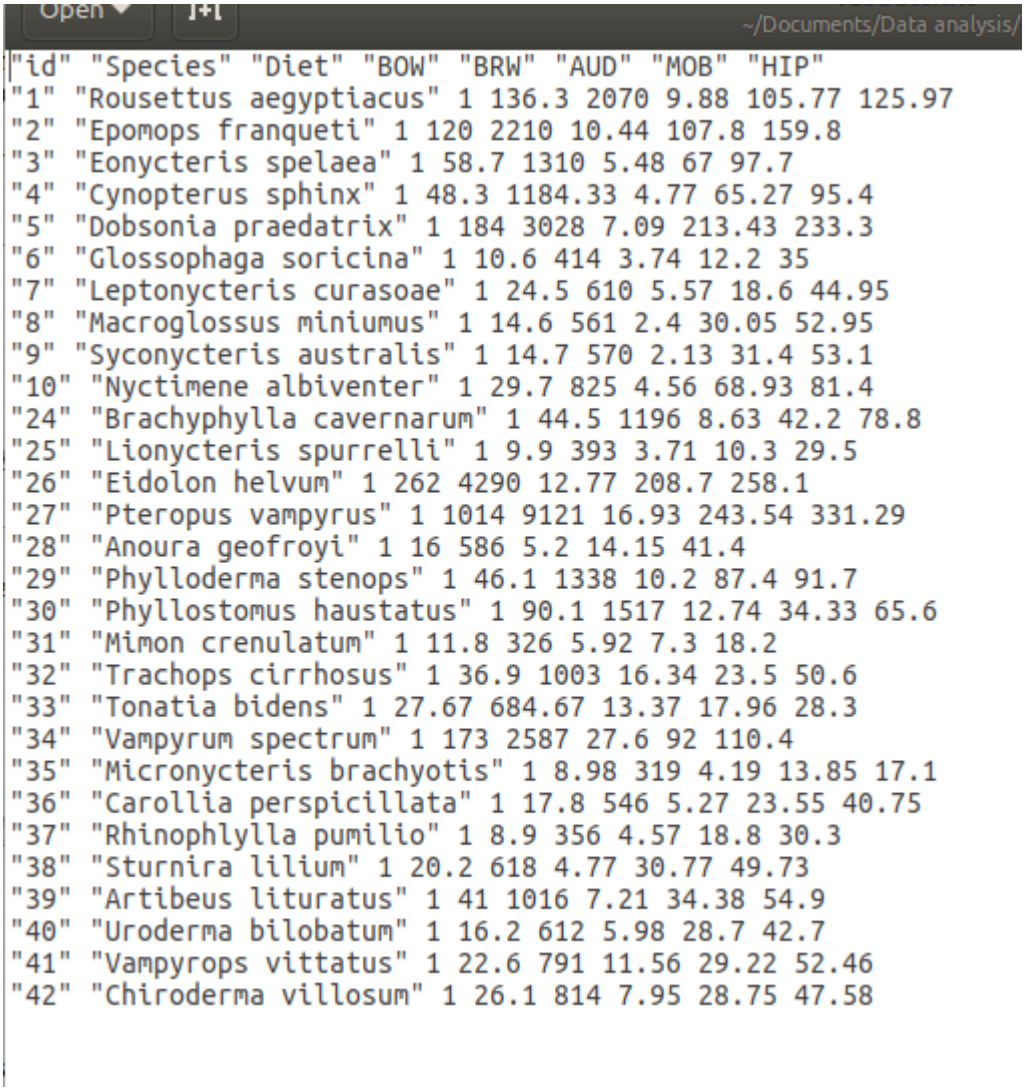
Using a linear model in a study can be decomposed into several steps :

- Find a tool to explore the dataset and to use the linear model regression
- Clean the dataset and keep only the useful data
- Look for some correlation between dataset's variables
- Make some hypothesis
- Run the linear model
- Correct the model if necessary
- Make some conclusion on the model and his relevance.

The picture bellow shows a good example of the linear regression model (blue)



In this example, we are going to use the following dataset:

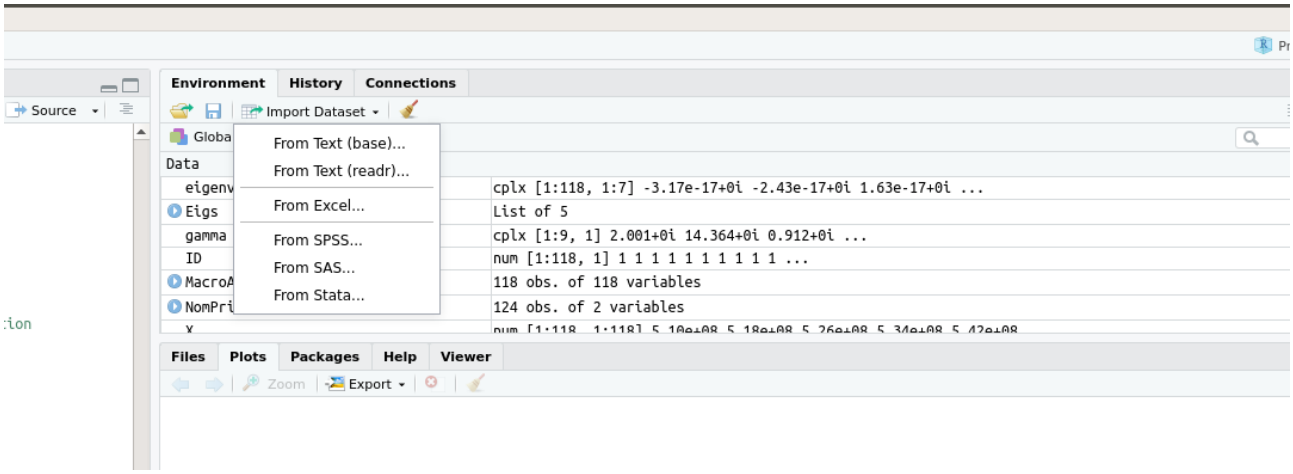



```
"id" "Species" "Diet" "BOW" "BRW" "AUD" "MOB" "HIP"
"1" "Rousettus aegyptiacus" 1 136.3 2070 9.88 105.77 125.97
"2" "Epomops franqueti" 1 120 2210 10.44 107.8 159.8
"3" "Eonycteris spelaea" 1 58.7 1310 5.48 67 97.7
"4" "Cynopterus sphinx" 1 48.3 1184.33 4.77 65.27 95.4
"5" "Dobsonia praedatrix" 1 184 3028 7.09 213.43 233.3
"6" "Glossophaga soricina" 1 10.6 414 3.74 12.2 35
"7" "Leptonycteris curasoae" 1 24.5 610 5.57 18.6 44.95
"8" "Macroglossus miniumus" 1 14.6 561 2.4 30.05 52.95
"9" "Syconycteris australis" 1 14.7 570 2.13 31.4 53.1
"10" "Nyctimene albiventer" 1 29.7 825 4.56 68.93 81.4
"24" "Brachyphylla cavernarum" 1 44.5 1196 8.63 42.2 78.8
"25" "Lionycteris spurrelli" 1 9.9 393 3.71 10.3 29.5
"26" "Eidolon helvum" 1 262 4290 12.77 208.7 258.1
"27" "Pteropus vampyrus" 1 1014 9121 16.93 243.54 331.29
"28" "Anoura geofroyi" 1 16 586 5.2 14.15 41.4
"29" "Phylloderma stenops" 1 46.1 1338 10.2 87.4 91.7
"30" "Phyllostomus haustatus" 1 90.1 1517 12.74 34.33 65.6
"31" "Mimon crenulatum" 1 11.8 326 5.92 7.3 18.2
"32" "Trachops cirrhosus" 1 36.9 1003 16.34 23.5 50.6
"33" "Tonatia bidens" 1 27.67 684.67 13.37 17.96 28.3
"34" "Vampyrum spectrum" 1 173 2587 27.6 92 110.4
"35" "Micronycteris brachyotis" 1 8.98 319 4.19 13.85 17.1
"36" "Carollia perspicillata" 1 17.8 546 5.27 23.55 40.75
"37" "Rhinophlylla pumilio" 1 8.9 356 4.57 18.8 30.3
"38" "Sturnira lilium" 1 20.2 618 4.77 30.77 49.73
"39" "Artibeus lituratus" 1 41 1016 7.21 34.38 54.9
"40" "Uroderma bilobatum" 1 16.2 612 5.98 28.7 42.7
"41" "Vampyrops vittatus" 1 22.6 791 11.56 29.22 52.46
"42" "Chiroderma villosum" 1 26.1 814 7.95 28.75 47.58
```

This dataset is composed of 7 variables :Diet, body masses (BOW), brain masses (BRW) and volumes of three brain regions (main olfactory bulb MOB, hippocampus HIP, auditory nuclei AUD) for 29 bats. In this example, we are trying to find a link between all these variables.

Step 1: Loading the dataset

First you have to install Rstudio and then import the dataset.



This will copy the dataset in a variable called “Tabbats”. The same effect can be obtained by running the command (shown in terminal window):

```
Tabbats <- read.csv("Tabbats.txt", sep="")
```

Step 2: Clean the dataset

In this step, you have to remove the data not useful for a regression linear model with the following command :

```
str(Tabbats)
TabBats=Tabbats[,(-1:-3)]
str(TabBats)
```

We removed 3 attributes:

- The ID because it’s irrelevant;
- Species because we cannot do a correlation within
- Diet, because they are all phytophagy(1).

To do that we use the following command:str(Tabbats)TabBats=Tabbats[,(-1:-3)] which shifts the array variable by 2 position and copy it into a new variable called “TabBats” (R variable names are case sensitive).

The result is a new table TabBats without the columns 1 to 3 of the table Tabbats. We are left with 5 variables: BOW, BRW, AUD. MOB and HIP.

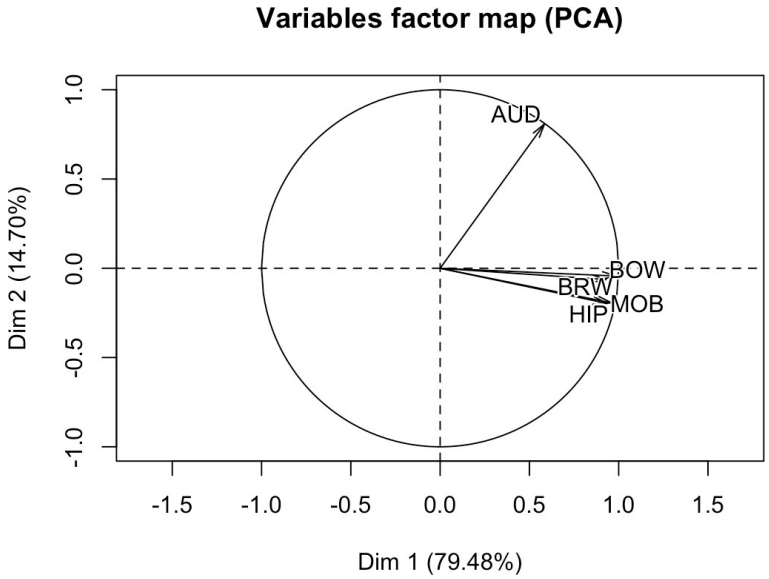
Step 3: Look for some correlation between dataset’s variables

Before use the linear regression, you have to look if the variables have a link between themselves. If the variables are not correlated, it’s not useful to do a linear regression. For that use the principals component analysis (PCA) method which creates a Variables factor map. The Variables factor map presents a view of the projection of the observed variables into the plane spanned by the first two principal components. This shows us the structural relationship between the variables and the components, and helps us name the components. The projection of a variable vector onto the component axis allows us to directly read the correlation between the variable and the component.

To do that in R, you need to install the library FactoMineR and use the command result ← PCA(dataframe) to show the circle of revolution. The following commands need to be excuted:

```
install.packages("FactoMineR")
library(FactoMineR)
result <- PCA(TabBats)
```

This should produce the graph below:



On this graph we note that:

- The axes of BOW and BWR are almost aligned;
- The axes of HIP and MOB are almost aligned;
- The four axes BOW, BWR, HIP and MOB are close.

So from the graph results:

- BOW and BWR are highly correlated;
- HIP and MOB are very correlated but less than the two attributes below;
- BOW, BWR, HIP and MOB are also correlated.

To see the content of our variable “result”, we use the command print(result) which lists the information confirming the graph hypothesis

```
print(result)
```

We find some interesting information:

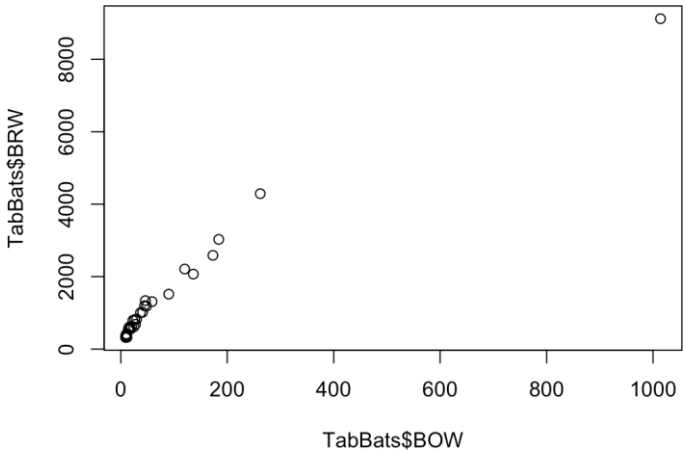
- The eigenvalues;
- The coordinates for the variables;
- The correlations variables – dimensions;
- The contributions of the variable.

Step 4: Find the equation between variables BOW and BRW

We now take the 2 attributes which are the most correlated (TabBats$BRW,TabBats$BOW) and use the command “plot” to show the link between the 2 attributes in a graph. Interpret the result.

```
plot(TabBats$BOW,TabBats$BRW)
```

We can see a scatter graph which links the two variables BOW and BRW with the plot(TabBats$BOW,TabBats$BRW) command:



We can see there is a kind of linear link between the two variables. Almost all the individuals verifed this tendency (we got one point isolated in the right top corner of the graph).
The assumed equation of this model is : Y = B1X +B0

Using the command lm(attribut21~attribute2), we can find the coefficient of the linear regression and plot this regression. And write the equation of this model.

To know the coefficients of the equation of the regression model, we use the following command:

```
mod=lm(TabBats$BRW~TabBats$BOW)
print(mod)
```

We get the Equation:y = 9 x + 623.4. This linear regression is very close of our variables. It confirms that BOW and BRW are correlated.

The command summary will give you more information on the linear regression as the residuals , coefficients, standard error… With the help of the commad plot, summary and abline(coef(mod),col='red'), interpret this information and make a hypothesis for improving the model.

```
plot(TabBats$BOW,TabBats$BRW)
abline(coef(mod),col='red')
summary(mod)
```

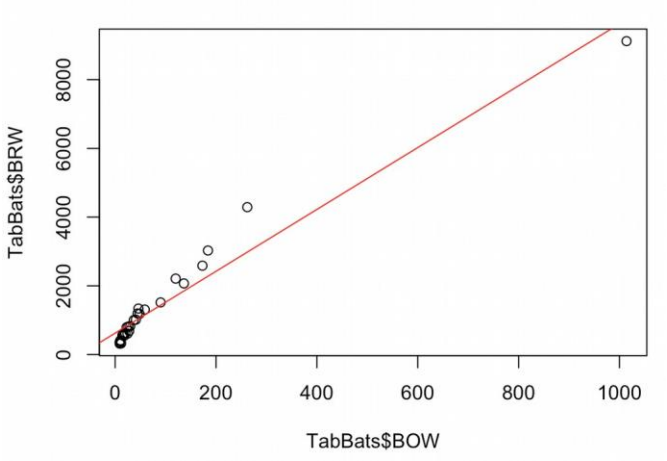
The following summary is obtained:

```
Residuals:
    Min      1Q  Median      3Q     Max
-628.32 -233.94  -65.74  158.26 1308.59

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 623.4469    81.4762   7.652 3.14e-08 ***
TabBats$BOW   8.9999     0.3972  22.659  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 396.9 on 27 degrees of freedom
Multiple R-squared:   0.95,    Adjusted R-squared:  0.9482
F-statistic: 513.4 on 1 and 27 DF,  p-value: < 2.2e-16
```

Step 5: Correct the model

The residual standard error is high which means that our data model isn't right or some points are disturbing it. The p-value is almost equal to 0, which gives us a first clue of the close relation between BRW and BOW. The R-squared value is also very high (0,95 on a scale from 0 to 1) which confirm that we got a strong correlation between our two variables.

In conclusion, the residual standard error is too high; we can't trust our regression model.



On this diagram we can see the linear regression of our dataset going through all our points. As we tell below, we should delete points to reduce the residual standard error in order to have a better regression model.

9- Apply your hypothesis and modify your dataframe and make a new conclusion about the 2 attributes

```
tab=TabBats[-14,]
plot(tab$BOW,tab$BRW)
mod2=lm(tab$BRW~tab$BOW)
summary(mod2)

plot(TabBats$BOW,TabBats$BRW)
abline(coef(mod),col='red')
abline(coef(mod2),col='blue')
```

Exercise

Repeat the analysis for variables HIP and MOB.