batman

Benjamin Cathelineau

09/12/2021

Presentation

```
myData <- read.table(file = "bats.csv",sep = ";",skip = 3,header = T)
names(myData)
```

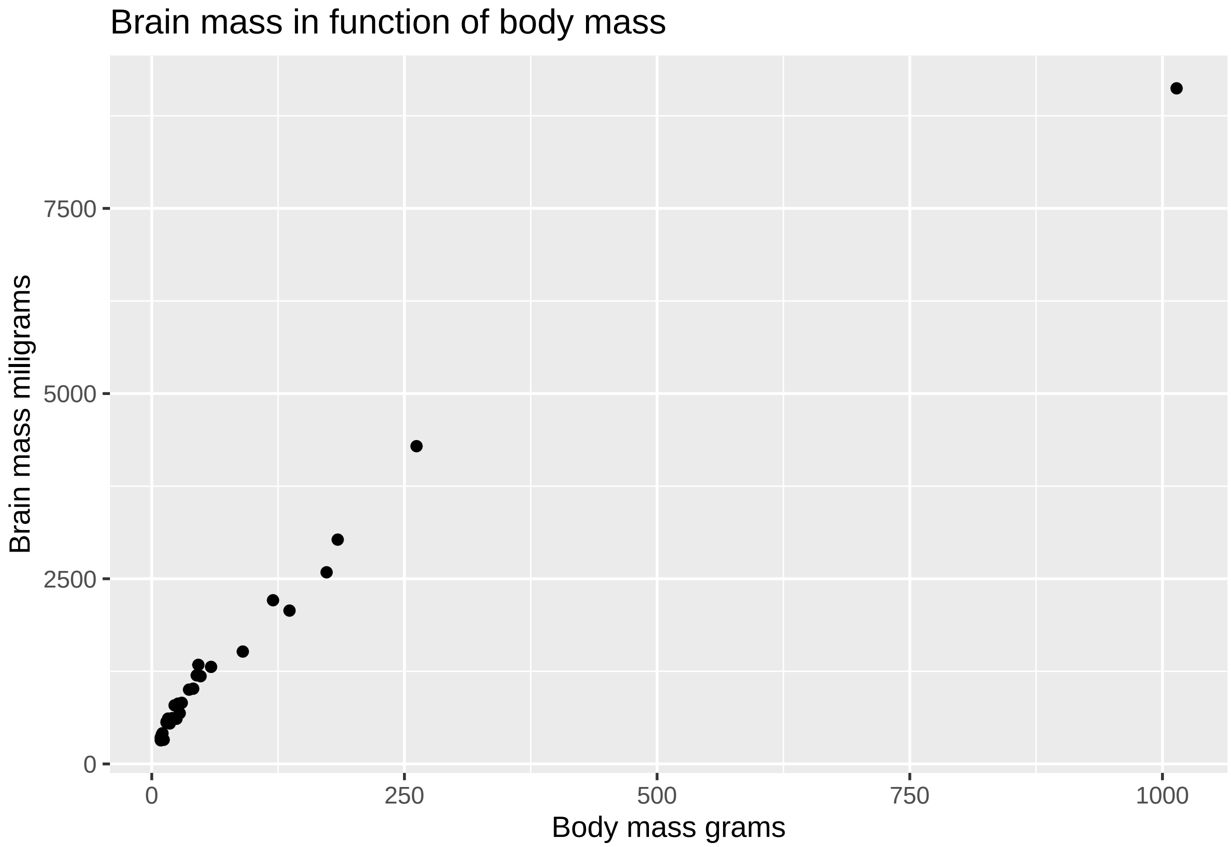
```
## [1] "Species" "Diet"    "Clade"   "BOW"     "BRW"     "AUD"     "MOB"
## [8] "HIP"
```

Study of the relationship between brain weight and body masss

Brain mass, body mass relation in phytophagous

```
phyto = myData[(myData$Diet==1),]
```

```
library(ggplot2)
ggplot(phyto, aes(x=BOW,y=BRW))+
  geom_point()+
  ggtitle("Brain mass in function of body mass")+
  xlab("Body mass grams")+
  ylab("Brain mass miligrams")
```



Linear regression

```
regression=lm(BRW~BOW,data=phyto)
regression
```

```
##
## Call:
## lm(formula = BRW ~ BOW, data = phyto)
##
## Coefficients:
## (Intercept)          BOW
##       623.4          9.0
```

In mathematical form : $brw = 623.4 + 9 \times bow$

```
summary(regression)
```

```
##
## Call:
## lm(formula = BRW ~ BOW, data = phyto)
##
## 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 ***
## 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
```

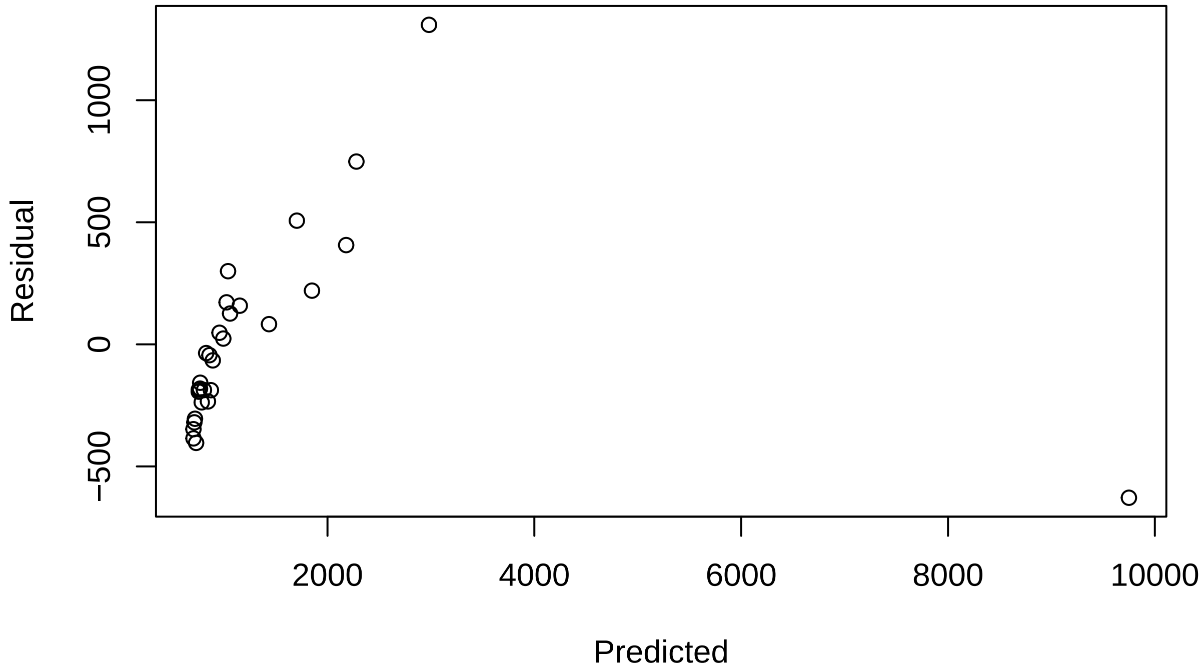
This model is of very high quality, the $p - values$ are very low for both the intercept and *bow*. R^2 is very close to one The $H0$ hypothesis is to test if the body mass has no impact on the brain weight. For $H0$ to be true the multiplicator, currenlty equal to 9 would need to be very close to 0 In other words, the relation between brain weight and body mass is very clear. ### Analysis of variance

```
anova(regression)
```

```
## Analysis of Variance Table
##
## Response: BRW
##           Df   Sum Sq  Mean Sq F value    Pr(>F)
## BOW        1 80888380 80888380  513.42 < 2.2e-16 ***
## Residuals 27  4253838   157550
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this table we have additional information such as the sum of residual squares. This is the sum of the difference between the prediction (from the model) and the empirical values, each of theses value being squared.

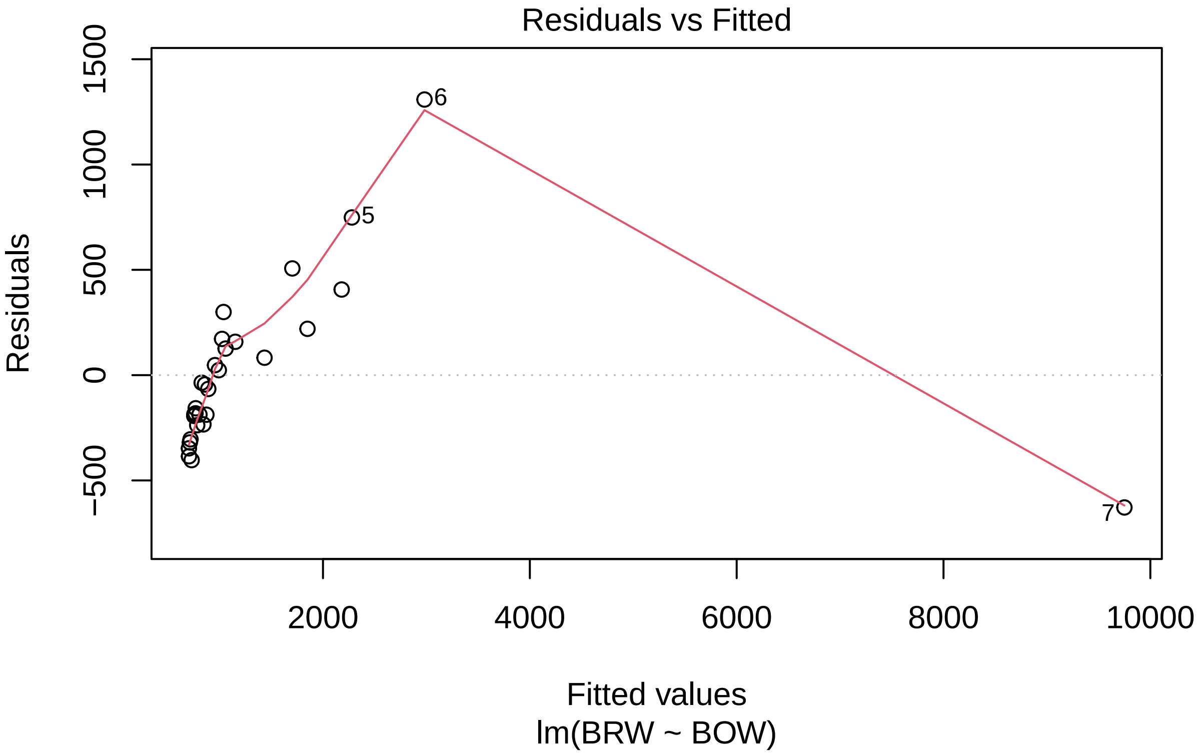
```
plot(regression$fitted.values,regression$residuals,xlab="Predicted",ylab="Residual")
```

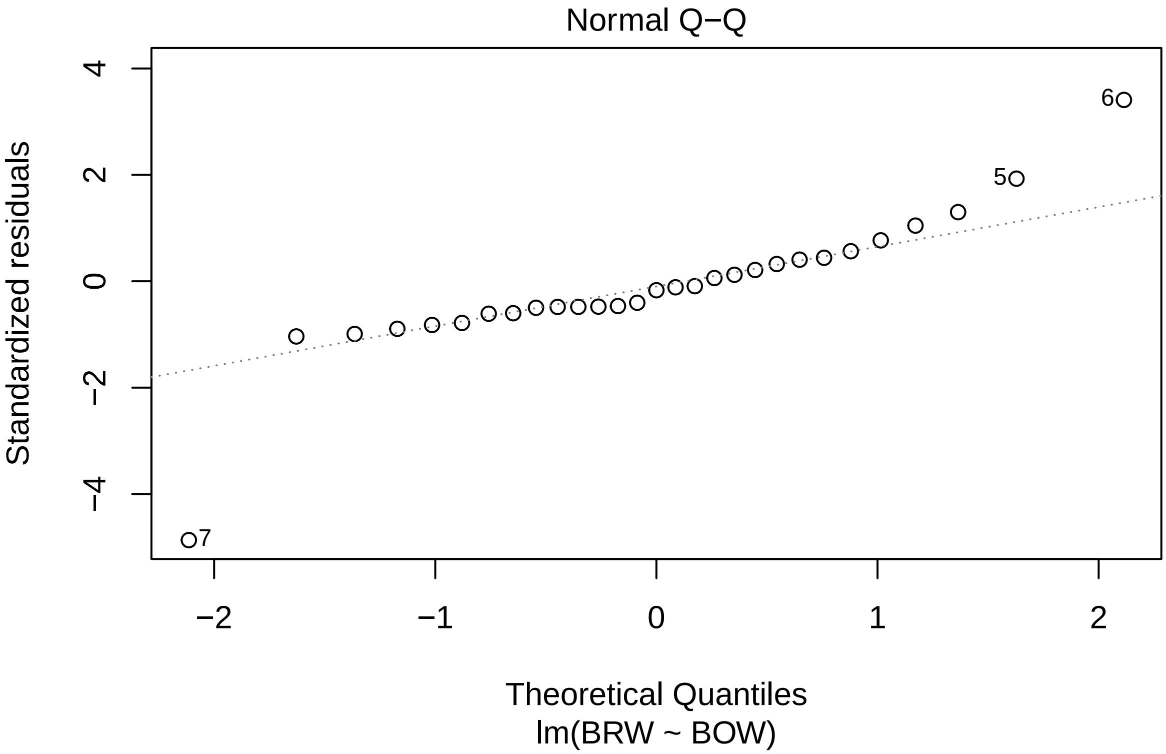


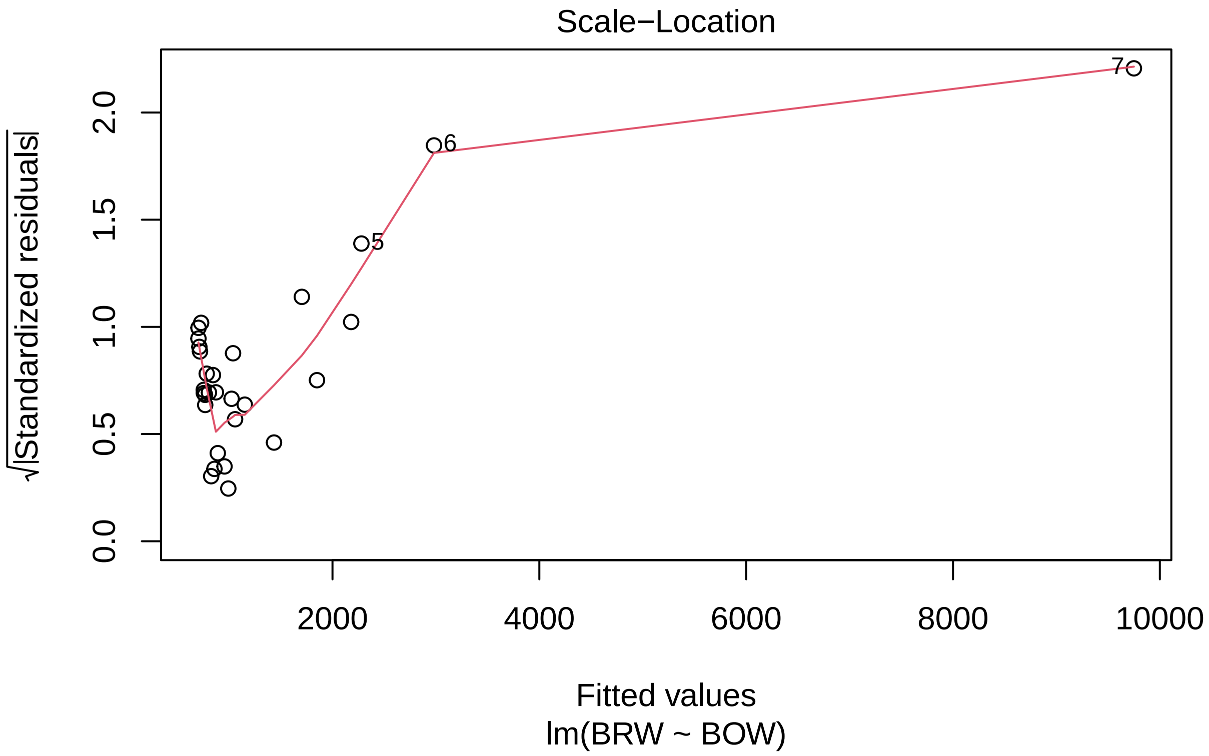
Most of the point are concentrated in the [-500;0] residual, [0;1000] predicted region. From this graph, ignoring the outlier at 10000 predicted, it is apparent that our model tend to get smaller prediction slightly wrong in the small direction and larger prediction slightly wrong in the large direction. In other words, there seems to be some sort of pattern.

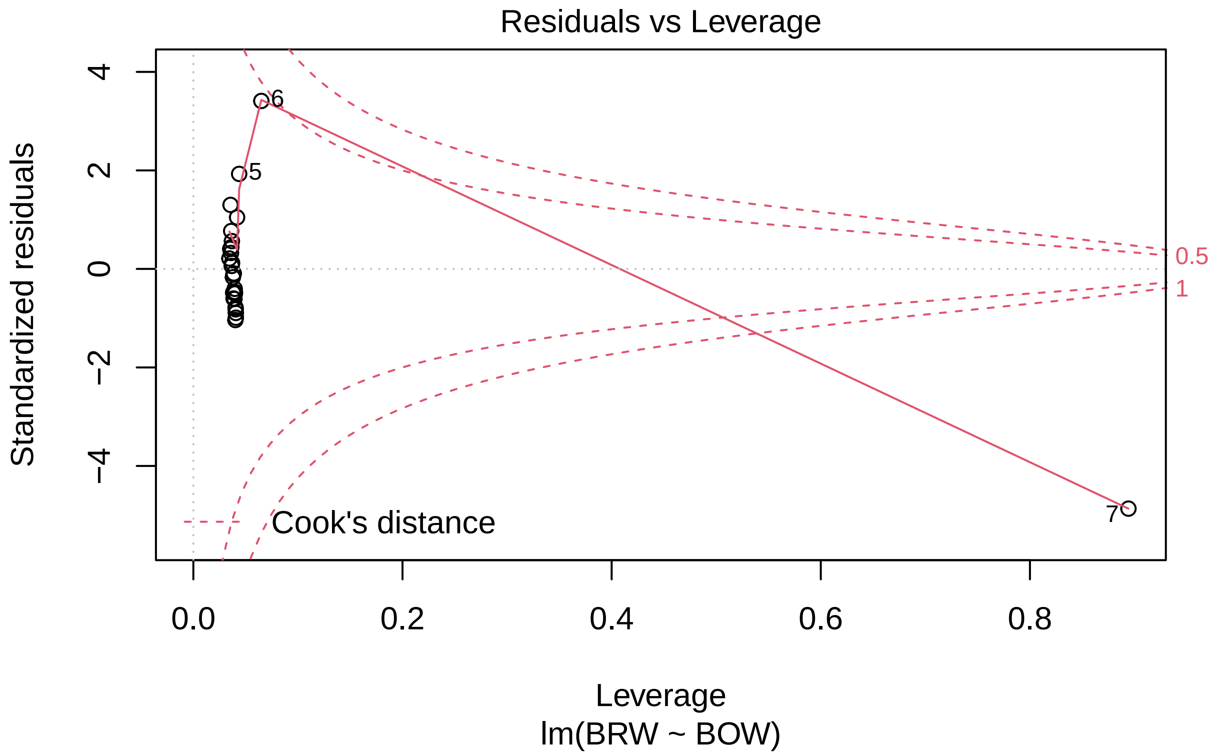
We can also note that there is more individual of lower weight than of higher weight. In other words, the data is not balanced.

```
plot(regression)
```


Residuals vs Fitted
Residuals
1500
1000
500
0
−500
6
5
7
2000
4000
6000
8000
10000
Fitted values
lm(BRW ~ BOW)


Normal Q−Q
Standardized residuals
4
2
0
−2
−4
6
5
7
−2
−1
0
1
2
Theoretical Quantiles
lm(BRW ~ BOW)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2000
4000
6000
8000
10000
Fitted values
lm(BRW ~ BOW)
7
6
5



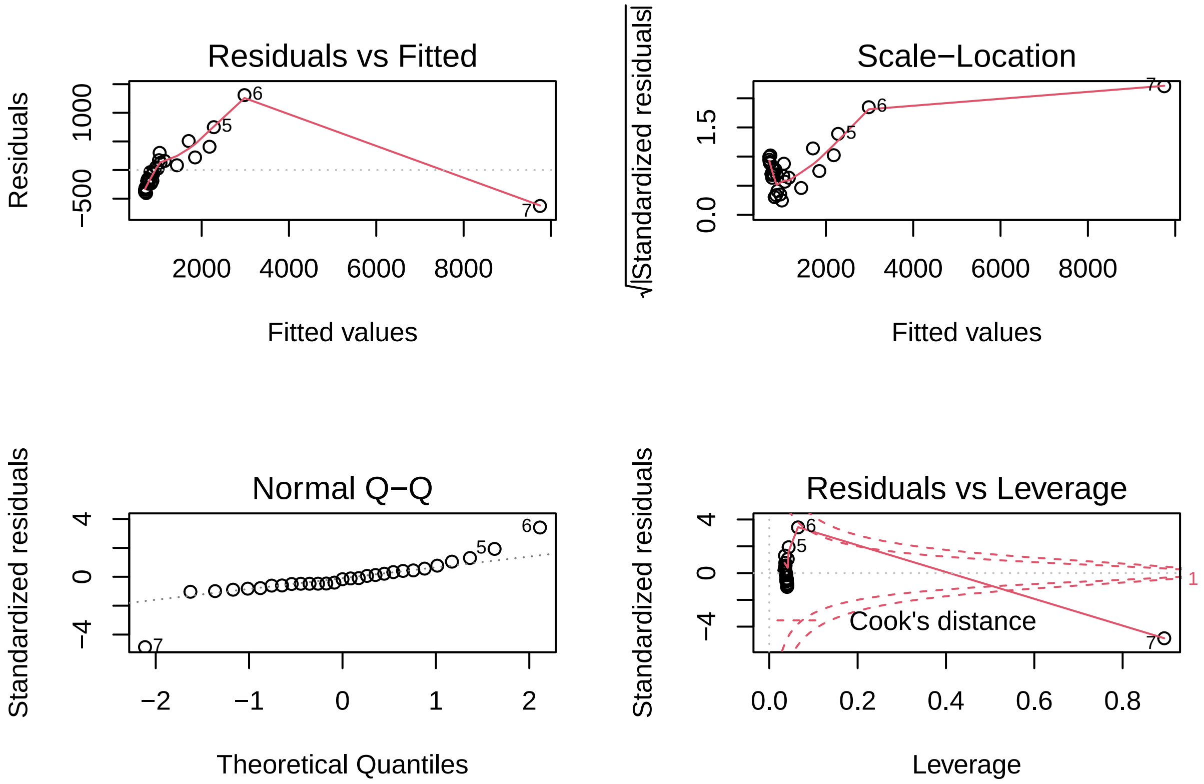
Now we remove the problematic individual “batman”.

```
phytobis=phyto[which(phyto$BRW<8000),]
regressionbis=lm(BRW ~ BOW,data=phytobis)
summary(regressionbis)
```

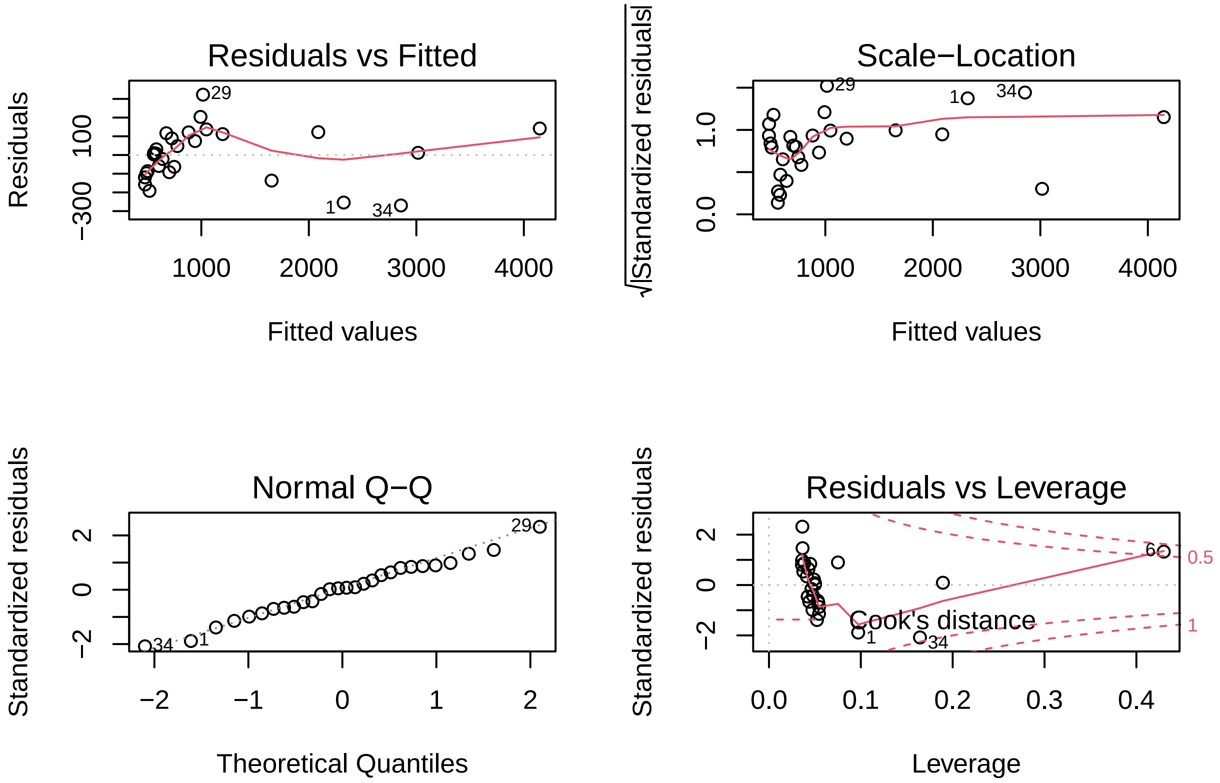
```
##
## Call:
## lm(formula = BRW ~ BOW, data = phytobis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -269.76  -93.33    8.73  112.93  322.55
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 346.5452    35.4920   9.764 3.48e-10 ***
## BOW          14.5099     0.4285  33.860  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 141.8 on 26 degrees of freedom
## Multiple R-squared:  0.9778, Adjusted R-squared:  0.977
## F-statistic:  1147 on 1 and 26 DF,  p-value: < 2.2e-16
```

We obtain a very different model as soon as we remove batman. Now the mathematical formula looks like $brw = 346.54 + 14.5 \times bow$. This show that the batman had a massive impact on the model, which might makes us think that it is likely that the batman was an error in the data.

```
par(mfcol=c(2,2))
plot(regression)
```



```
plot(regressionbis)
```



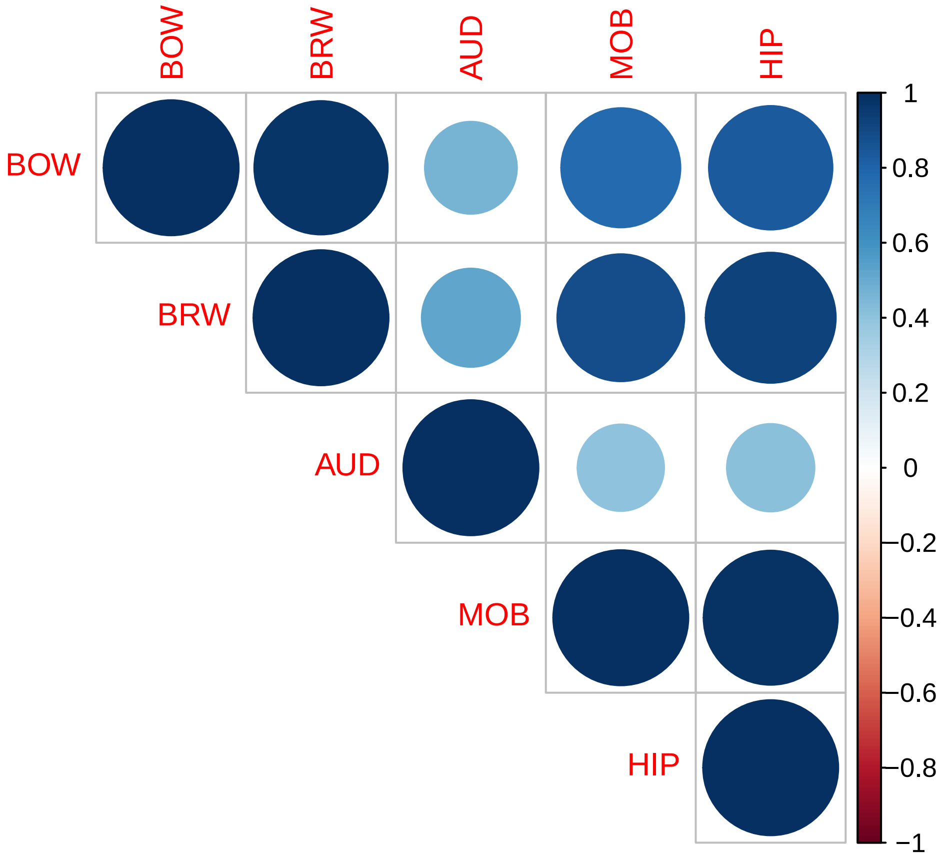
The validity plots look much better with the second model where the batman is removed. Notably, the Normal Q-Q fits a normal distribution almost perfectly, as opposed to before. The other plot look better as well, I don't understand them perfectly but they look shrinked for the most part.

Study of the contribution to the total weight of each part of the brain

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
phytoNum=phyto[,c(4:8)]
mat.cor=cor(phytoNum)
corrplot(mat.cor, type= "upper")
```



From this nice plot we see that all variable are perfectly correlated with themselves (which make sense).

More interestingly, we see the previously studied brain mass, body mass correlation. We also see a interesting olfactory zone volume (MOB), volume of the hipocampus (HIP) correlation.

Pearson tests

```
cor.test(phyto$BRW,phyto$HIP)
```

```
##
##  Pearson's product-moment correlation
##
## data:  phyto$BRW and phyto$HIP
## t = 12.91, df = 27, p-value = 4.574e-13
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8502663 0.9658107
## sample estimates:
##       cor
## 0.9276811
```

```
cor.test(phyto$BRW,phyto$MOB)
```

```
##
##  Pearson's product-moment correlation
##
## data:  phyto$BRW and phyto$MOB
## t = 9.7964, df = 27, p-value = 2.203e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7644185 0.9442114
## sample estimates:
##       cor
## 0.8834215
```

```
cor.test(phyto$BRW,phyto$AUD)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  phyto$BRW and phyto$AUD
## t = 3.2338, df = 27, p-value = 0.003215
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2007495 0.7497021
## sample estimates:
##       cor
## 0.5283792
```

The brain weight is highly correlated with the olfactory zone volume (MOB) and with the volume of the hipocampus (HIP) but not with the auditory part of the brain (AUD). As stated before, HIP and MOB are themselves correlated together, which explains that BRW is correlated with both.

Regression model

```
regm=lm(BRW~AUD+MOB+HIP,data=phytobis)
summary(regm)
```

```
##
## Call:
## lm(formula = BRW ~ AUD + MOB + HIP, data = phytobis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -268.55  -68.84    9.88   61.66  375.34
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -312.692     76.628  -4.081  0.00043 ***
## AUD           47.989      6.067   7.910 3.85e-08 ***
## MOB           -2.444      3.257  -0.750  0.46034
## HIP           15.981      2.960   5.399 1.52e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 158.5 on 24 degrees of freedom
## Multiple R-squared:  0.9744, Adjusted R-squared:  0.9712
## F-statistic: 304.5 on 3 and 24 DF,  p-value: < 2.2e-16
```

```
anova(regm)
```

```
## Analysis of Variance Table
##
## Response: BRW
##           Df   Sum Sq  Mean Sq F value    Pr(>F)    
## AUD        1  6817133  6817133 271.210 1.397e-14 ***
## MOB        1 15409397 15409397 613.040 < 2.2e-16 ***
## HIP        1   732653   732653  29.148 1.519e-05 ***
## Residuals 24   603265    25136                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$brw = -312.692 + 47,989 \times aud + -2.444 \times mob + 15,981 \times hip$

From the ANOVA result we see that this model is of very high quality, as R^2 is close to 1. Most coefficient of the model are good and reliable, at the exception of of MOB, which has a way to high p value, as well as a low coefficient (-2.444) which does not impact the model a lot.

Removing higly correlated variable

An hypothesis to explain the fact that MOB is not well integrated in our model is that it is highly correlated with HIP, which is already a part of our model. HIP and MOB are collinear. We can use the previously mentioned pearson test to check this:

```
cor.test(phyto$MOB,phyto$HIP)
```

```
## 
##  Pearson's product-moment correlation
## 
## data:  phyto$MOB and phyto$HIP
## t = 30.297, df = 27, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9692151 0.9933030
## sample estimates:
##       cor 
## 0.9856097
```

Theses 2 variable are indeed extremely correlated. Therefore, we should remove one of them from the model.