batman

Benjamin Cathelineau

09/12/2021

Presentation

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

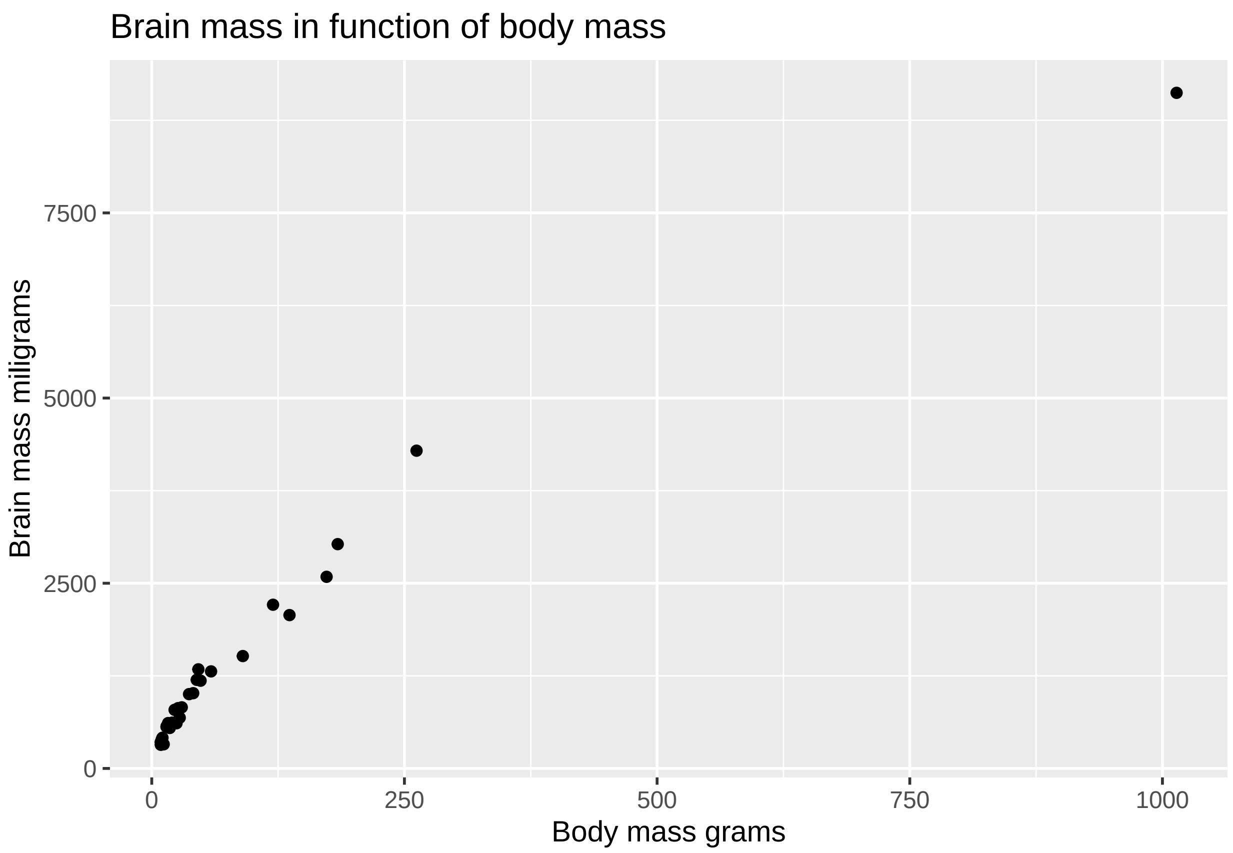
```
## [1] "Species" "Diet"    "Clade"   "BOW"     "BRW"     "AUD"     "MOB"
## [8] "HIP"
```

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, the coefficient of determination, is very close to 1, which is good. F-Statistic, $\frac{MSM}{MSE}$, is 513.4, which is very large, which is good. The $H0$ hypothesis of this test is true if the body mass has no impact on the brain weight. For $H0$ to be true the $p-values$, would need to be closer to 1 In other words, we can reject this hypothesis and the relation between brain weight and body mass is very clear.

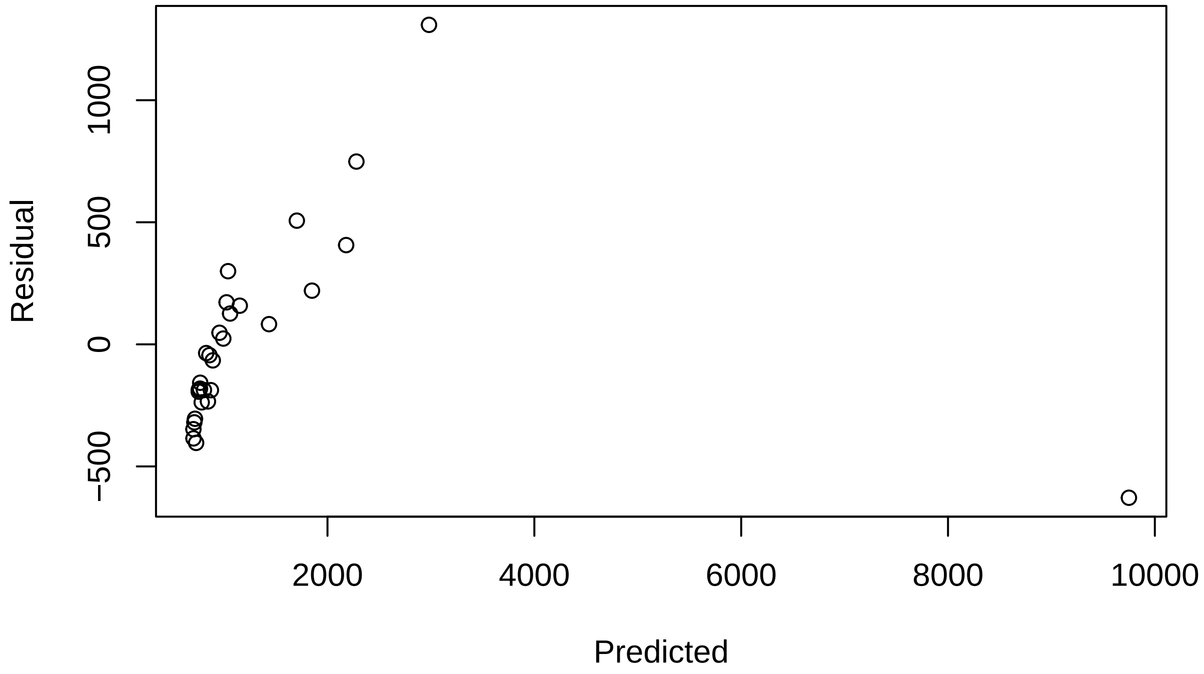
Analysis of variance

```
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.

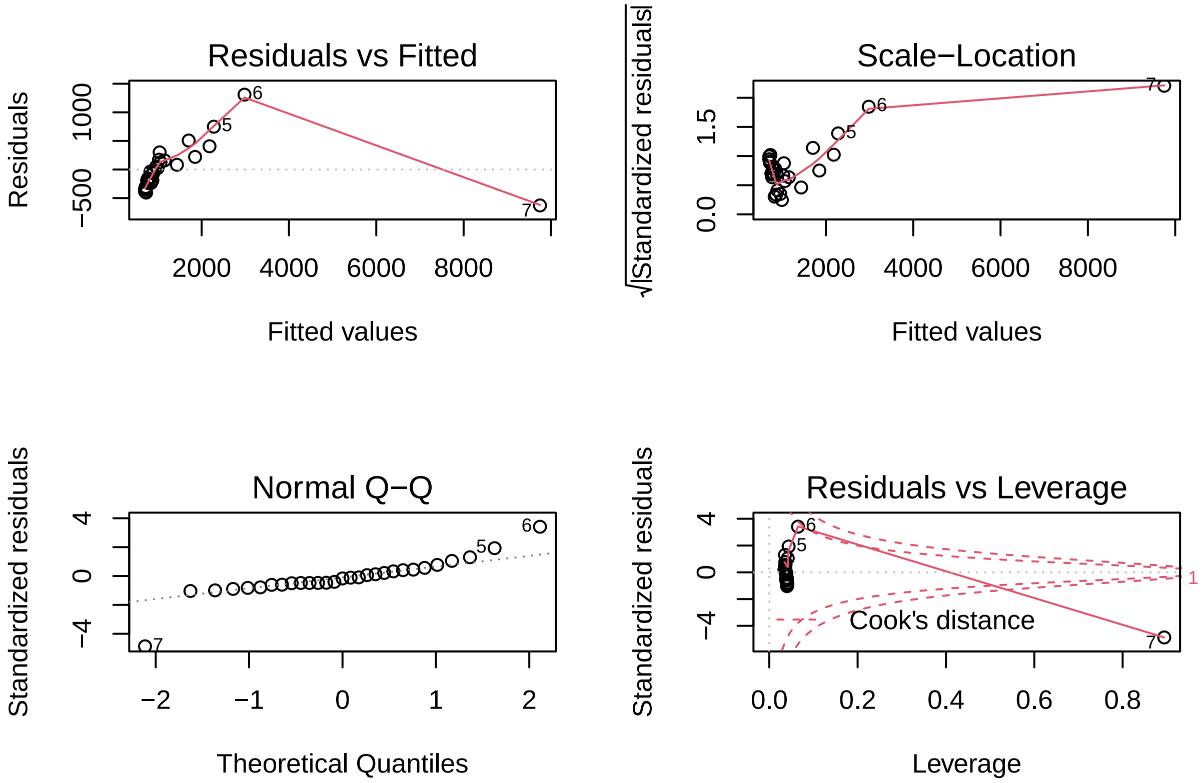
```
plot(regression$fitted.values,regression$residuals,xlab="Predicted",ylab="Residual")
```



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.

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



Now we remove the problematic individual “batman, which is individual number 7.

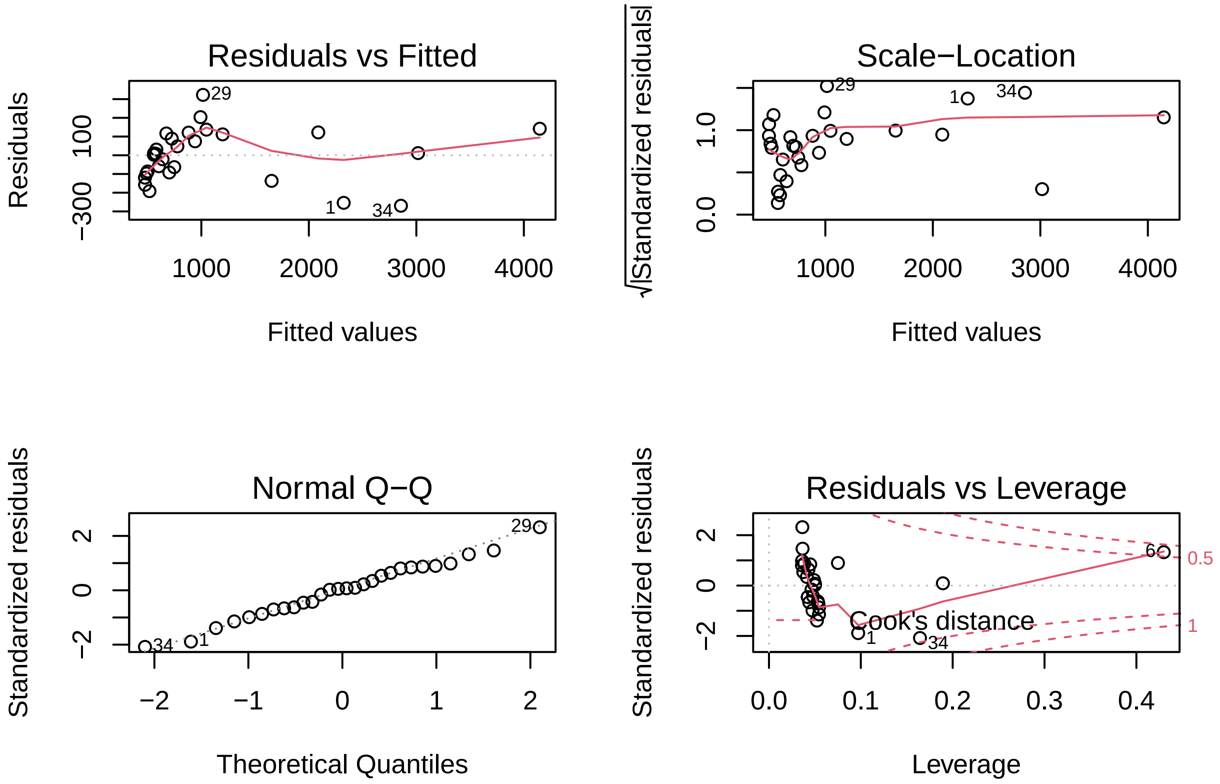
```
phytobis=phyto[which(phyto$BRW<8000),]
regressionbis=lm(BRW ~ BOW,data=phytobis)
summary(regressionbis)
```

```
##
## 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(regressionbis)
```



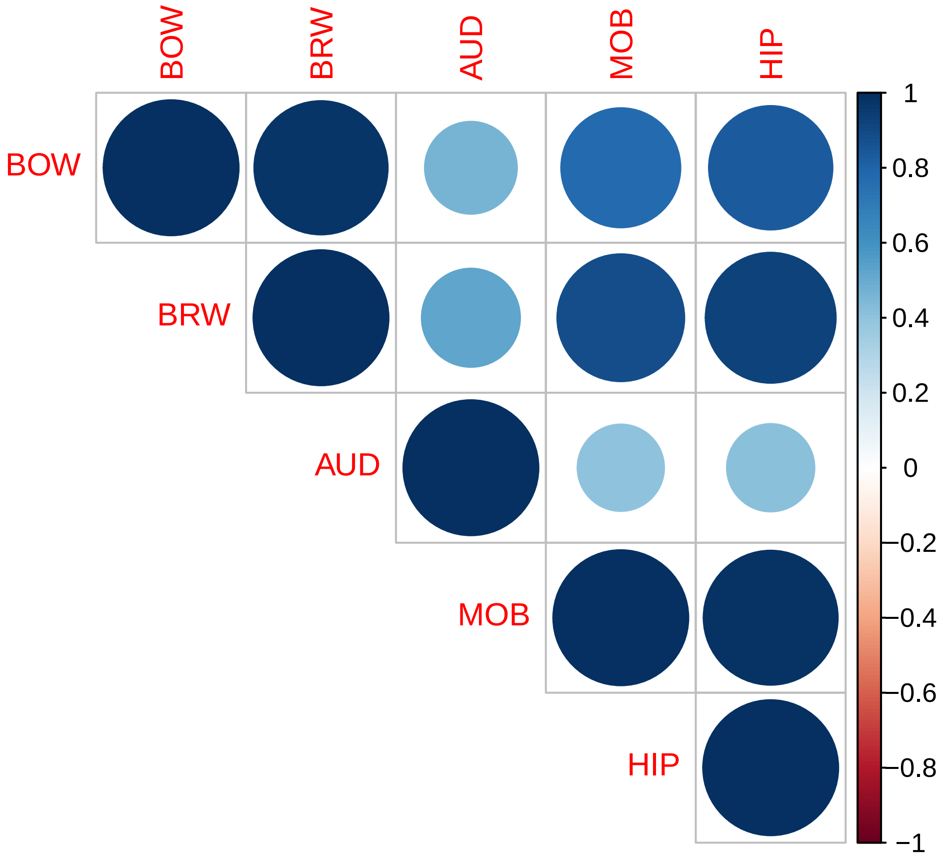
The validity plots look much better with the second model where the batman is removed. Notably, the Normal Q-Q fits a normal distribution almost perfectly, as opposed to before. In other words, the errors follow a normal distribution. The other plot look better as well, I don't understand them perfectly 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 variables are indeed extremely correlated.

Therefore, we should remove one of them from the model.

We can do a new linear regression without HIP.

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

```
##
## Call:
## lm(formula = BRW ~ AUD + MOB, data = phytobis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -451.78 -109.14   25.33   95.86  598.49
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -39.4858    83.8989  -0.471    0.642
## AUD          49.6348     8.8342   5.618 7.59e-06 ***
## MOB          14.8403     0.8739  16.981 3.08e-15 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 231.2 on 25 degrees of freedom
## Multiple R-squared:  0.9433, Adjusted R-squared:  0.9388
## F-statistic:   208 on 2 and 25 DF,  p-value: 2.627e-16
```

```
anova(regmbis)
```

```
## Analysis of Variance Table
##
## Response: BRW
##           Df   Sum Sq  Mean Sq F value    Pr(>F)
## AUD        1  6817133  6817133  127.57 2.601e-11 ***
## MOB        1 15409397 15409397  288.37 3.082e-15 ***
## Residuals 25  1335917    53437
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Now we see that MOB is indeed relevant, which was hidden before by its collinearity with HIP.

The step command

```
reg0=lm(BRW~1,data = phyto)
stepreg= step(reg0,scope = BRW ~AUD + MOB +HIP, direction = "forward")
```

```
## Start:  AIC=433.88
## BRW ~ 1
##
##        Df Sum of Sq      RSS    AIC
## + HIP   1  73272731 11869487 378.74
## + MOB   1  66447848 18694370 391.92
## + AUD   1  23770396 61371823 426.39
## <none>              85142218 433.88
##
## Step:  AIC=378.74
## BRW ~ HIP
##
```

```
##        Df Sum of Sq      RSS    AIC
## + MOB   1   2846939  9022548 372.79
## + AUD   1   2013783  9855704 375.35
## <none>              11869487 378.74
##
## Step:  AIC=372.79
## BRW ~ HIP + MOB
##
##        Df Sum of Sq     RSS    AIC
## + AUD   1   1910121 7112426 367.89
## <none>              9022548 372.79
##
## Step:  AIC=367.89
## BRW ~ HIP + MOB + AUD
```

```
summary(stepreg)
```

```
##
## Call:
## lm(formula = BRW ~ HIP + MOB + AUD, data = phyto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1272.35  -287.80    24.65   209.64  1659.56
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1003.952    213.483  -4.703 8.05e-05 ***
## HIP            44.351      7.999   5.544 9.18e-06 ***
## MOB           -29.243      9.417  -3.105  0.00468 **
## AUD            52.819     20.385   2.591  0.01574 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 533.4 on 25 degrees of freedom
## Multiple R-squared:  0.9165, Adjusted R-squared:  0.9064
## F-statistic: 91.42 on 3 and 25 DF,  p-value: 1.317e-13
```

This seems to be some sort of algorithm to select the best model. I read a detailed explanation there.

However, in my opinion, it's not working really well here: As we have previously shown MOB and HIP are collinear, and here it seems that the model is trying to cancel them out by putting a negative coefficient in front of MOB. It is my opinion that, because MOB is almost perfectly collinear with HIP, we should avoid having both in the model.

We can retry the same command without HIP or without MOB:

```
reg0=lm(BRW~1,data = phyto)
stepreg2 = step(reg0,scope = BRW ~AUD + HIP , direction = "forward")
```

```
## Start:  AIC=433.88
## BRW ~ 1
##
```

```
##         Df Sum of Sq      RSS    AIC
## + HIP    1  73272731 11869487 378.74
## + AUD    1  23770396 61371823 426.39
## <none>               85142218 433.88
##
## Step:  AIC=378.74
## BRW ~ HIP
##
##         Df Sum of Sq      RSS    AIC
## + AUD    1   2013783  9855704 375.35
## <none>               11869487 378.74
##
## Step:  AIC=375.35
## BRW ~ HIP + AUD
```

```
summary(stepreg2)
```

```
##
## Call:
## lm(formula = BRW ~ HIP + AUD, data = phyto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1304.96  -273.00    99.07   207.66  2300.52
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -703.649    219.688  -3.203  0.00358 **
## HIP           19.941      1.711  11.658 7.94e-12 ***
## AUD           54.220     23.524   2.305  0.02941 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 615.7 on 26 degrees of freedom
## Multiple R-squared:  0.8842, Adjusted R-squared:  0.8753
## F-statistic: 99.31 on 2 and 26 DF,  p-value: 6.7e-13
```

```
reg0=lm(BRW~1,data = phyto)
stepreg2 = step(reg0,scope = BRW ~AUD + MOB , direction = "forward")
```

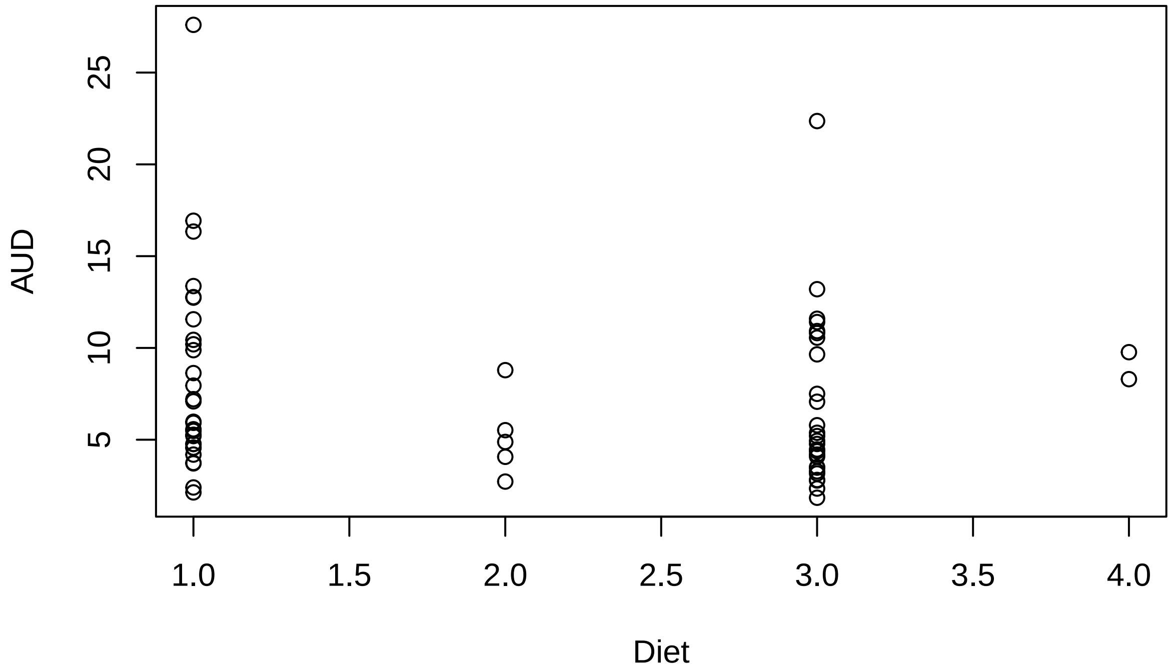
```
## Start:  AIC=433.88
## BRW ~ 1
##
##         Df Sum of Sq      RSS    AIC
## + MOB    1  66447848 18694370 391.92
## + AUD    1  23770396 61371823 426.39
## <none>               85142218 433.88
##
## Step:  AIC=391.92
## BRW ~ MOB
##
##         Df Sum of Sq      RSS    AIC
## + AUD    1   2836653 15857717 389.14
## <none>              18694370 391.92
##
## Step:  AIC=389.14
## BRW ~ MOB + AUD
```

```
summary(stepreg2)
```

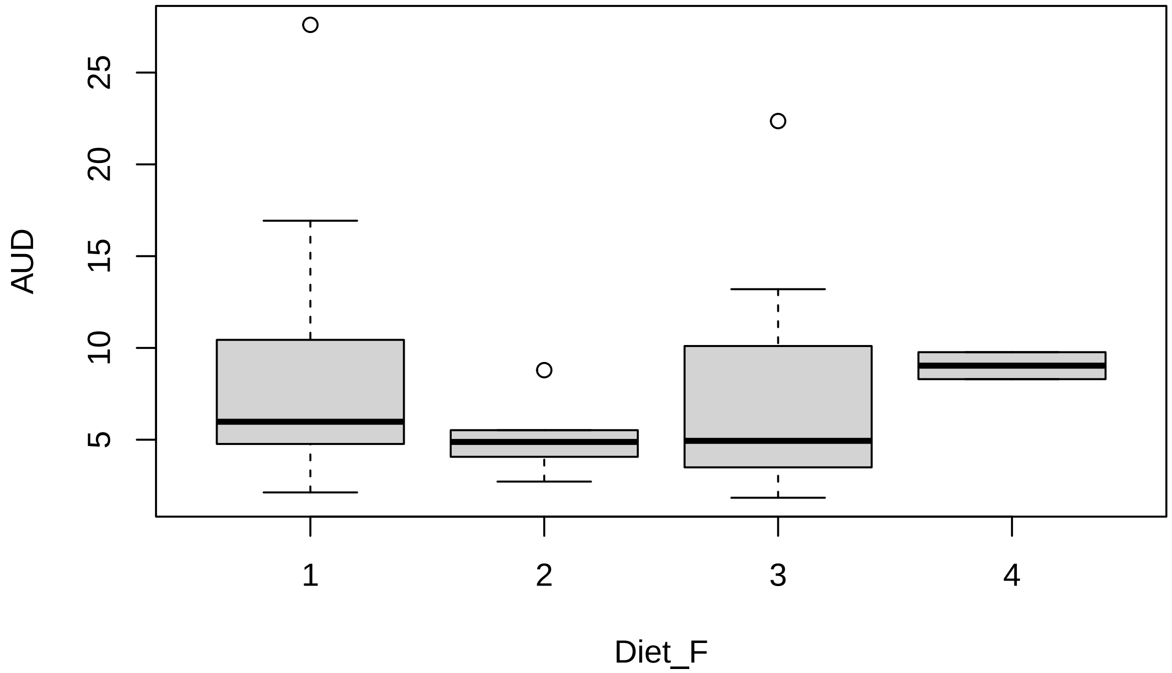
```
##
## Call:
## lm(formula = BRW ~ MOB + AUD, data = phyto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1712.55  -171.18    71.45   208.73  3085.82
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -422.979    272.332  -1.553   0.1325
## MOB           22.065      2.554   8.639 4.07e-09 ***
## AUD           64.049     29.699   2.157   0.0405 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 781 on 26 degrees of freedom
## Multiple R-squared:  0.8138, Adjusted R-squared:  0.7994
## F-statistic:  56.8 on 2 and 26 DF,  p-value: 3.245e-10
```

Link between volum of the auditory part and diet

```
myData$Diet_F = as.factor(myData$Diet)
with(myData,plot(AUD~Diet))
```



```
with(myData,plot(AUD~Diet_F))
```



I'm not sure, but I think we should a the first graph (with the points). The second graph is strange to me because it doesn't not take into account the number of points, so we don't know how confident we should be. If you look at diet 4 on graph 2 you might think that there is a very high chance than, if a bat has diet 4 then it will have a volume 10 AUD. However, there are only 2 points at diet number 4…

Nonetheless the second graph add informations about the distributions, (quartiles and median), so maybe we should keep both ?

Regression analysis

```
lm = lm(AUD~Diet_F,data=myData)
summary(lm)
```

```
##
## Call:
## lm(formula = AUD ~ Diet_F, data = myData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.179 -3.226 -1.341  2.530 19.291
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.3093     0.9040   9.192 5.48e-13 ***
## Diet_F2      -3.1133     2.3573  -1.321    0.192
## Diet_F3      -1.5886     1.3019  -1.220    0.227
## Diet_F4       0.7257     3.5591   0.204    0.839
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.868 on 59 degrees of freedom
## Multiple R-squared:  0.04512,    Adjusted R-squared:  -0.003434
## F-statistic: 0.9293 on 3 and 59 DF,  p-value: 0.4323
```

```
anova(lm)
```

```
## Analysis of Variance Table
##
## Response: AUD
##           Df  Sum Sq Mean Sq F value Pr(>F)
## Diet_F     3   66.07  22.023  0.9293 0.4323
## Residuals 59 1398.26  23.699
```

It seems that there is no clear linear relation between the AUD volume and the diet. The Sum sq residuals is much higher than the Sum sq explained by the model. This is an indicator of a bad linear model. F-Statistic bellow 1 is also very bad. The analysis of variance also tells us that our model is not significant.

We could conclude that our hypothesis is wrong, in other that insectivorous insect do not tend to have a larger AUD volume.

However, I feel like we are ignoring something : the body weight What if insectivorous bats are smaller on average ? Then even if they have a large AUD volume for their size, it would still be smaller than the AUD volume of other species.

To illustrate the idea let’s consider an example: let’s say that human A has 43 size feet (european sizing) and 150cm height. Human B has 45 size feet and 177cm height. Then, with the previous method we would conclude that human B has larger feet. But really, if we look at the ratio, we see that human A has exeptionnaly large feet.

Therefore, what we really need to look at, in my opinion, is to ratio AUD volume to body weight.

```
library(dplyr)
```

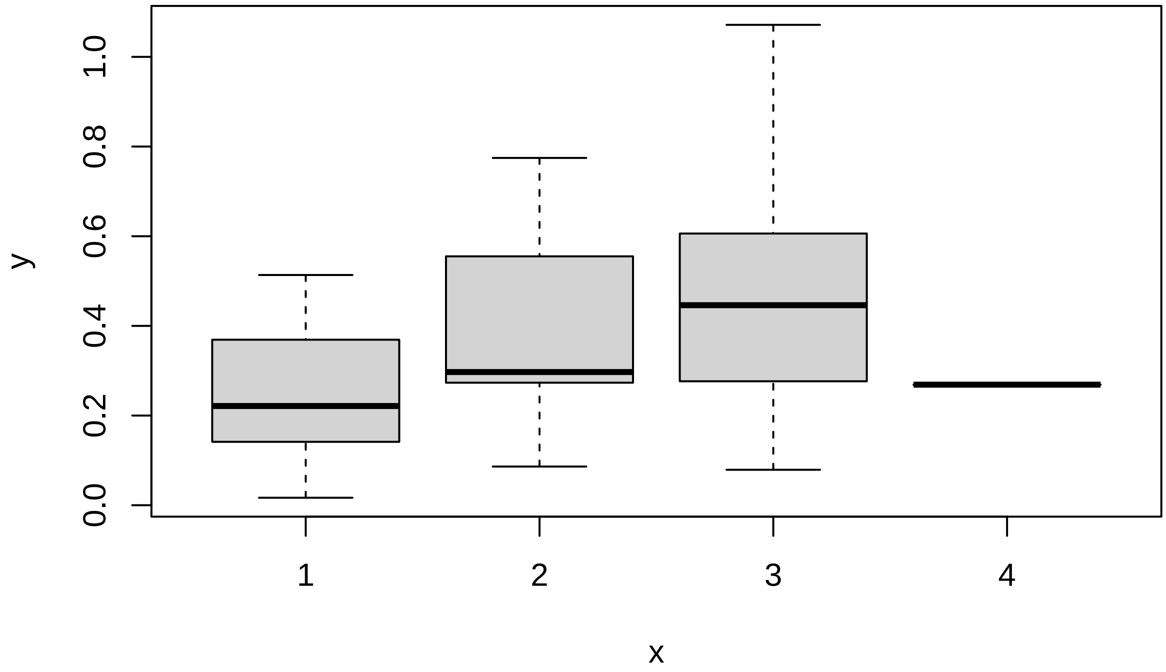
```
##
## Attachement du package : ’dplyr’
```

```
## Les objets suivants sont masqués depuis ’package:stats’:
##
##     filter, lag
```

```
## Les objets suivants sont masqués depuis ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```
myData_with_ratio=myData %>% mutate(aud_to_bow_ratio=AUD/BOW)

plot(myData_with_ratio$Diet_F,myData_with_ratio$aud_to_bow_ratio)
```



```r
lmratio = lm(aud_to_bow_ratio~Diet_F,data=myData_with_ratio)
summary(lmratio)
```

```
##
## Call:
## lm(formula = aud_to_bow_ratio ~ Diet_F, data = myData_with_ratio)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38598 -0.13702 -0.02146  0.12857  0.60641
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.24881    0.03670   6.780 6.36e-09 ***
## Diet_F2      0.14845    0.09569   1.551 0.126165
## Diet_F3      0.21621    0.05285   4.091 0.000132 ***
## Diet_F4      0.02007    0.14447   0.139 0.890008
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1976 on 59 degrees of freedom
## Multiple R-squared:  0.2272, Adjusted R-squared:  0.1879
## F-statistic: 5.783 on 3 and 59 DF,  p-value: 0.001552
```

```
anova(lmratio)
```

```
## Analysis of Variance Table
##
## Response: aud_to_bow_ratio
##           Df  Sum Sq  Mean Sq F value   Pr(>F)
## Diet_F     3 0.67751 0.225835   5.783 0.001552 **
## Residuals 59 2.30405 0.039052
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now we see that the median AUD volume is much higher for diet type 3, which are the insectivorous bats. This time, the analysis of variance tells us that our model is significant.