TP : is Batman somewhere ?

Oussama Oulkaid

December, 2021

```
# The environment
library(tidyverse)
library(ggplot2)
library(corrplot)
```

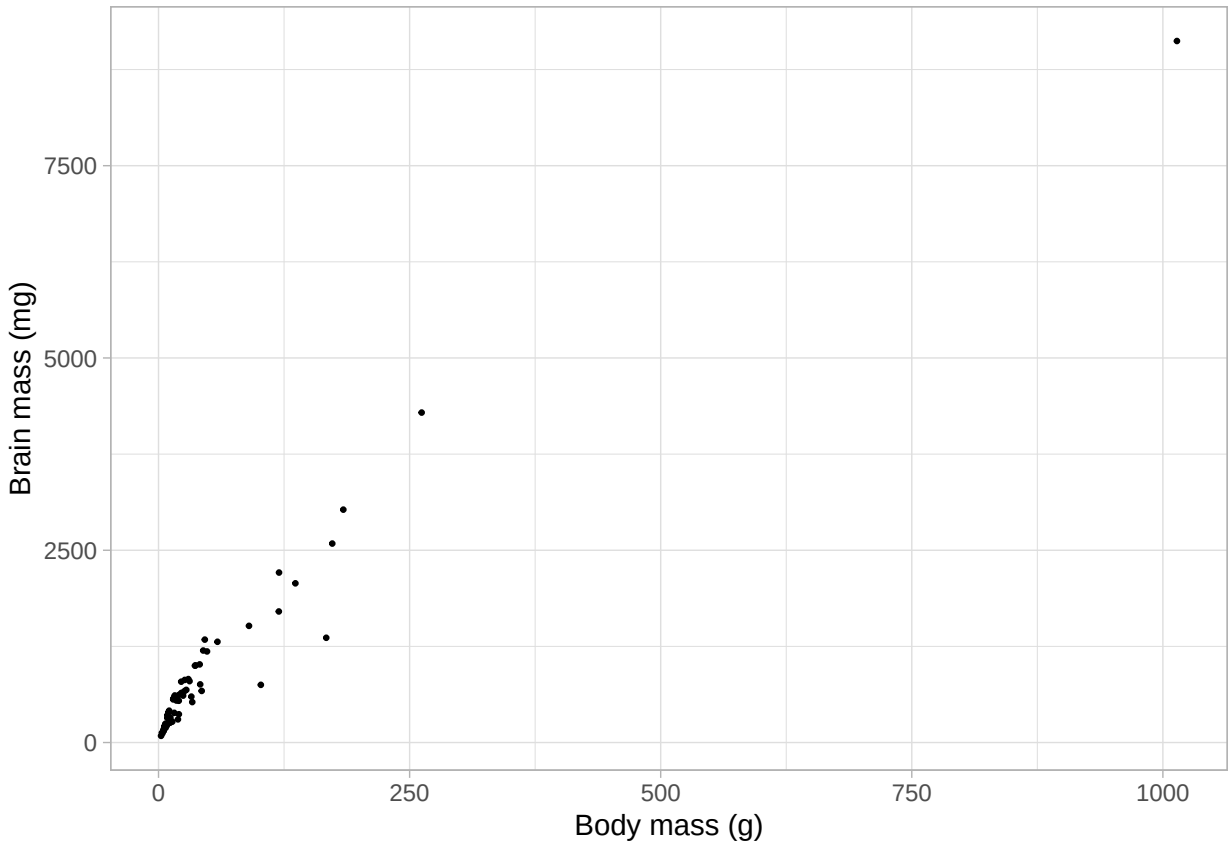
Dataframe

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

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

2 - Relationship between brain weight and body mass

```
# Focusing only on the phytophagous
phyto = myData[(myData$Diet==1),]
ggplot(myData, aes(x=BOW, y=BRW)) + geom_point(size = 0.5) +
  xlab("Body mass (g)") + ylab("Brain mass (mg)") + theme_light()
```



We can spot an outlier element that has both a huge body and brain mass compared to the whole distribution.

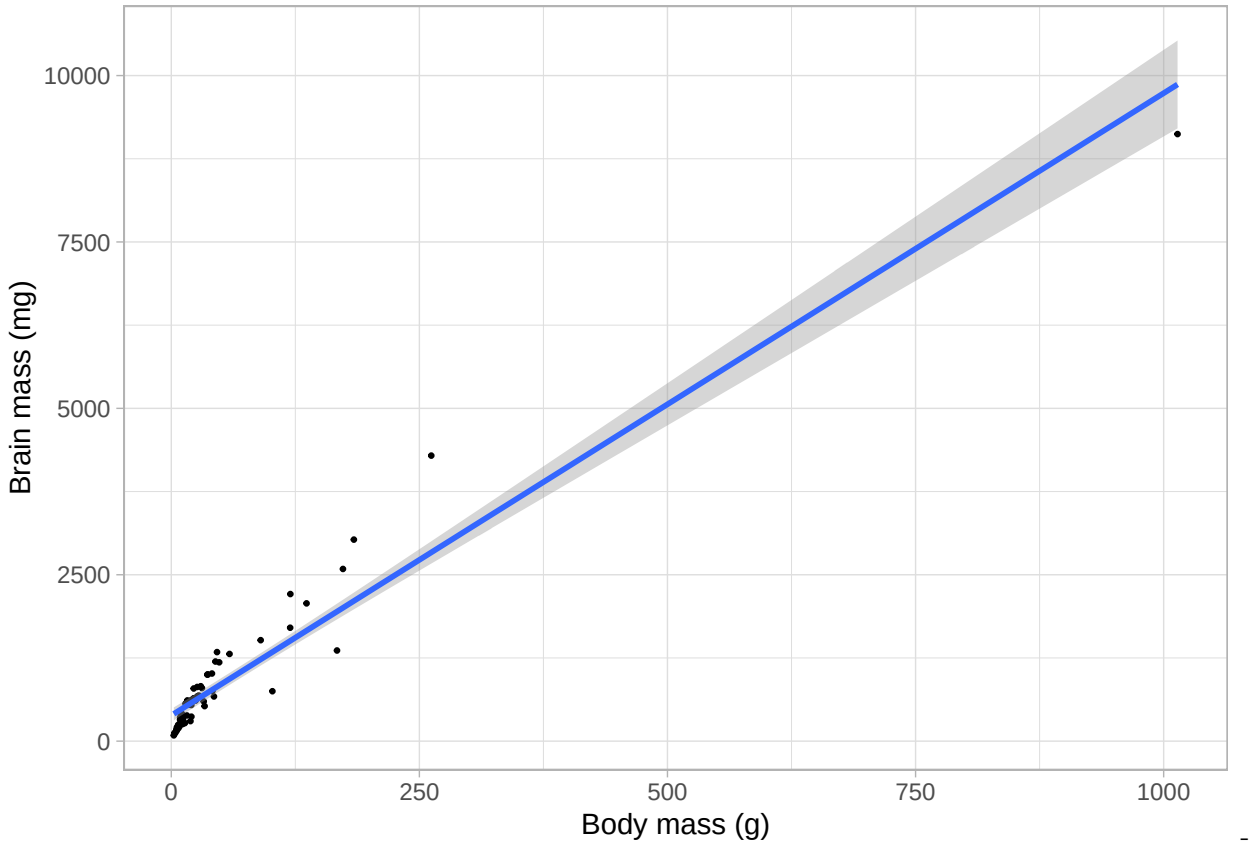
- We launch the simple linear regression. The estimated model has the following form: $BRW = \hat{\beta}_1 + \hat{\beta}_2 \times BOW + \epsilon$

```
# Simple regression model: BRW = b1 + b2*BOW + error
#                          BRW = 623.4469 + 8.9999*BOW + error
reg1 = lm(BRW ~ BOW, data=phyto)
summary(reg1)
```

```
##
## 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
```

- From the summary we have the numerical form of the model: $BRW = 623.4469 + 8.9999 \times BOW + \epsilon$ (where `623.4469` is the estimate of the intercept $\hat{\beta}_1$).
- We also see that the value of the coefficient of determination $\hat{\beta}_2$ (`~9`) is significantly different from zero.
- At the end of the row we got three stars, which is the highest level of significance of the variable `BOW` in the model.
- In addition, the very low `p-value` (`< 2.2e-16`) reflects the fact that the variable `BOW` has a big influence on `BRW`. Thus, the relationship between brain weight and body mass can be estimated to be linear.
- The H0 hypothesis would be to say that the coefficient $\hat{\beta}_2$ is null (meaning that `BOW` has no influence of the value of `BRW`). This hypothesis is false.
- we draw the regression line:

```
ggplot(myData, aes(x=BOW, y=BRW)) + geom_point(size = 0.5) +
  stat_smooth(method="lm", se=TRUE) +
  xlab("Body mass (g)") + ylab("Brain mass (mg)") + theme_light()
```



Analysis of the variance table:

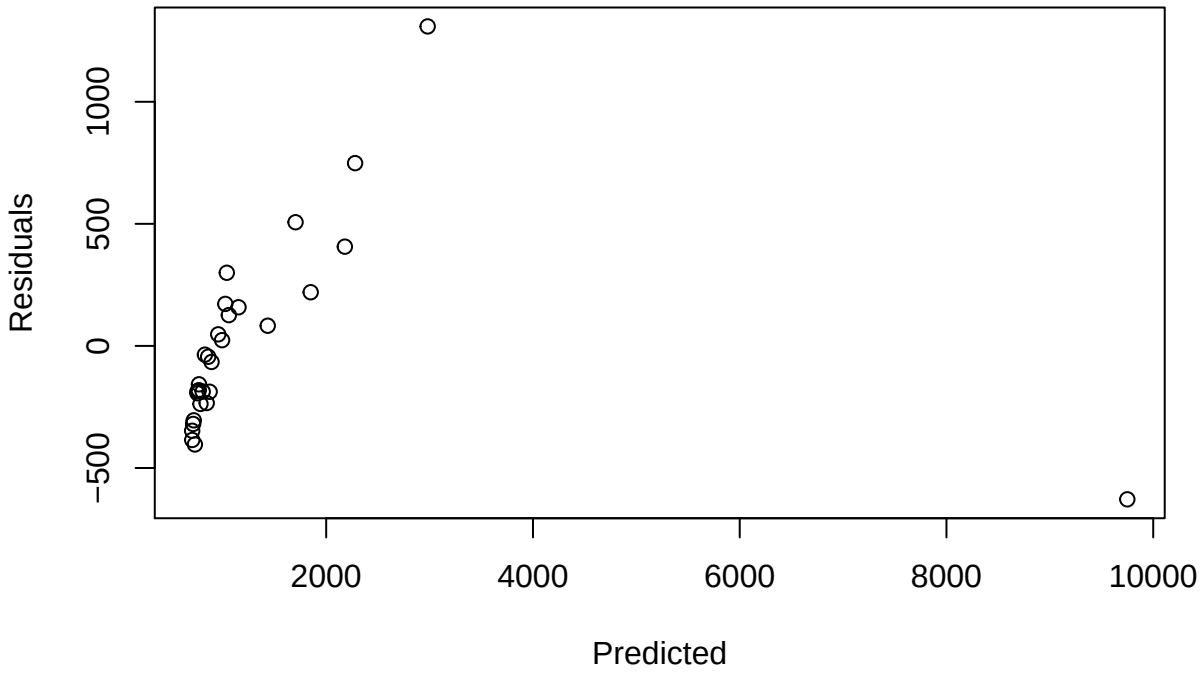
```
anova(reg1)
```

```
## 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
```

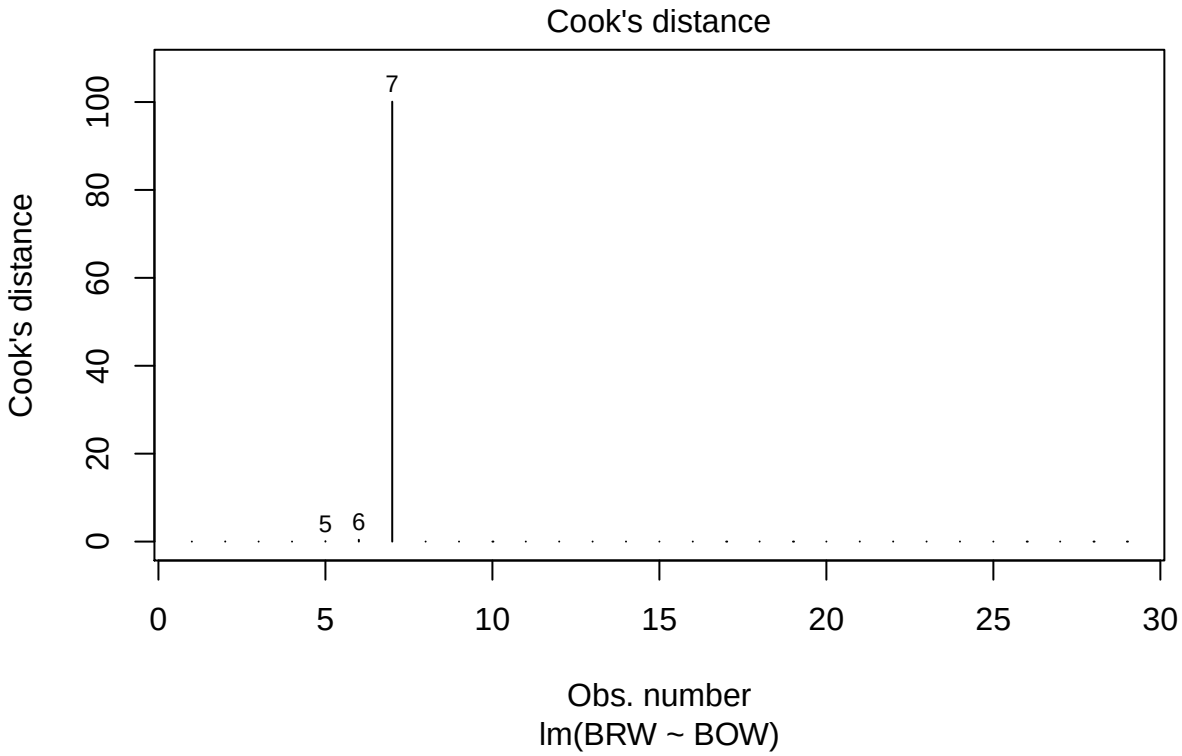
- Additional information in the table: SSE (sum of squares of the error) and SSM (sum of squares of the model).
- The sum of squared residuals is `4253838`.
- We draw the graph of the residuals with respect to the predicted values:

```
plot(reg1$fitted.values, reg1$residuals, xlab="Predicted", ylab="Residuals")
```



- The graph shows that the value of residuals generally increases for higher values of predicted brain weight. At a first glance, we can assume that the outlier element (with a predicted value of brain weight around `1000`) caused a deviation in the linear regression model, because the model tries to include all the elements of the dataset.
- The Cook's distance graph shows that the seventh element in the dataset has the largest distance. It corresponds to the outlier previously cited.

```
plot(reg1, 4) #Cook's distance
```



- We redo the analysis without this individual and we compare the results obtained:

```
which(phyto$BRW>8000)
```

```
## [1] 7
```

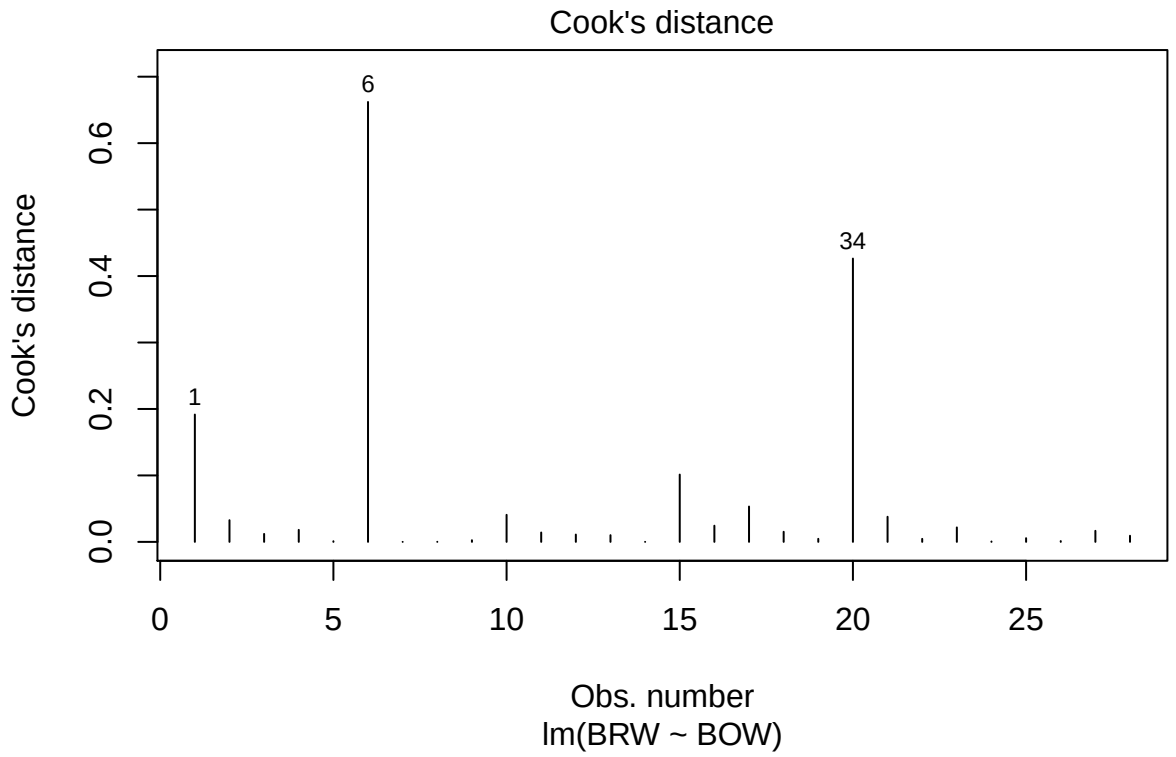
```
phytobis = phyto[which(phyto$BRW<8000),]
reg2 = lm(BRW ~ BOW, data=phytobis)
summary(reg2)
```

```
##
## 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
```

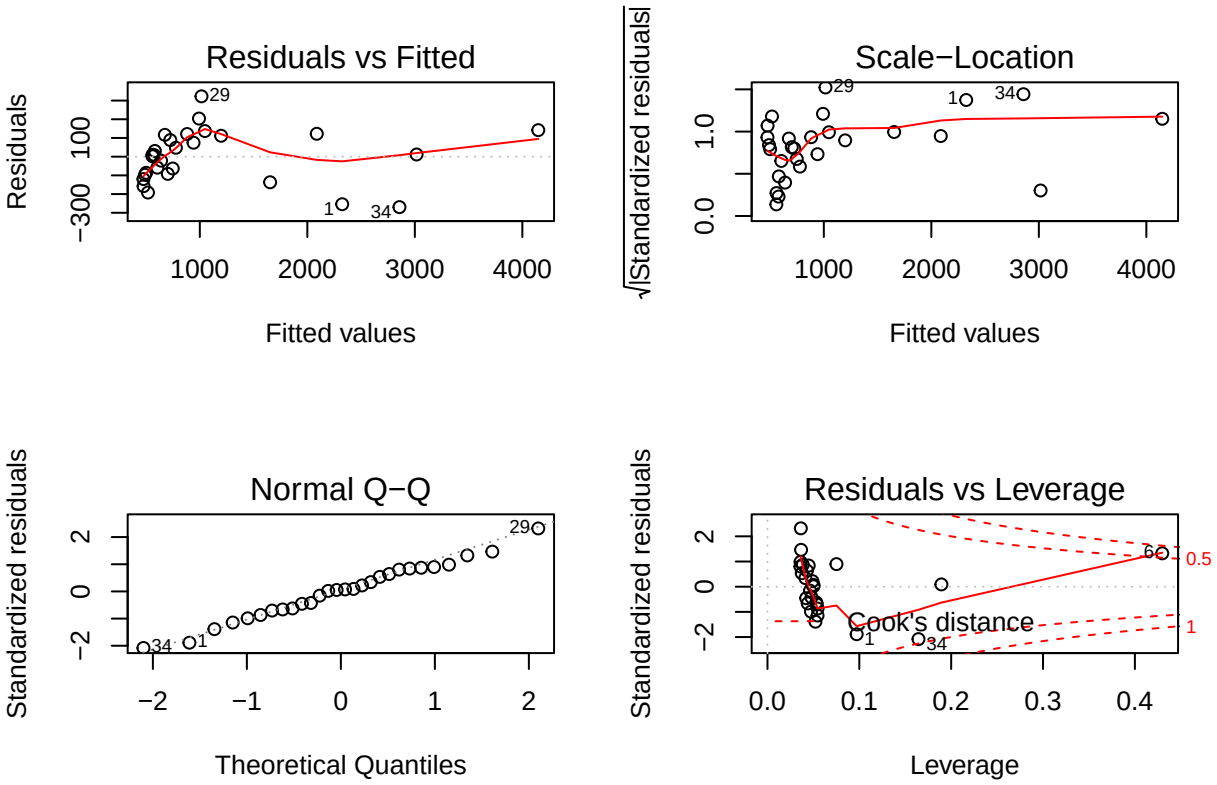
- Now, we clearly notice an improvement of the relevance of the model ; the coefficient of determination is much more significant ($\hat{\beta}_2$). it's value is `14.5099` (was `8.9999` for `reg1`). This goes along with a decrease of the intercept ($\hat{\beta}_1$).

- Thus, our assumption was valid ; removing the outlier from the linear regression model is helpful.
- The Cook's distance corresponding to the new model is:

```
plot(reg2, 4) #Cook's distance
```




```
plot(reg2)
```



- One advantage is that the second model allows to have a wider view of the distribution of the sample.

–>

3 - Contribution of each part of the brain to the total weight

–>

- 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
```

- between `BRW` and `HIP` : the correlation is the highest (`p-value = 4.574e-13`).
- between `BRW` and `MOB` : the correlation is high (`p-value = 2.203e-10`).
- between `BRW` and `AUD` : the correlation is lower (`p-value = 0.003215`).
- Multiple regression model: $BRW = \hat{\beta}_1 + \hat{\beta}_2 \times AUD + \hat{\beta}_3 \times MOB + \hat{\beta}_4 \times HIP + \epsilon$

```
# BRW = b1 + b2*AUD + b3*MOB + b4*HIP + error
# BRW = -312.692 + 47.989*AUD - 2.444*MOB + 15.981*HIP
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
```

- The numerical model: $BRW = -312.692 + 47.989 \times AUD - 2.444 \times MOB + 15.981 \times HIP + \epsilon$
- The `p-value` is very small. We can tell that the variable `BRW` is influenced by the `AUD` and `HIP`. `MOB` on the other hand seems to not have a significant coefficient ($\hat{\beta}_3$) compared to ($\hat{\beta}_2$ and $\hat{\beta}_4$).

```
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
```

- We run the following function :

```
reg0 = lm(BRW ~ 1, data = phyto)
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
```

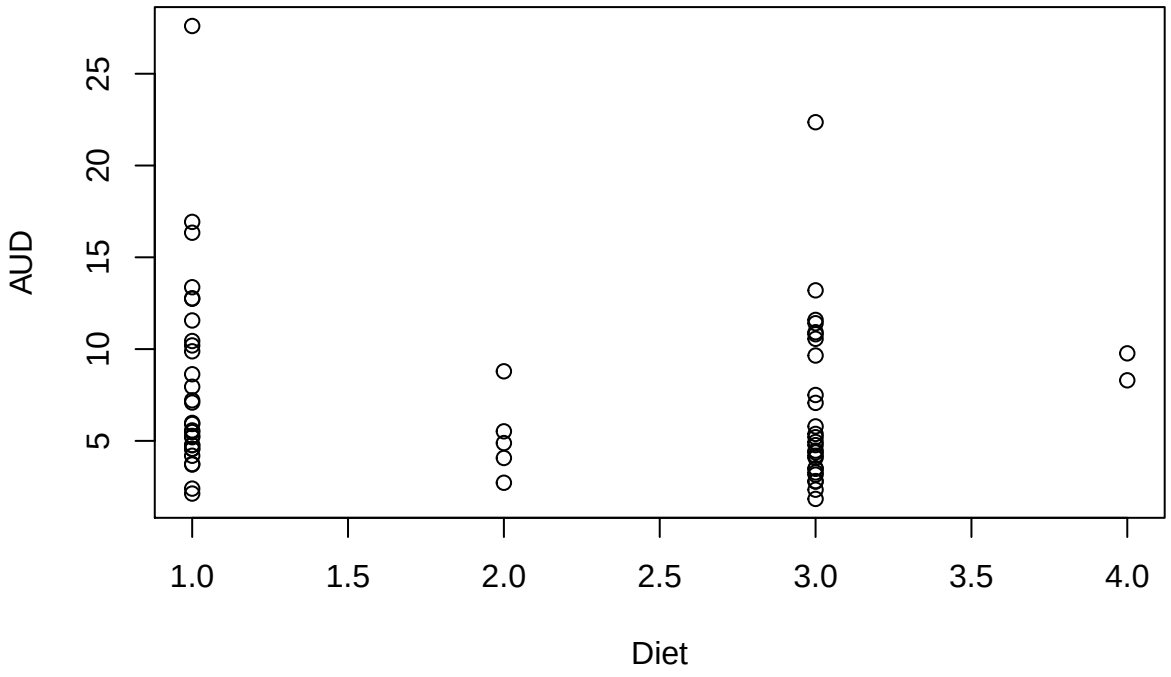
```
##
## Call:
## lm(formula = BRW ~ HIP + MOB + AUD, data = phyto)
##
## Coefficients:
## (Intercept)          HIP          MOB          AUD
##    -1003.95        44.35       -29.24        52.82
```

- The purpose of this function is to start from a linear regression model (`reg0` in this case) and perform some analysis by adding at each time a new variable. It seems that the variables are added in a specific order such that the variable with the highest `p-value` is added first, and so on.

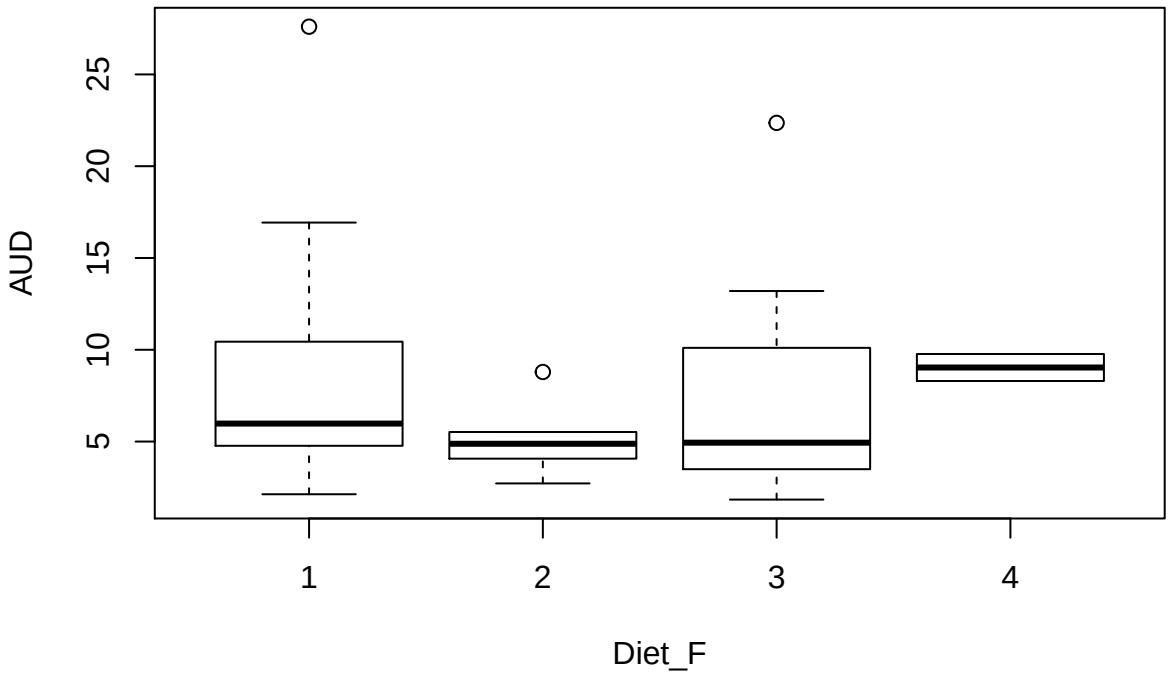
- *"The AIC is designed to find the model that explains the most variation in the data, while penalizing for models that use an excessive number of parameters.. The lower the AIC, the better the model fit."* (source: https://www.statology.org/aic-in-r/)
- If we consider this definition, we can assume that the multivariate model (`BRW~AUD + MOB + HIP`) fits better.

4 - Link between volume of tthe auditory part and diet

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




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



- I think it's preferable to look at the first plot. Especially that for some diet categories the number of samples is not representative (2 and 4). So looking only at the factors graph would be misleading.
- 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
```

- Conclusion : there is no correlation between the auditory brain volume and diet ; the sum of squared residuals is way more large than the sum of squares of the model. In addition, the corresponding `p-value` is relatively high.