Labwork-Batman

Kwate Dassi Loic

12/9/2021

Labwork: Is Batman somewhere?

Data Loading

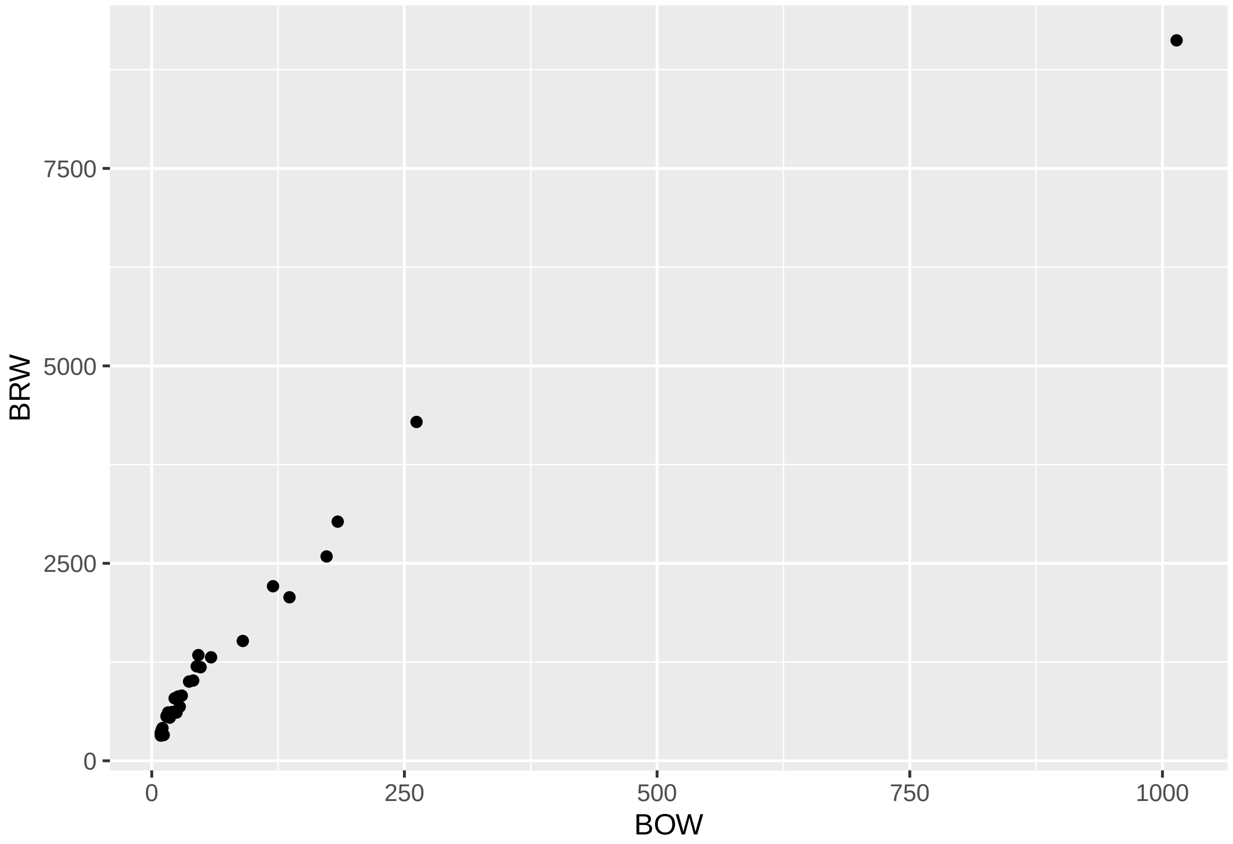
```
library(ggplot2)
data = read.table(file="bats.txt", sep=";", skip=3, header=T)
names(data)
```

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

Study of the relationship between brain weight and body mass

- Scatter plot of the function $BRW = f(BOW)$

```
phyto = data[(data$Diet == 1),]
ggplot(phyto, aes(x=BOW, y=BRW)) +
  geom_point()
```



- Approximate the function $BRW = f(BOW)$ with a simple linear regression. That is, the approximated function will be the following:

$$BRW = \beta_0 + \beta_1 \times BOW$$

```
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
```

Interpretation of the first linear regression

- Value of the intercept term: 623.449
- The $p - values$ of the statistic test are the following: 2.2×10^{-16} and 3.14×10^{-8}
- Hypothesis H_0 of the test: $\beta_0 = 0$ and $\beta_1 = 0$
- Both the $p - values$ are below the common threshold which is 5% therefore with can reject the null hypothesis
- Regarding the value of the adjusted R-squared, 0.9482, which is above the 0.8, we can said that the simple linear regression is expressive and enough to approximate the true mapping between the body mass and brain mass

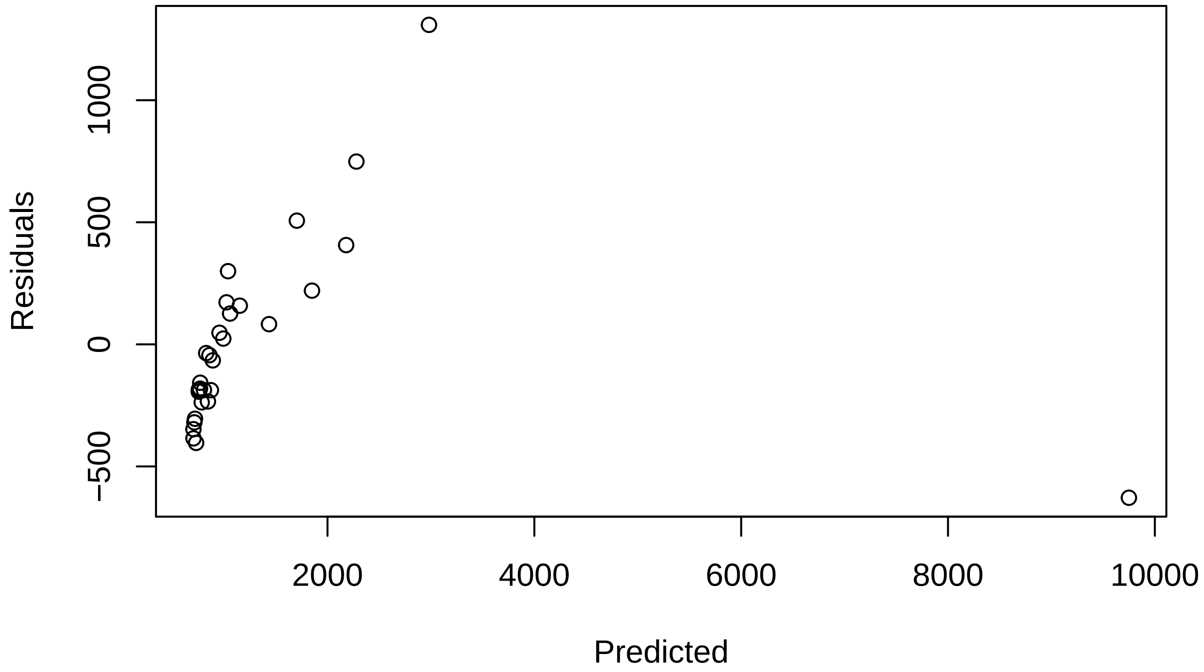
Analysis of variance of the simple linear regression

```
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
```

The analysis of variance provide supplementary information on variance, namely, the total sum of squares and the mean of the sum of squares.

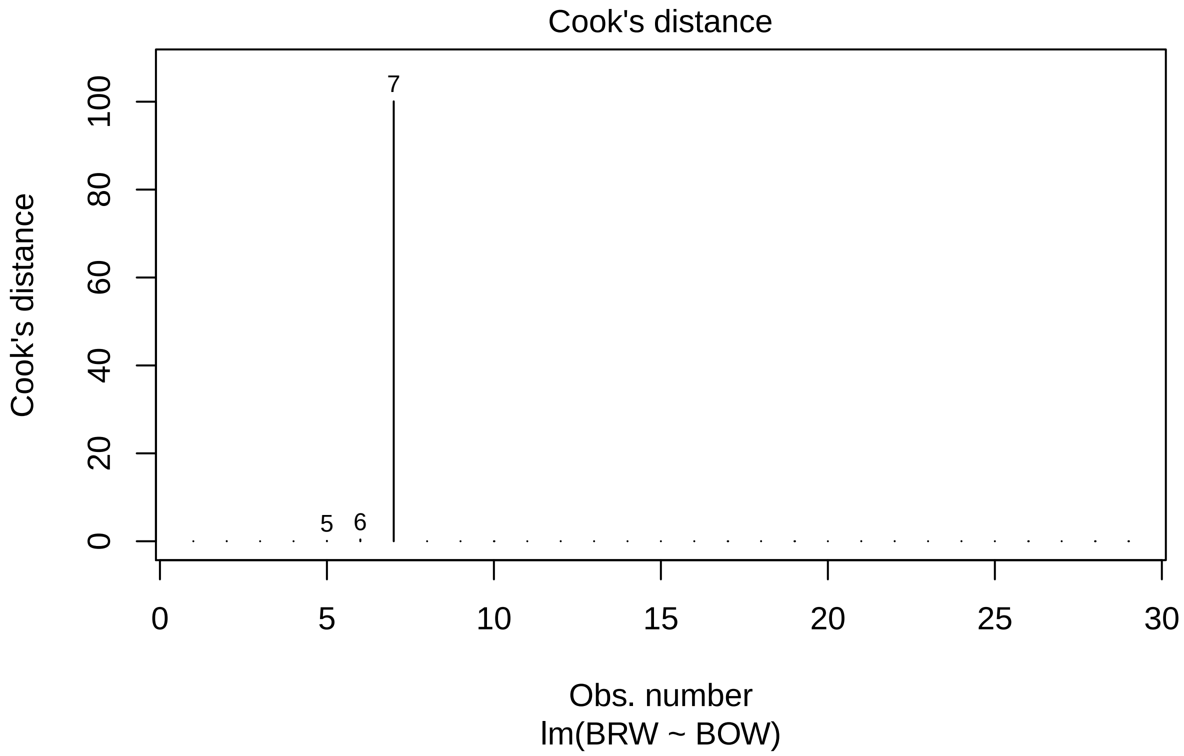
```
plot(reg1$fitted.values, reg1$residuals, xlab="Predicted", ylab="Residuals")
```



Based on the trend obtained from this previous plot, we can say that the residuals tend to increase for larger fitted values.

Regression without larger fitted values

```
plot(reg1, 4)
```



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

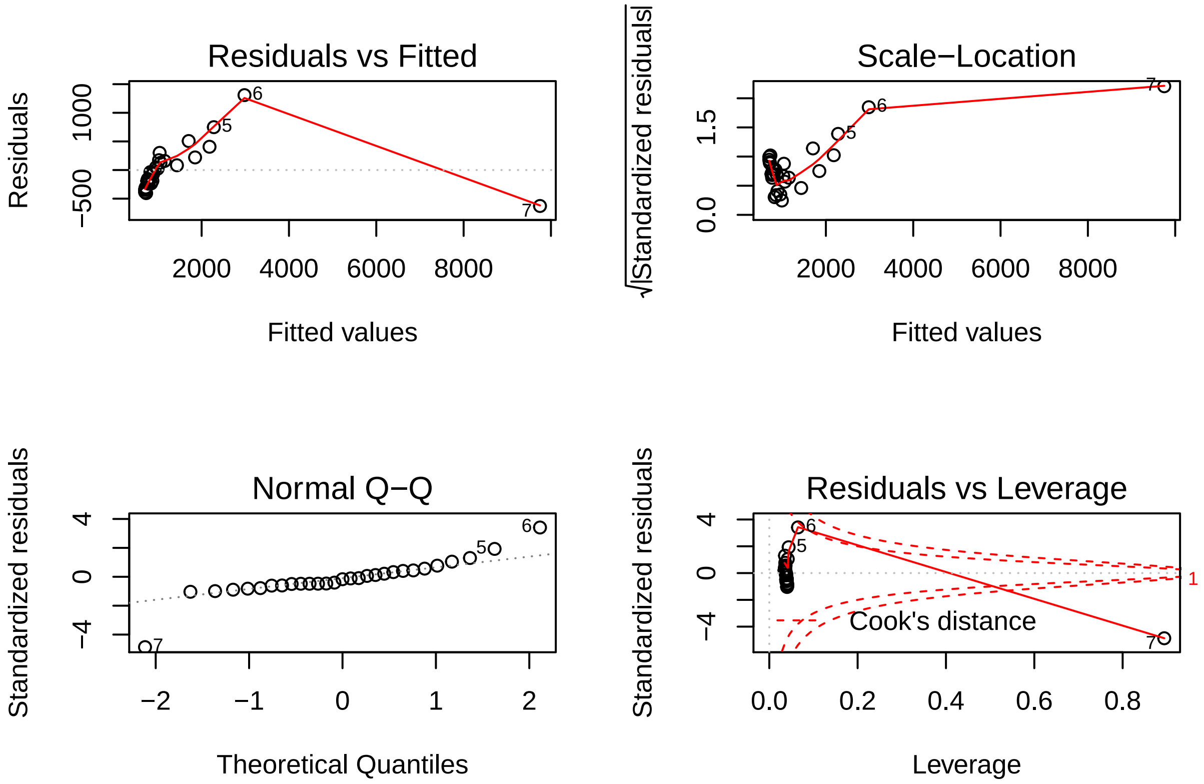
```
## [1] 7
```

```r
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
```

```
par(mfcol=c(2, 2))
plot(reg1)
```



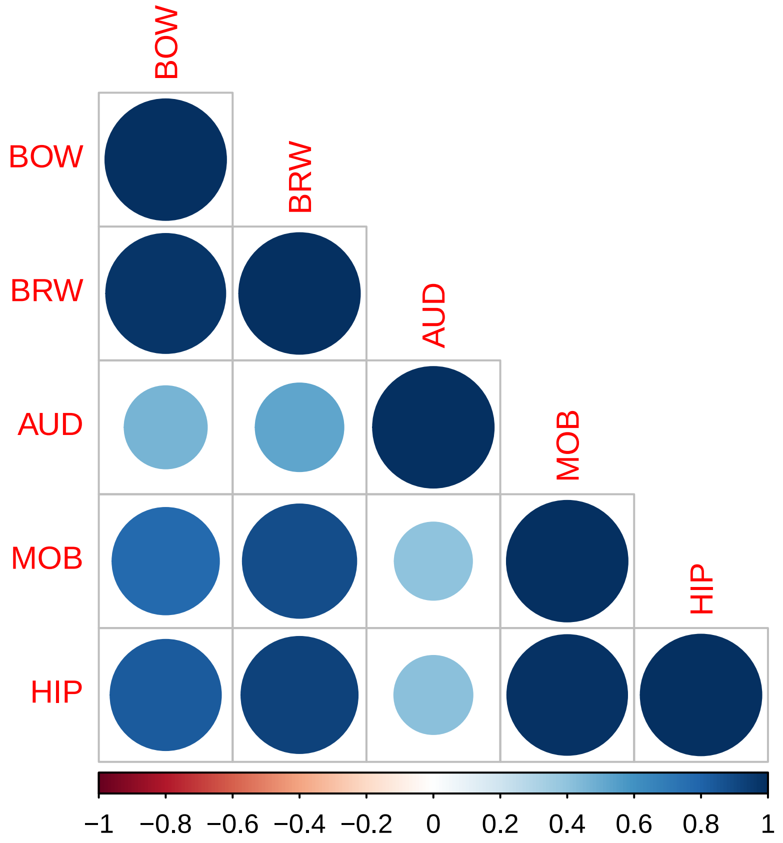
Study of the contribution to the total weight of each part of the brain

Correlation analysis between explicative variables

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

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



```
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
```

Conclusion: It is worth considering the variables HIP, MOB, and AUD to explain the weight of the body since the pairwise correlation between those variables and the target one is none negligible.

```
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
```

Interpretation of the muti-variable linear regression

- Mathematical model: $BRW = \beta_0 + \beta_1 \times AUD + \beta_2 \times MOB + \beta_3 \times HIP$.
- The multi-variate linear regression implemented can be considered as valid because the explained variance is close to the total variance. This is explained by the fact that the value of the adjusted correlation, 0.9712 is close to 1. This model provides a better explanation of the body's brain that the first simple linear regression.
- The values of the coefficients are the following : -312.69, 47.98, -2.44, and 15.98, ascribed to the variable **Intercept**, **AUD**, **MOB**, and **HIP** respectively.
- The model is not confident enough regarding the worthiness of the variable **MOB**. The $p - value$ is high enough (far above 5%) to reject the null hypothesis.

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

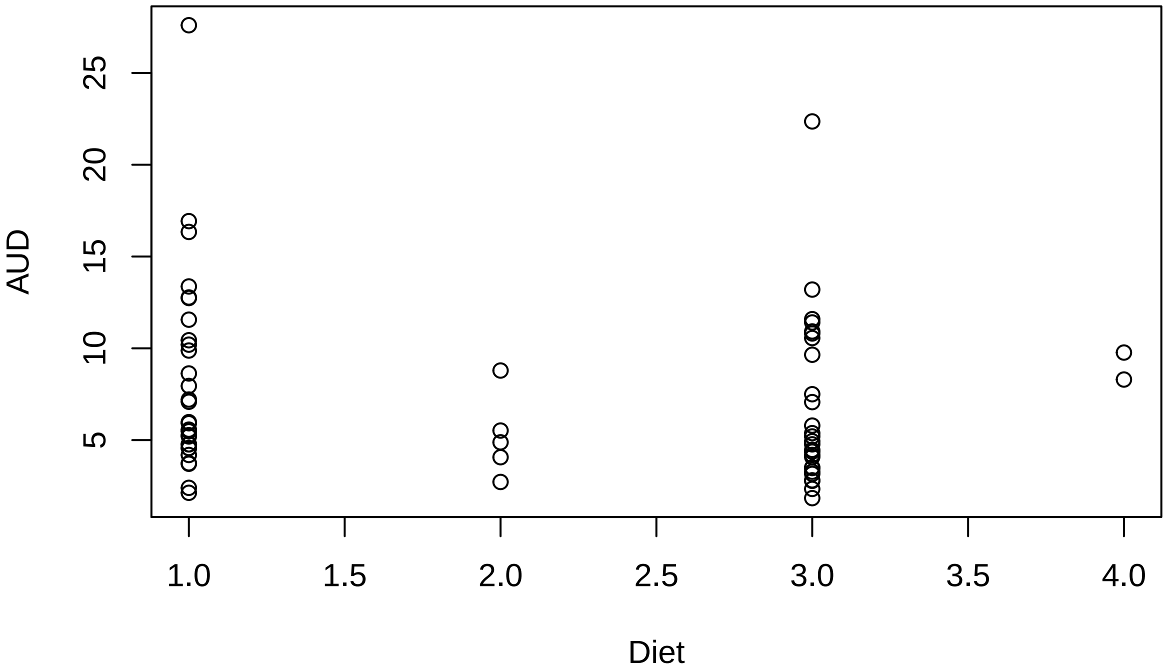
```
## Start:  AIC=433.88
## BRW ~ 1
##
##        Df Sum of Sq      RSS    AIC
## + HIP   1  73272731 11869487 378.74
## + MOB   1  66447848 18694370 391.92
## + AUD   1  23770396 61371823 426.39
## <none>              85142218 433.88
##
## Step:  AIC=378.74
## BRW ~ HIP
##
##        Df Sum of Sq      RSS    AIC
## + MOB   1   2846939  9022548 372.79
## + AUD   1   2013783  9855704 375.35
## <none>              11869487 378.74
## - HIP   1  73272731 85142218 433.88
##
## Step:  AIC=372.79
## BRW ~ HIP + MOB
##
##        Df Sum of Sq      RSS    AIC
## + AUD   1   1910121  7112426 367.89
## <none>               9022548 372.79
## - MOB   1   2846939 11869487 378.74
## - HIP   1   9671823 18694370 391.92
##
## Step:  AIC=367.89
## BRW ~ HIP + MOB + AUD
##
##        Df Sum of Sq      RSS    AIC
## <none>               7112426 367.89
## - AUD   1   1910121  9022548 372.79
## - MOB   1   2743277  9855704 375.35
## - HIP   1   8745291 15857717 389.14
```

```
##
## Call:
## lm(formula = BRW ~ HIP + MOB + AUD, data = phyto)
##
## Coefficients:
```

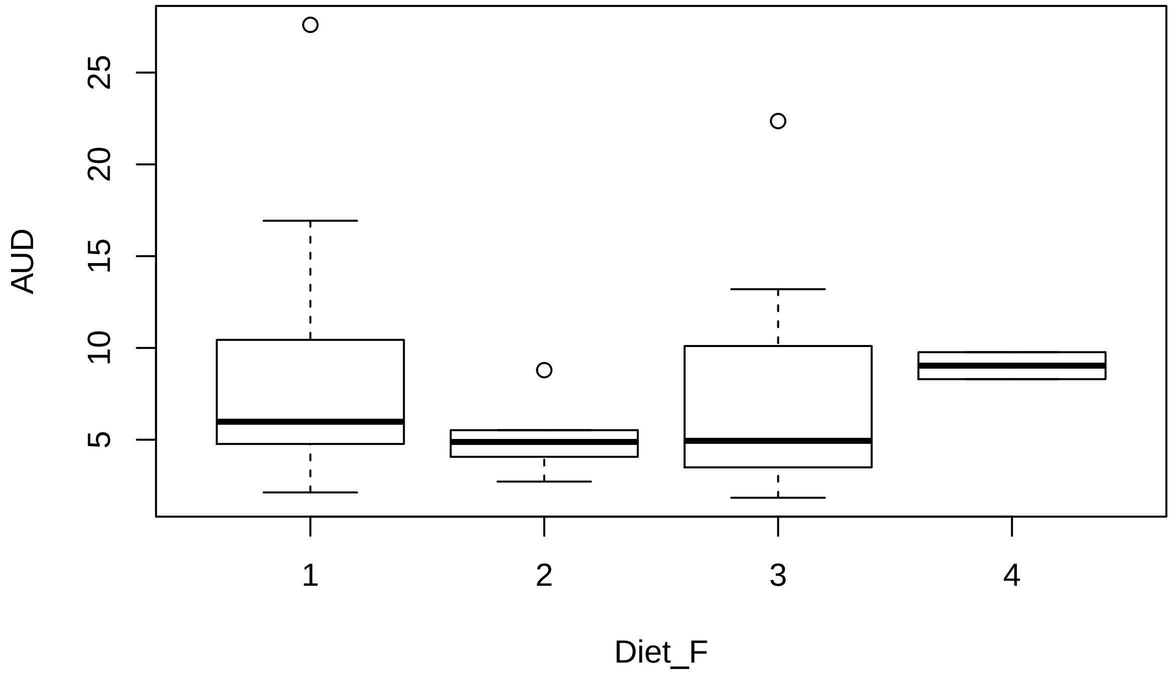
```
## (Intercept)          HIP          MOB          AUD
##    -1003.95        44.35       -29.24        52.82
```

- The preceding test aim to find the subset of variable that provide a better explainability of the target variable by iteratively add a new variable in the model till there is no significant improvement according to metric used to test the performance of the selected model, that is, the AIC score.
- The conclusion that could be drawn out from this test is the all the variables AUD, MOB, HIP are noteworthy for the prediction of BRW # Link between volume of the auditory part and diet Plot of the function $AUD = f(DIET)$

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



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



The graph with the variable **Diet** as a factor explains the data better than the graph with the original type. Besides the plot of the mapping, it's also provides the confidence interval of the approximation of the the mean of each Diet category.

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

Based on the results of the anova, especially on the $p - value$, we can conclude that the the simple linear regression is not suited to model the mapping between the volume of the auditory and the diet. This is not surprising because of previous plot doesn't favor the drawing of any straight line to fit the cloud of the point.