TP: Is Batman Somewhere

Olaniyan Folajimi and Carlos Vargas

December 9, 2021

1. Preamble

The goal of this exercise is to implement the estimation and classical tests for simple and multivariate regression models and analysis of variance. The data is obtained from a study on the brain size of bats. The abbreviations given in the file are described in the article like Diet categories (1 = phytophage; 2 = gleaner; 3 = aerial insectivore; 4 = vampire), BOW = body mass, BRW = brain mass, AUD = auditory nuclei volume, MOB = main olfactory bulb volume ; HIP = hippocampus volume.

```
library(ggplot2)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

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

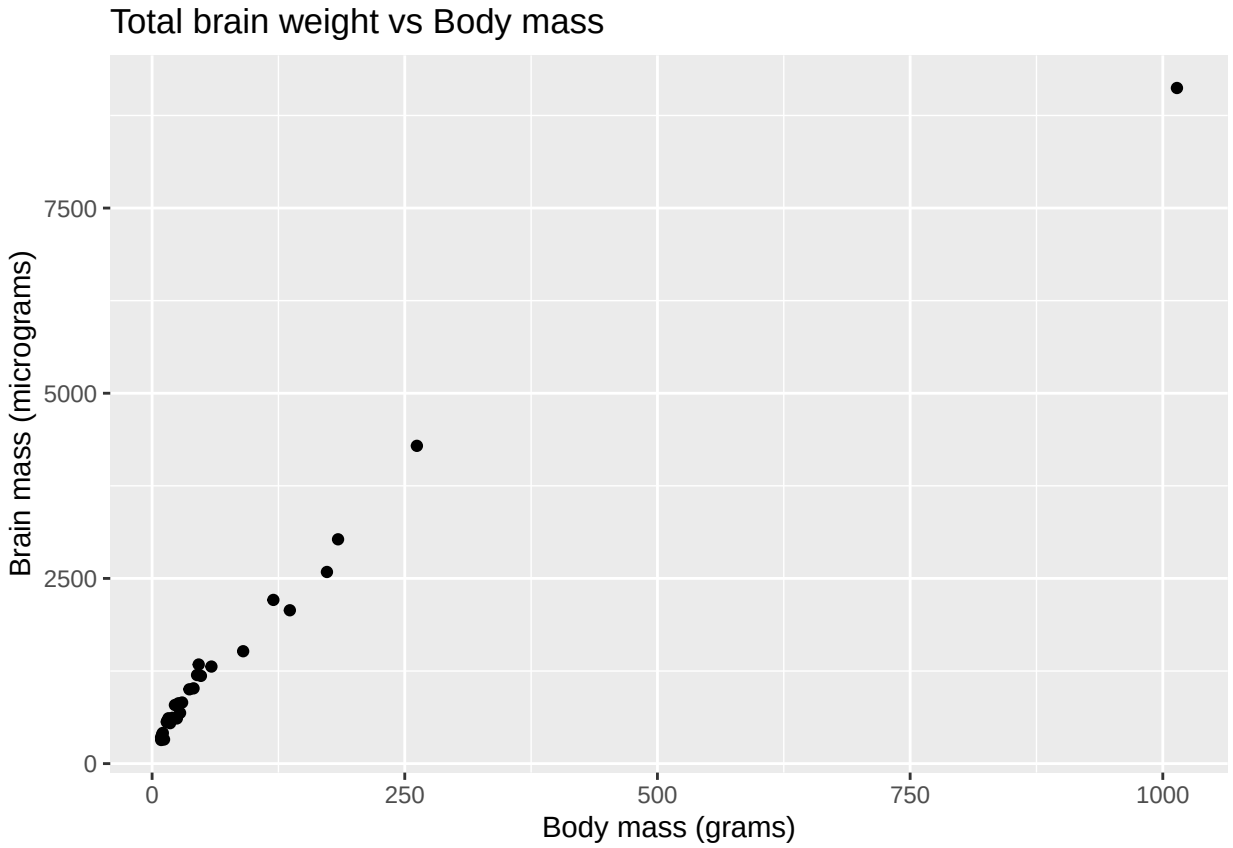
```
## [1] "Species" "Diet"    "Clade"   "BOW"     "BRW"     "AUD"     "MOB"
## [8] "HIP"
```

```
dim(myData)
```

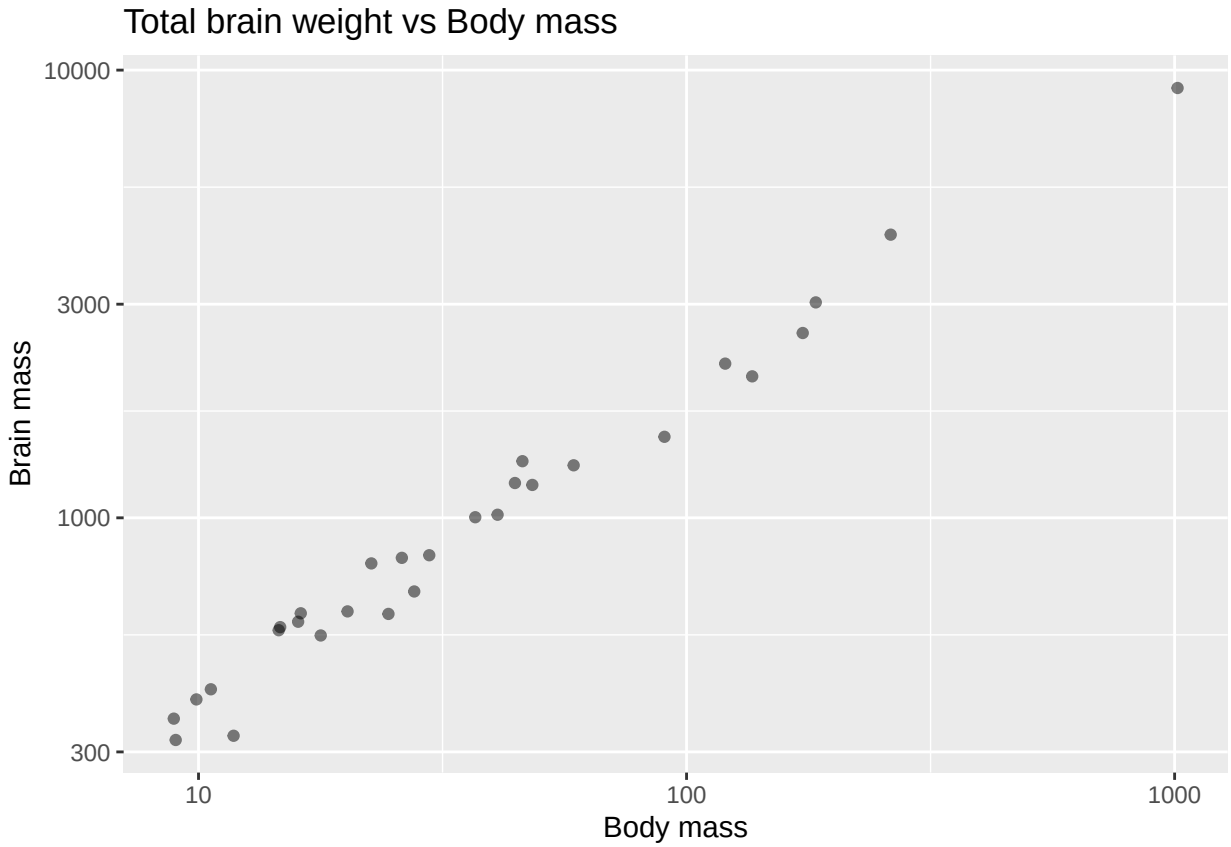
```
## [1] 63  8
```

2. Study of the relationship between brain weight and body mass

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

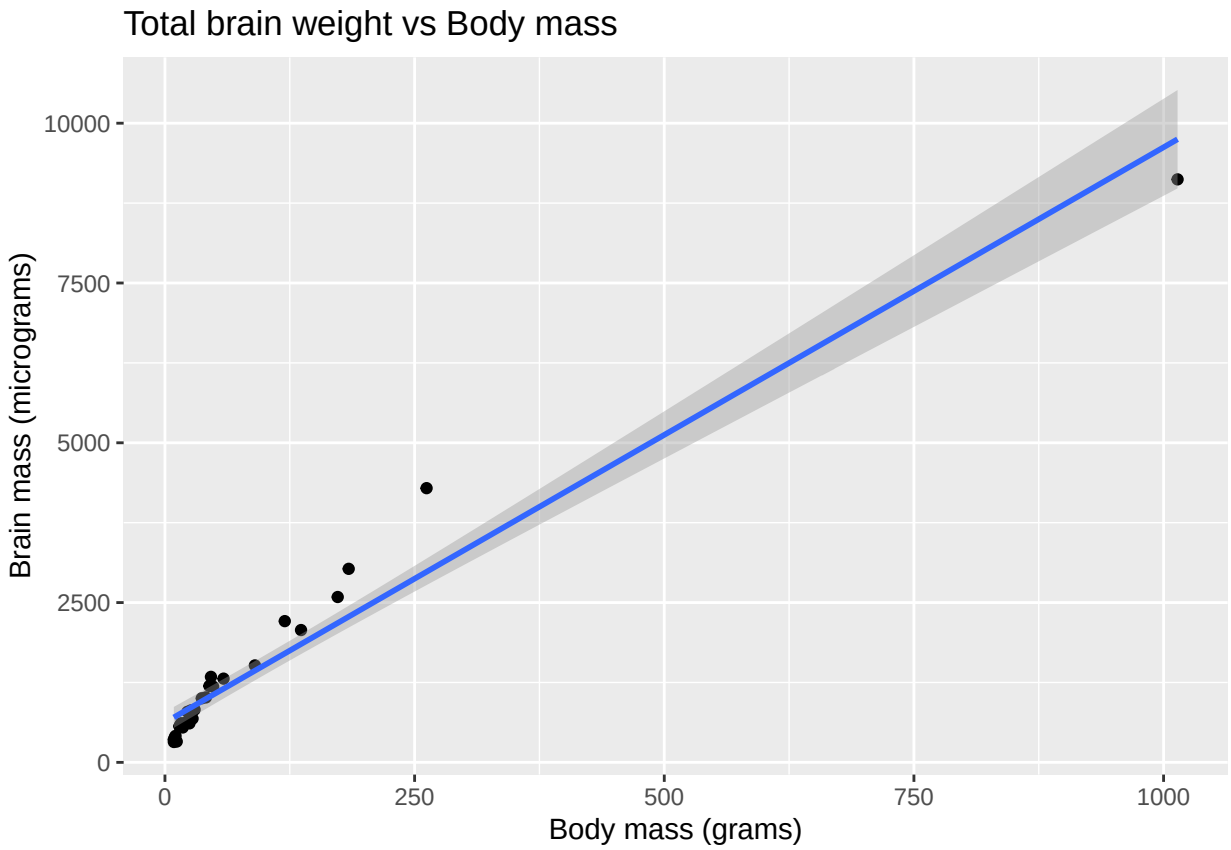


The initial graph shows a linear relationship between the brain mass and body mass. The data also appear to be clustered as 95% of the data points have a body mass below 250. In the following graph, we try to observe a more spreadout representation by taking the logarithm of the data.

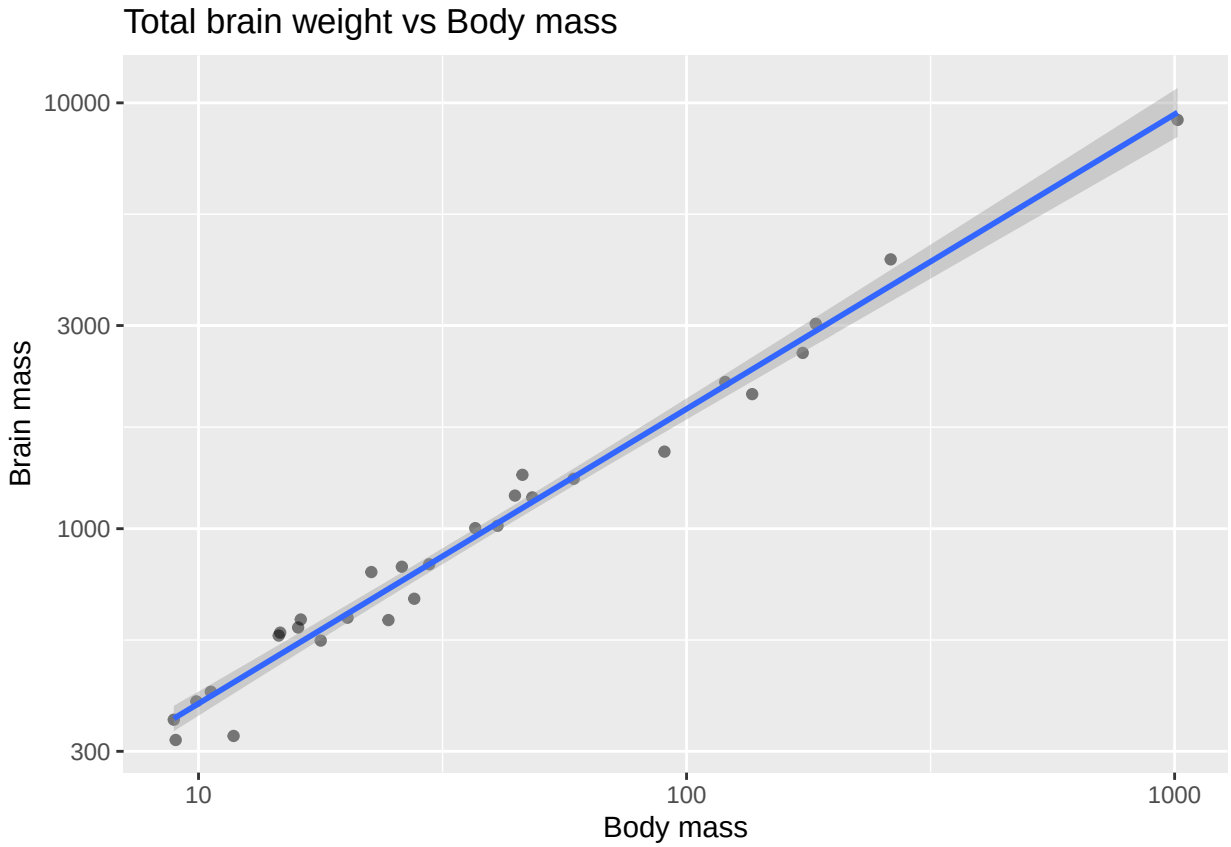


We can also include fit regression lines to both graphs as below:

```
## `geom_smooth()` using formula 'y ~ x'
```




```
## `geom_smooth()` using formula 'y ~ x'
```




```
reg1 = lm(BRW ~ BOW, data=phyto)
```

Mathematical form

In general, the expression of the model estimated for linear regression is: $Y = \hat{\beta}_1 + \hat{\beta}_2 \times X + \epsilon$ R. Specifically, R computes

$$BRW = \hat{\beta}_1 + \hat{\beta}_2 \times BOW + \epsilon$$

Regression summary

```
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
```

1. Intercept: $\hat{\beta}_1 = 623.4469$.
2. $\hat{\beta}_2 = 8.9999$.
3. Test statistics: $T_{\hat{\beta}_1} = 7.652$ and $T_{\hat{\beta}_2} = 22.659$.
4. The p-value $< 2.2 \times 10^{-16} \ll 0.05$ showing that BOW has an influence on BRW.
5. The null hypothesis H0 is $\hat{\beta}_1 = \hat{\beta}_2 = 0$. In this case it is rejected as $\hat{\beta}_2 \neq 0$ and the p-value is also very small.
6. Based on the above observations, we can say that body mass has an effect on brain weight and the relationship is linear.
7. Coefficient of determination: $R^2 = 0.95$ which tells that almost all the variation is explained by the model.

Analysis of variance (ANOVA)

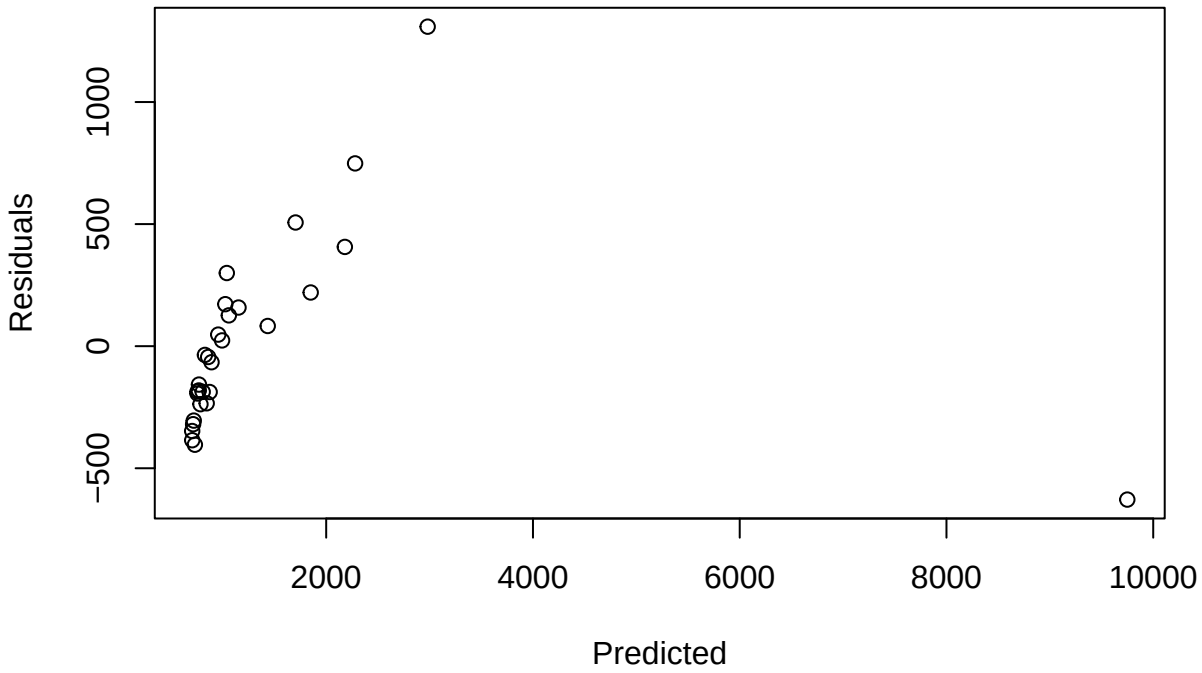
```
anova(reg1)
```

```
## 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 provides additional information such as Sum of Squared Explained (SSE) = 80888380 and Sum of Residual Squares (SRS) = 4253838.

Graph of residuals

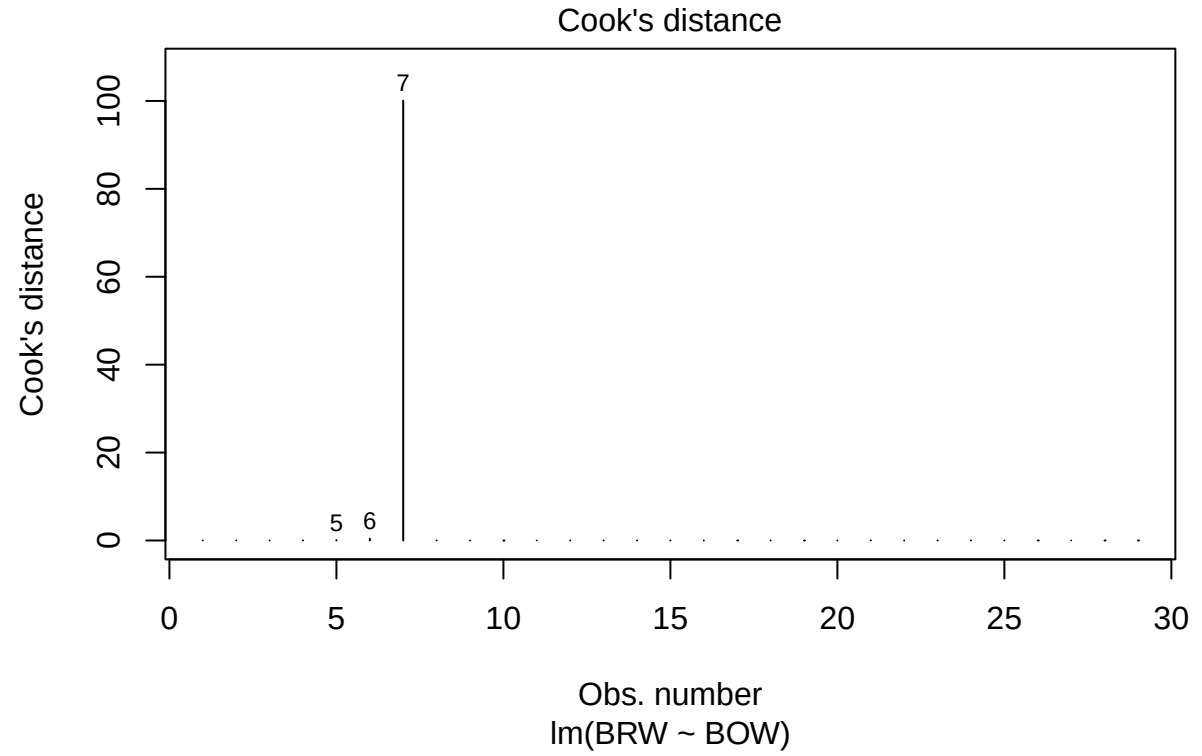
```
plot(reg1$fitted.values, reg1$residuals, xlab="Predicted", ylab="Residuals")
```



The graph shows a pattern where the residual increases as the predicted brain weight increases. In generally, this graph should be without structure. Also, we can observe that the model predicts almost $10000\mu g$ for one of the samples and this appears to be outside the general range of values. We assume this is an outlier and we can confirm this by drawing a plot the Cook's distance.

Cook's distance

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



The

graph shows that observation 7 has a Cook's distance that is orders of magnitude above all other samples. There is a high probability that this is the outlier sample so we will remove from consideration. We also redo the analysis without it.

New Analysis

```
phyto[7,]
```

```
##                Species Diet Clade  BOW  BRW   AUD    MOB    HIP
## 7  Pteropus  vampyrus     1     I 1014 9121 16.93 243.54 331.29
```

Here we see that sample 7 has a brain weight of $9121 \mu g$.

```
phytobis=phyto[which(phyto$BRW<8000),]
```

We plot the new data as shown below:

```
ggplot(phytobis, aes(x=BOW, y=BRW)) +
  geom_point() +
  geom_smooth(method = "lm") +
  ggtitle("Total brain weight vs Body mass") +
  labs(x="Body mass (grams)", y="Brain mass (micrograms)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```


```
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
```

We look at the regression analysis again:

1. Intercept: $\hat{\beta}_1 = 346.5452$.
2. $\hat{\beta}_2 = 14.5099$.
3. Test statistics: $T_{\hat{\beta}_1} = 9.764$ and $T_{\hat{\beta}_2} = 33.860$.
4. The p-value $< 2.2 \times 10^{-16} \ll 0.05$ showing that BOW has an influence on BRW.
5. Coefficient of determination: $R^2 = 0.9778$ which tells that almost all the variation is explained by the model.

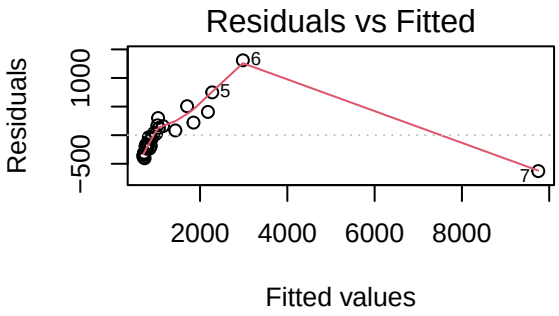
In general, the fit of the data appears to be better that in the previous analysis. The $\hat{\beta}_2$ is higher showing an increased influence of body mass on the response.

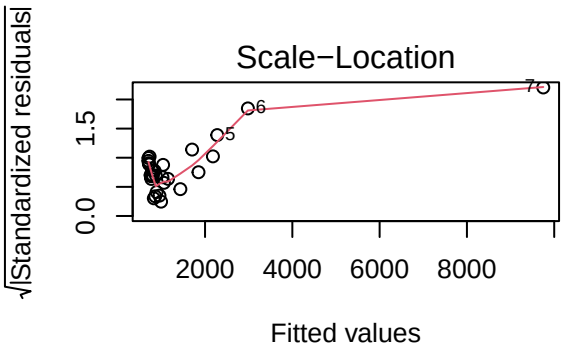
Comparison of diagnostic graphs

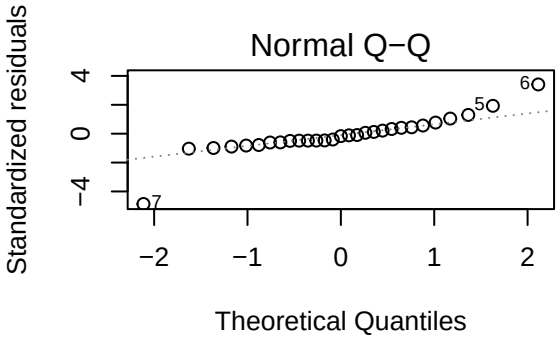
We continue the rest of our analysis by comparing the diagnostic graphs of the regression analysis with ans without the outlier data.

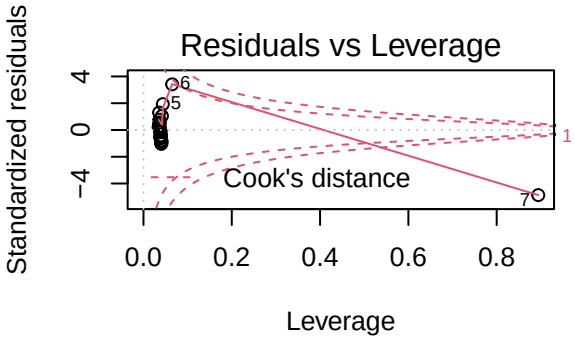
With outlier:

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



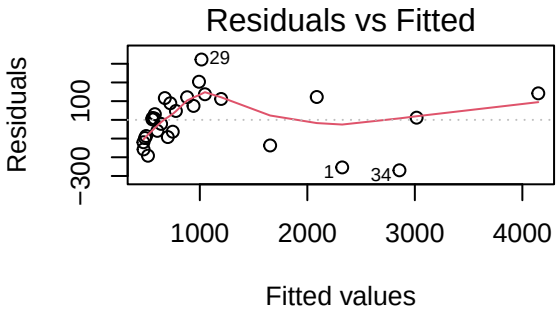


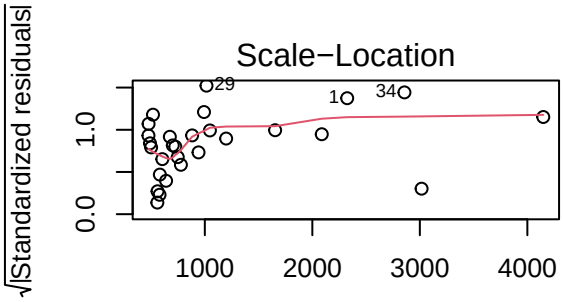


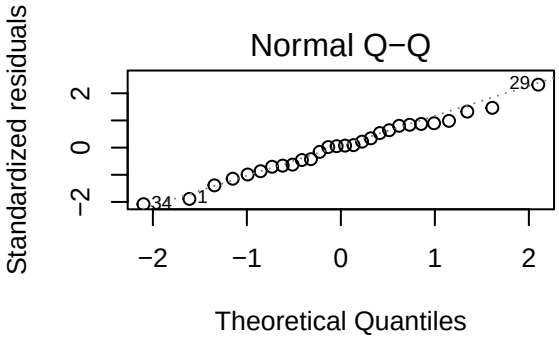


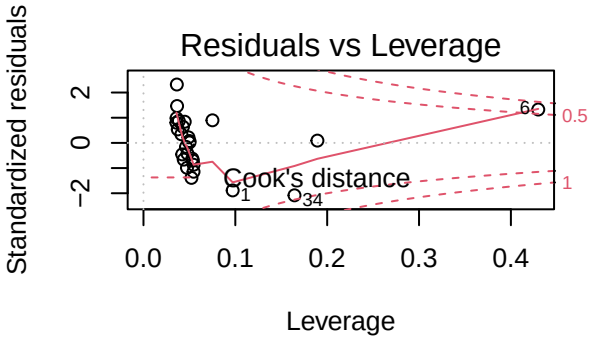
Without outlier:

```
par(mfcol=c(2,2))
plot(reg2)
```









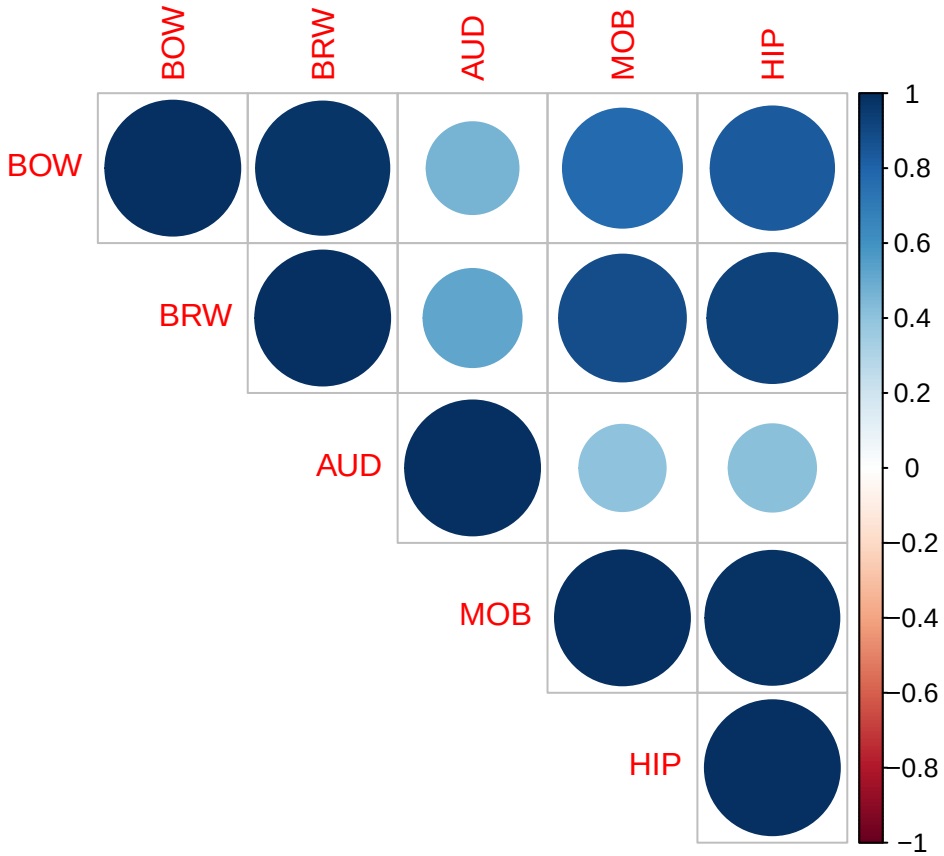
We make the following observations when comparing Graphs 2 and 3 for reg1 and reg2s:

1. We observe in graph 2.1 (Residual vs Fitted), the data points have a rather linear pattern and outlier that alters the pattern. Graph 3.1 does not define a clear pattern.
2. For the Normal Q-Q plots, in general, the plot should follow the bi-sector i.e. the points should le on the bisecting line. While all points are close to the bisector in both Graphs 2 and 3, there is more deviation in the Graph 2 (2.2) especially at the top right. This deviation is not present in the Grap 3 (3.2)
3. For linear regression, the Scale Location graph should be without a clear structure and this is better presented in Graph 3 (3.3)

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

In this part, we try to understand the contribution of each part of the brain to the total weight. The variable to explain is the total weight of the brain (variable BRW). The potentially explanatory variables are the volume of the auditory part of the brain (variable AUD), the volume of the olfactory zone (MOB), and the volume of the hippocampus (HIP).

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



In the following, we will check the impact of combining measurements of different parts of the brain. We do this by check this my computing the correlation of each variable to the prediction.

```
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
```

In general, the smaller the p-value, the higher the correlation of the exploratory variable to the brain weight. From the above results, we see that hippocampus (HIP) and olfactory zone (MOB) variables have very small p-values and correlation above 0.8834 which shows that there is a correlation between those variables and the response. On the other hand, although the p-value of the auditory (AUD) variable is less than 0.05 at 0.03, the correlation is not as strong with only 0.5284 correlation. Therefore, we conclude that HIP and MOB are the variables that produce the highest impact on the response.

Mathematical form

The expression of the model estimated for linear regression is:

$$BRW = \hat{\beta}_1 + \hat{\beta}_2 \times AUD + \hat{\beta}_3 \times MOB + \hat{\beta}_4 \times HIP + \epsilon$$

.

```
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
```

Again we do an analysis of the results of ANOVA by looking at key properties:

1. Intercept: $\hat{\beta}_1 = -312.692$.
2. $\hat{\beta}_2 = 47.989$, $\hat{\beta}_3 = -2.444$, $\hat{\beta}_4 = 15.981$.
3. Multiple regression equation: $BRW = -312.692 + 47.989 \times AUD + -2.444 \times MOB + 15.981 \times HIP + \epsilon$.
4. Test statistics: $T_{\hat{\beta}_1} = 7.652$ and $T_{\hat{\beta}_2} = 22.659$.
5. The p-values for HIP and AUD are 3.85×10^{-8} and 1.52×10^{-5} which are very small values showing that they have an influence on the output.
6. However, MOB has a p-value of 0.46034 indicating that it does not influence the output and should be removed from the model.
7. Based on the above observations, we can say that AUD and HIP have an effect on brain mass.

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

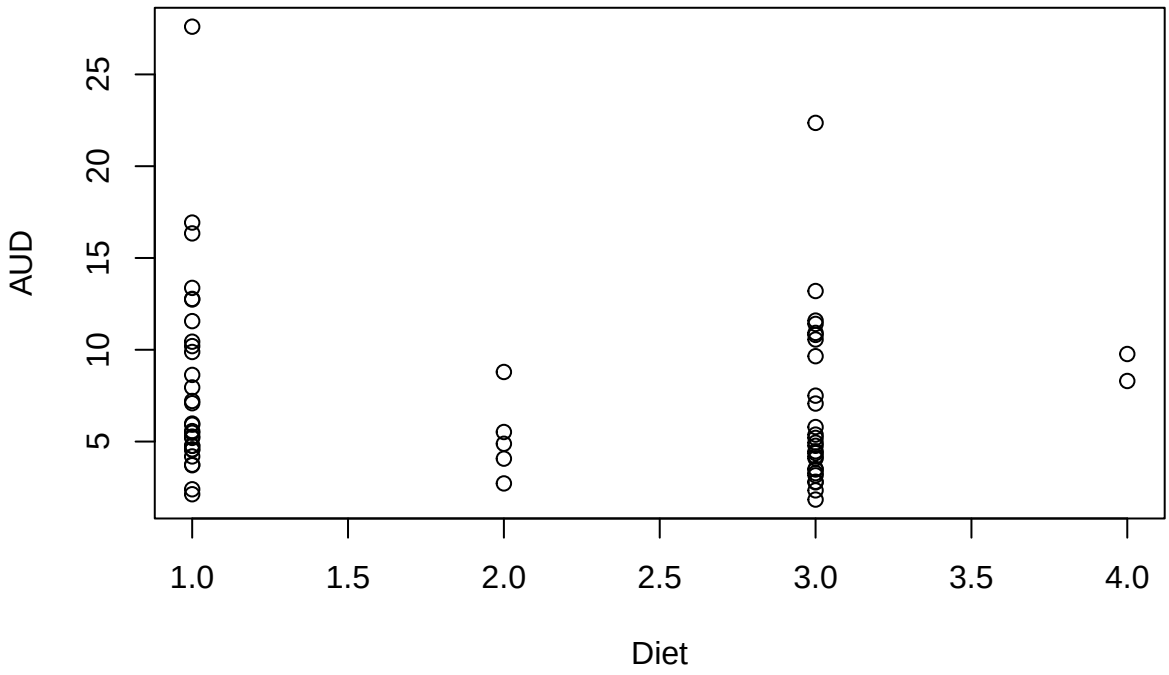
```
## Start:  AIC=433.88
## BRW ~ 1
##
##        Df Sum of Sq      RSS    AIC
## + HIP   1  73272731 11869487 378.74
## + MOB   1  66447848 18694370 391.92
## + AUD   1  23770396 61371823 426.39
## <none>              85142218 433.88
##
## Step:  AIC=378.74
```

```
## 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
```

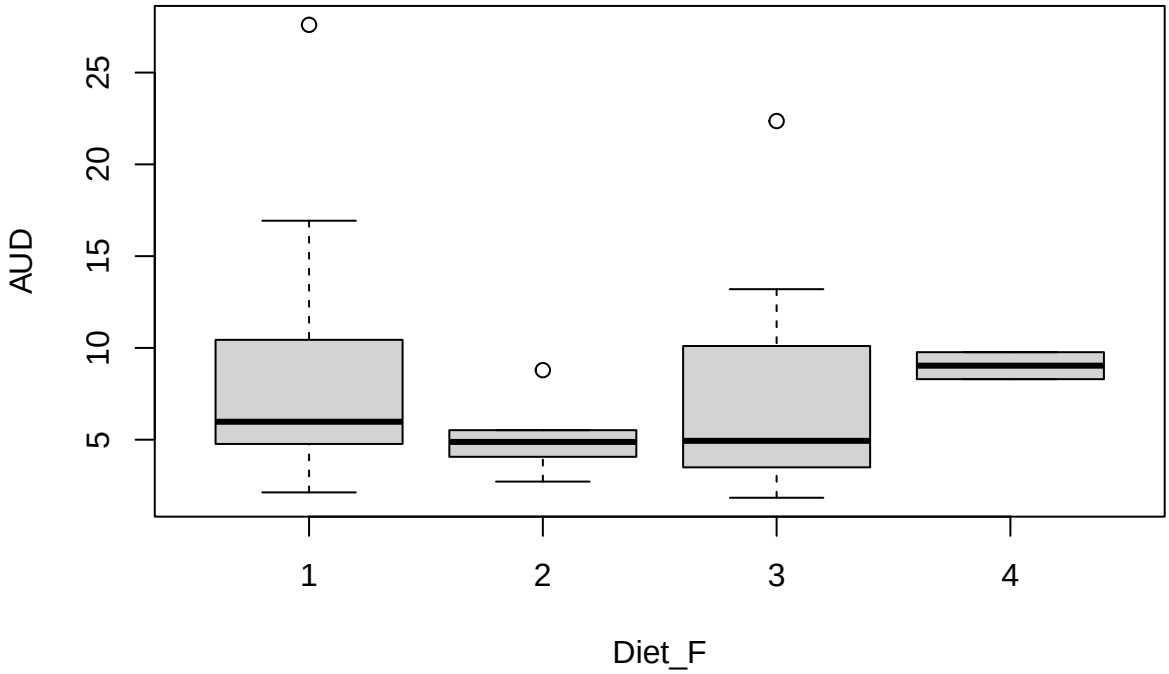
The function `step` chooses a model using the **Akaike information criterion (AIC)** in a stepwise algorithm. Starting from an initial model (in this case `reg0`), the algorithm finds the best model by incrementally adding parameters to the model and selecting the model with the lowest AIC. In this case the best model is `BRW ~ HIP + MOB + AUD`.

4. Link between volume of the auditory part and diet

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




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



While both graphs attempt to present the same data, the first graph shows the individual data points - making it possible to quickly observe important properties of each dietary category - and the second graph uses box plots that are not well-suited to the data. In particular, categories 2 and 4 have very few data points making it difficult to correctly interpret.

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

The p-value of ANOVA is $0.4323 > 0.05$. This indicates that the diet may not have an impact on the auditory volume of the bats. Nonetheless, without further statistical tests, it is difficult to make strong conclusions. Also, the previous graphs show that a straight line may not fit the data correctly so a linear model may not be appropriate.