[II.2313] Analyse de données - TP 3

isep
École d'ingénieurs du numérique

Rémi Biolley – Thierry Lincoln

A. Step by step linear regression using the "bats" data

A.1 Correlation circles

1. First of all, let's open the file. To do so we use the *read.table()* command specifying that the file contains a header

We obtain the following result:

	id	Species	Diet	BOW	BRW	AUD	MOB	HIP
1	1	Rousettus aegyptiacus	1	136.3	2070	9.88	105.77	125.97
2	2	Epomops franqueti	1	120	2210	10.44	107.8	159.8
3	3	Eonycteris spelaea	1	58.7	1310	5.48	67	97.7
4	4	Cynopterus sphinx	1	48.3	1184.33	4.77	65.27	95.4
5	5	Dobsonia praedatrix	1	184	3028	7.09	213.43	233.3
6	6	Glossophaga soricina	1	10.6	414	3.74	12.2	35
7	7	Leptonycteris curasoae	1	24.5	610	5.57	18.6	44.95
8	8	Macroglossus miniumus	1	14.6	561	2.4	30.05	52.95
9	9	Syconycteris australis	1	14.7	570	2.13	31.4	53.1
10	10	Nyctimene albiventer	1	29.7	825	4.56	68.93	81.4
11	24	Brachyphylla cavernarum	1	44.5	1196	8.63	42.2	78.8
12	25	Lionycteris spurrelli	1	9.9	393	3.71	10.3	29.5
13	26	Eidolon helvum	1	262	4290	12.77	208.7	258.1
14	27	Pteropus vampyrus	1	1014	9121	16.93	243.54	331.29
15	28	Anoura geofroyi	1	16	586	5.2	14.15	41.4
16	29	Phylloderma stenops	1	46.1	1338	10.2	87.4	91.7
17	30	Phyllostomus haustatus	1	90.1	1517	12.74	34.33	65.6
18	31	Mimon crenulatum	1	11.8	326	5.92	7.3	18.2
19	32	Trachops cirrhosus	1	36.9	1003	16.34	23.5	50.6
20	33	Tonatia bidens	1	27.67	684.67	13.37	17.96	28.3
21	34	Vampyrum spectrum	1	173	2587	27.6	92	110.4
22	35	Micronycteris brachyotis	1	8.98	319	4.19	13.85	17.1
23	36	Carollia perspicillata	1	17.8	546	5.27	23.55	40.75
24	37	Rhinophylla pumilio	1	8.9	356	4.57	18.8	30.3
25	38	Sturnira lilium	1	20.2	618	4.77	30.77	49.73
26	39	Artibeus lituratus	1	41	1016	7.21	34.38	54.9
27	40	Uroderma bilobatum	1	16.2	612	5.98	28.7	42.7
28	41	Vampyrops vittatus	1	22.6	791	11.56	29.22	52.46
29	42	Chiroderma villosum	1	26.1	814	7.95	28.75	47.58

2. *Comment on the data*

The dataframe is composed of 29 lines and 8 columns.

To every line is assigned an id. This value won't be useful for our study. We also observe that every value in the *Diet* column is set to 1 (which means that the studied bats are all herbivorous, good news for the ones who were afraid for their blood), therefore we won't get any information from this variable.

Finally, we see that the dataframe contains quantitative and qualitative information. This may become a problem as we progress in our study and we'll need to keep an eye on this point.

3. *Prompt the classes of the different attributes using the command str(tab). Remove from tab all attributes that may not be useful for a correlation analysis.*

We need to get rid of the variables which aren't useful for our study. The function *str()* will be a great help to have a more precise idea of what variable isn't necessary to keep. This function gives information about the structure of our dataframe.

There is what is returned by the function:

```
> str(tab)
'data.frame':   29 obs. of  8 variables:
 $ id     : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Species: Factor w/ 29 levels "Anoura geofroyi",..: 22 10 9 6 7 11 12 14 24 17 ...
 $ Diet   : int  1 1 1 1 1 1 1 1 1 1 ...
 $ BOW    : num  136.3 120 58.7 48.3 184 ...
 $ BRW    : num  2070 2210 1310 1184 3028 ...
 $ AUD    : num  9.88 10.44 5.48 4.77 7.09 ...
 $ MOB    : num  105.8 107.8 67 65.3 213.4 ...
 $ HIP    : num  126 159.8 97.7 95.4 233.3 ...
```

The variable *Species* contains factors. Anyway, we didn't need the information contained in this column. We can erase it. As we noticed earlier, we can also get rid of the *id* and the *Diet*.

We have two ways to remove these variables from our dataframe. The first one is to set the columns to *NULL*. But it is also possible to remove every unwanted column in one command line.

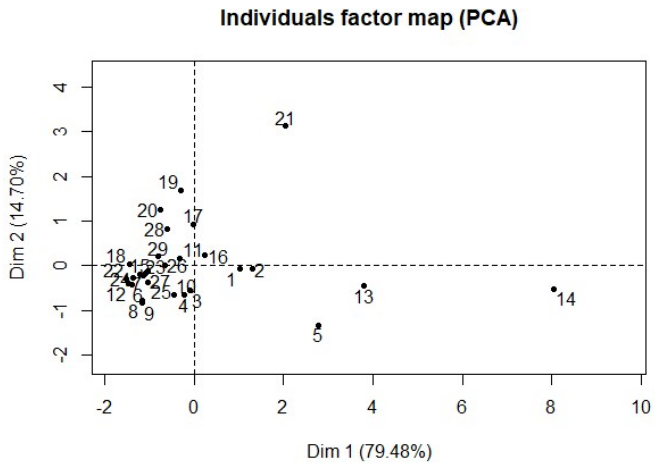
```
tab$id=NULL ; tab$Species=NULL

or

tab=tab[,-(1:3)]
```

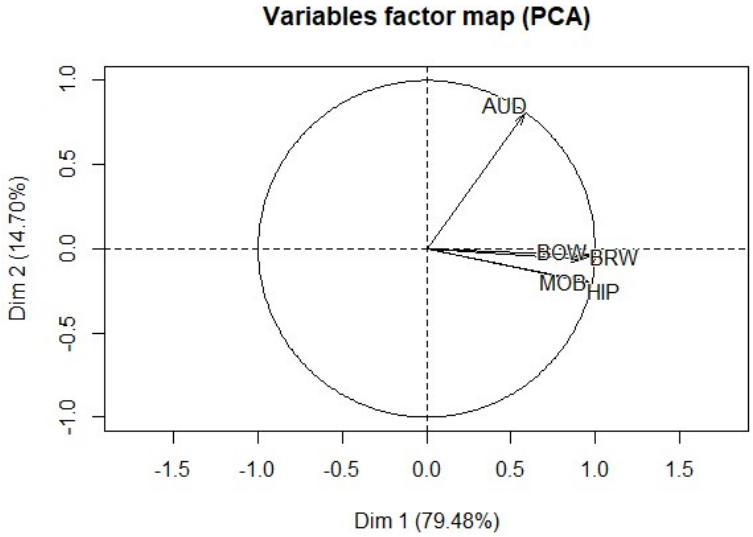
Now that our dataframe is cleaner, we may star our analysis.

4. *We want to do a quick Principal component analysis of this data set and draw the correlation circle in order to find correlated variables*

The command *PCA()* isn't directly available on RStudio. Therefore we need to install the appropriate library *FactoMineR* using the *install.packages()* function. Once it's done, we only have to call this library with the command line *library(FactoMineR)*.

We apply a PCA (Principal Component Analysis) to our dataframe. This gives us the two-following graphics:





The first graphic is used to reduce the dimension of our dataset. We can find redundancy in the correlated variables. This means we can find reduce the dimension of our dataset without losing that much information. The goal of the PCA is to maximize the variance of our data by creating new variables from our old ones. In the graphic, the DIM 1 axis is the new variable which has the highest variance. The DIM 2 axis is the one with the second highest variance.

To create those new variables, we use the highly correlated old variables. The new variables must be as less correlated as possible.

DM1 represents 79.48% of the variance while DM2 represents 14.70% of the variance. It means we can represent approximately 95% of the variance with only two dimensions.

On this graphic we can see that a lot of data are regrouped while some aren't gathered with the rest.

The second graphic is a circle of correlation. It uses the same axes as the previous graphic but also gives information about the correlation of the variables. Let's call ϑ the angle between 2 variables. If $\cos(\vartheta)$ is close to 1 or -1 then the two variables are highly correlated. On the contrary, if $\cos(\vartheta)$ is close to 0 we can say that the variables aren't that correlated.

The distance between the variables and the origin of the axes is also important as it symbolizes the quality of their representation. A distant variable is better represented than a close one.

Knowing this, we can say that the *BOW, BRW, HIP and MOB* variables are highly correlated. Nevertheless, *AUD* is orthogonal to the other variables. Therefore, *AUD* has a low correlation coefficient with the other variables.

```
> results
   name               description
1  "$eig"             "eigenvalues"
2  "$var"             "results for the variables"
3  "$var$coord"       "coord. for the variables"
4  "$var$cor"         "correlations variables - dimensions"
5  "$var$cos2"        "cos2 for the variables"
6  "$var$contrib"     "contributions of the variables"
7  "$ind"             "results for the individuals"
8  "$ind$coord"       "coord. for the individuals"
9  "$ind$cos2"        "cos2 for the individuals"
10 "$ind$contrib"     "contributions of the individuals"
11 "$call"            "summary statistics"
12 "$call$centre"     "mean of the variables"
13 "$call$ecart.type" "standard error of the variables"
14 "$call$row.w"      "weights for the individuals"
15 "$call$col.w"      "weights for the variables"
```

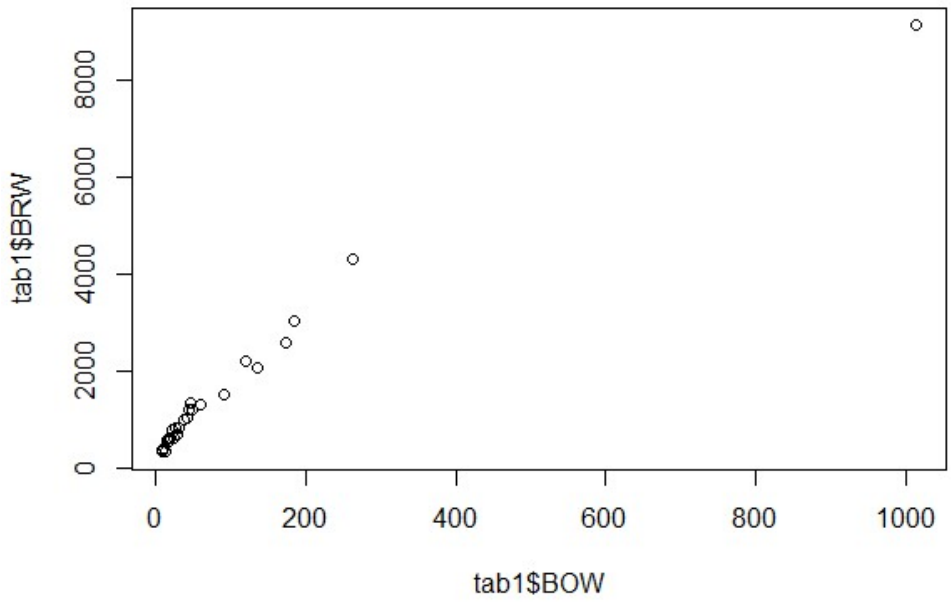
The variable *results* is a list containing variables in its first column and the description of their content in the second column. For example, by typing *results$eig* you can get the eigenvalue, the percentage of variance and the cumulative percentage of variance for every dimension. The information you can get through the variable *result* are useful information used to construct and analyze the PCA.

5. According to the previous question, which variables are the most correlated?

The angles between *MOB, HIP, BOW, BRW* being small, we can infer that the correlation between these variables is strong (cf the construction of the PCA plot). On the contrary *AUD* is lowly correlated with the other variables.

A.2 First linear regression using R

1. **We are interested in finding whether there is a link between the body mass of a bat (BOW) and its brain mass (BRW). When using *plot(tab1$BOW, tab1$BRW)*, the displayed graph is a scatter plot. The dots seem to be regrouped around a straight line. A linear relation may exist between those two variables. Only the data of the *Pteropus Vampyrus* doesn't seem to follow the same relation. We'll keep an eye on this last data.**



Write down the equation of the regression model that seems suited for this data set.

The regression that seems suited for this dataset is the « simple linear regression»

$$\hat{Y} = \widehat{\beta_1}X + \widehat{\beta_0}$$

The values that minimize the RSS are:

$$\widehat{\beta_1} = \frac{COV(X,Y)}{VAR(X)} = r \times \left(\frac{\sigma_Y}{\sigma_X}\right) \; and \; \widehat{\beta_0} = \mu_Y - \widehat{\beta_1}\mu_X$$

They minimize the errors between the predicted values and the observed values.

2. Use the command mod=lm(tab$BRW~tab$BOW) to start the linear regression. Display the variable mod to know the regression coefficients.

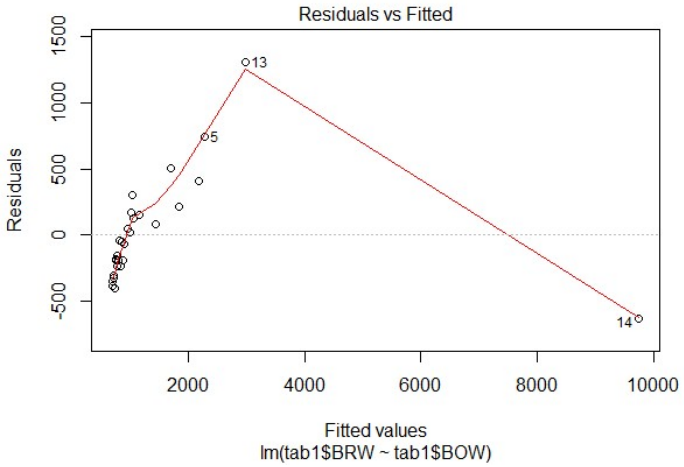
We want to see if there is a linear relation between the bat's mass and the mass of its brain. To do so, we start a linear regression. The command *lm()* gives us the coefficients of the estimated regression model. Using the names previously given to our coefficients, *lm(tab$BRW~tab$BOW)* tells us that $\widehat{\beta_1} = 9.0 \; and \; \widehat{\beta_0} = 623.4$

```
Coefficients:
(Intercept)     tab1$BOW
      623.4          9.0
```

3. Use the command plot(mod). Explain the significance of each diagram displayed. Based on the diagrams, what can you say on the validity of these regression results and on the data?

Plotting *mod* displays 4 different graphs. They give us information about the accuracy of our regression model.

- The "Residuals vs Fitted" graph:

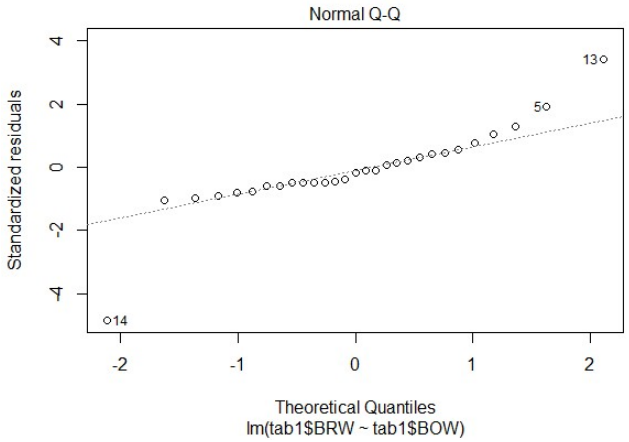


First of all, we have to define what a "residual" and a "fitted" are. A fitted is the value estimated by the linear regression, based on our data. A residual is the difference between the fitted value and the real one (the value stored in our dataset).

The "Residuals vs Fitted" shows us if the residuals have non-linear patterns. We suppose that there's a linear relation between *BRW* and *BOW,* so we expect to get a horizontal line with randomly distributed points around this line.

Unfortunately, we don't get such a line. The source of the problem seems to be the **14th** value of our data set.

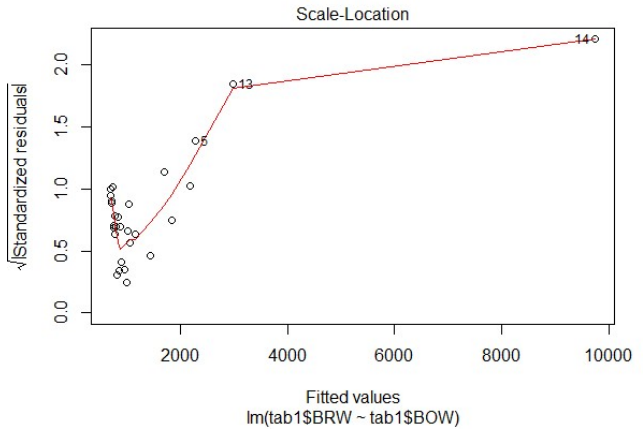
- The "Normal Q-Q" graph:



Thanks to this plot, we can determine if residuals are normally distributed. We get this plot by comparing the position of quantiles in the studied data set to the position of quantiles in the estimated data set.

We want the points to be aligned on the dashed line. That is the case except for the 14th and the 13th values.

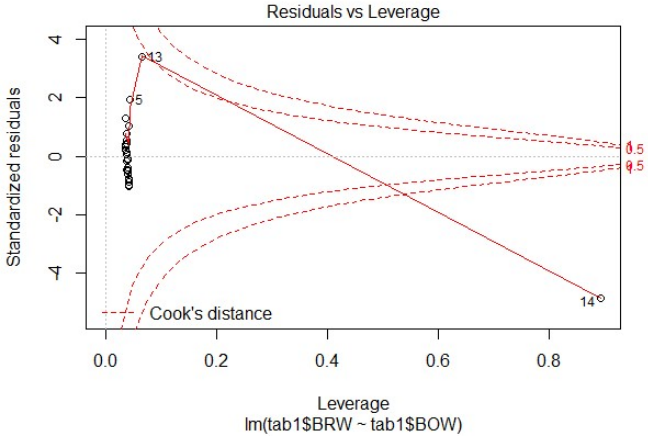
- The "Scale-Location" graph:



The "Scale-Location" graph helps us determine if our residuals have equal variance across the range. That means we can determine the homoscedasticity of our regression.

We want the red line to be horizontal to prove the homoscedasticity. In our case, the line is not horizontal and not even straight. Once again, the 14th value seems to be a factor of problem. The 13th value may be too but to a lesser extent.

- The "Residuals vs Leverage" graph:



This graph shows us the influence of every observation, which means how much our predicted model would change if we were to remove an observation. The influence can be determined thanks to the leverage. The leverage measures the amount by which the predicted value would change if the observation was shifted one unit in the y-direction. A high leverage leads to an important influence.

The observations which are influent are located in the top-right corner or in the bottom-right corner of our graph, outside the dashed lines which represent the Cook's distance.

Only one observation is beyond the dash lines and this observation is the 14th.

In the end, these regression results are not satisfying as we saw in the different graphs. Nevertheless, we noticed that the 14th observation seems to be the one degrading the accuracy of our model. Furthermore, this observation has a high influence on our model. We may try to remove this data from our data set in order to see if it can improve our model.

4. Use the command summary(mod). Explain and comment the results on the residuals. Explain what the R2 coefficient represent in this result. Conclude on the validity of the model.

```
Call:
lm(formula = tab1$BRW ~ tab1$BOW)

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 ***
tab1$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
```

The residuals are contained between –628.32 and 1308.59. This is a very wide interval, meaning the variance of the residuals may be too high. The median value is equal to –65.74. 25% of the residuals are lesser than –233.94, while 25% of them are higher than 158.26. As the median value isn't centered compared to the first and the third quartile *(distance between median and first quartile: 168.2 ; distance between median and third quartile: 224)*, we can say that the residuals aren't strictly following a normal distribution. That confirms what we saw in the Normal Q-Q graph.

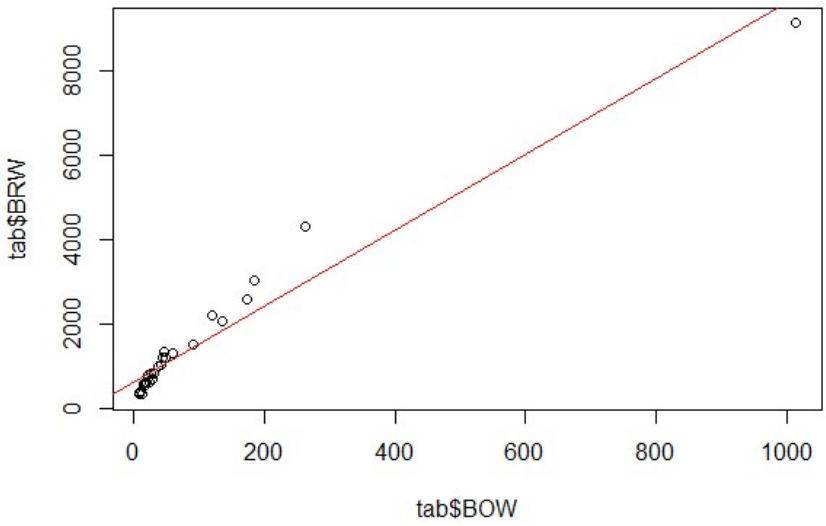
The R^2 coefficient is also called the determination coefficient. It measures the quality of the prediction of a linear regression. R^2 is written as a percentage between 0 and 100%, 0% meaning that the model isn't explaining the variability of the response data around its mean.

The adjusted R^2 coefficient takes account of the number of variables to see if adding or removing variables may improve the accuracy of our model. Its value is always lower than R^2.

In our case, the R^2 coefficient and the adjusted R^2 coefficient are very close to 1. This means we can be satisfied of our variables.

The model seems to be valid but can be somehow improved (particularly its homoscedasticity).

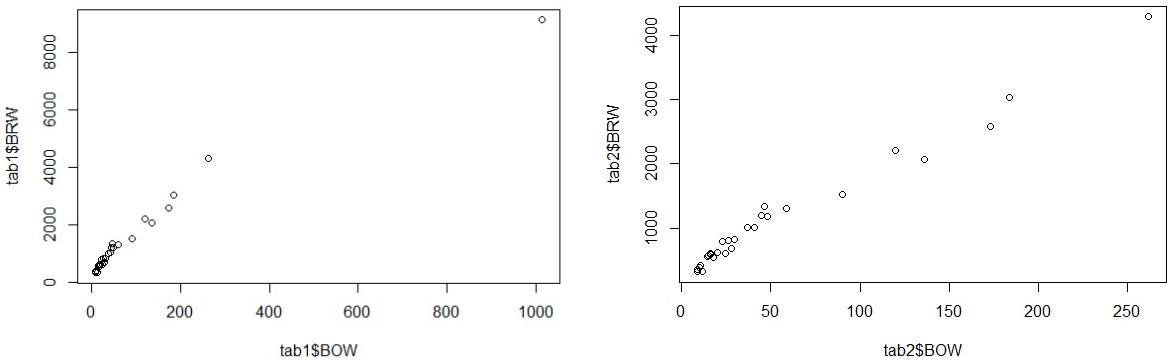
5. Use the following commands: plot(tabBOW,tabBRW) abline(coef(mod),col=”red”) Comment the resulting graph.



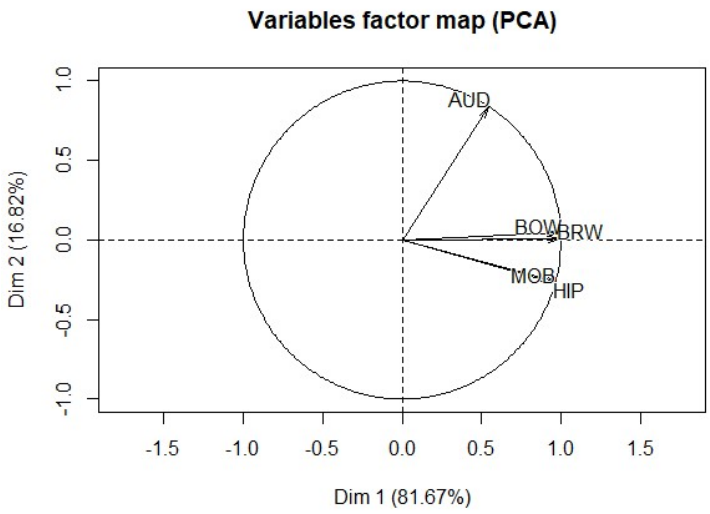
The estimated model is fitting our values well. We could be satisfied but it would be even better if the slope of the line was a little higher. Indeed, it seems a little bit too low as the line has to try to fit the value in the top-right corner. Getting rid of this value should help getting a better modelization. This value is the 14th, the one we were keeping an eye on and the one we wanted to remove when we were analyzing the fours graphs.

A.3 Second linear regression

1. Create a new array tab2 from which you will remove "Pteropus vampyrus".

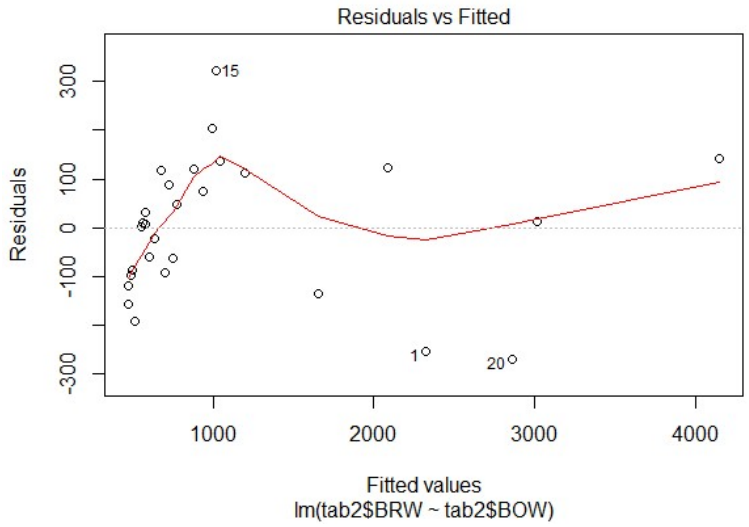


Now that we removed the values of "Pteropus vampyrus", the linear relation between *BOW* and *BRW* seems to be even more visible. Furthermore, we now represent 98.5% of the variance with two dimensions. We are even more precise with our two new dimensions than with our previous data set.

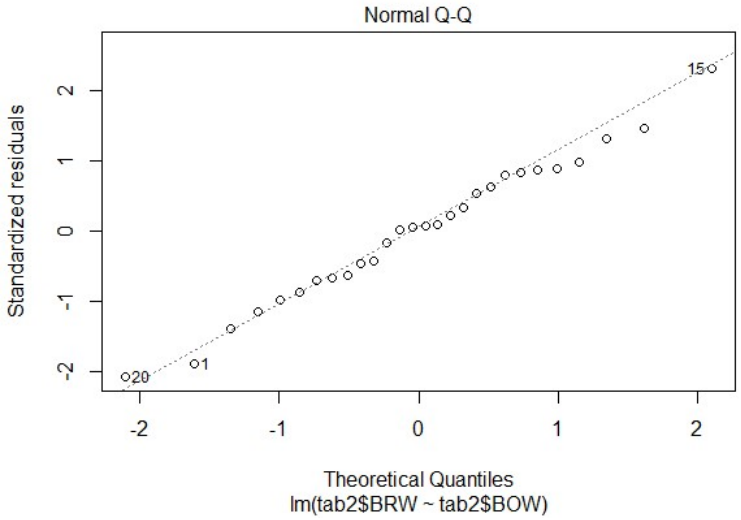


We can see on the correlation circle that the angle between BOW and BRW is wider than in our previous model. Still, the angle is very small which means the correlation coefficient remains high.

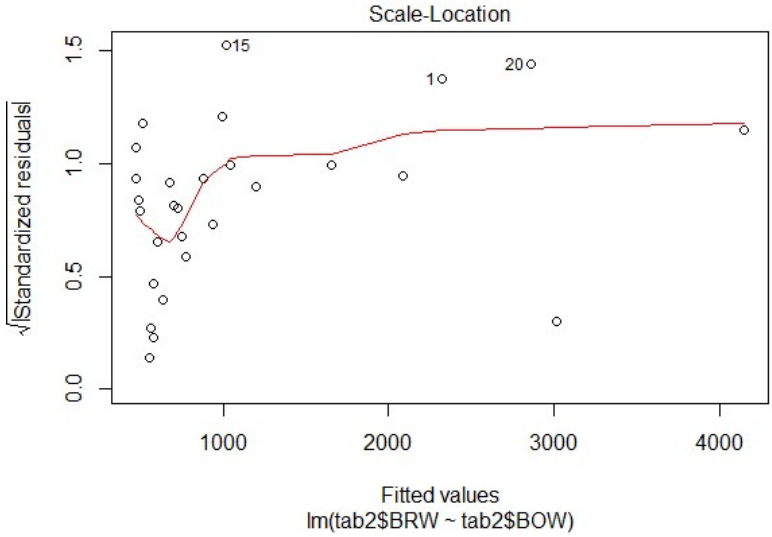
2. Do again questions B-2 to B-5 using tab2.



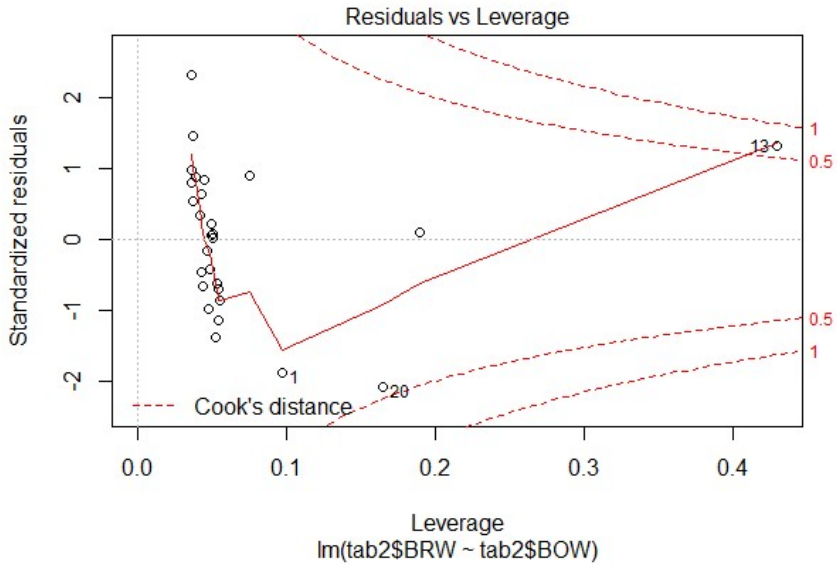
This Residuals vs Fitted graph is way more satisfying than the previous one. The line is nearly when its points have an abscissa above 1500. The problem is that, before the line gets horizontal, we have a lot of observations, meaning that it may be hard to correct the line.



The observations of the Normal Q-Q are nearly perfectly fitting the dashed line. Therefore, the residuals are normally distributed. This has been improved compared to our previous model.



Now that we got rid of the observation of "Petropus vampyrus", the Scale-Location graph is closer to a horizontal line. The line is horizontal after X-axis passes 1000. Before that moment, there are unwanted variations. In the same way as the Residuals vs Fitted graph, we can't really isolate an observation that would be responsible for these variations.



We notice that we don't have any observation beyond the Cook's distance. Only the 13th value is close to cross the line. This means the 13th has a bigger impact than the rest of the values on the model. Still, this impact is less important than the old 14th observation's one.

Regarding the graphs, the results are more satisfying than they used to be when we didn't have removed the "Petropus vampyrus". Still, we don't have perfect results for the Residuals vs Fitted and the Scale-location graphs. Nevertheless, we can't say without a risk that any of the remaining values is responsible for this gap between what we are expecting and what we're getting. It's even more true considering that the values don't have individually a huge impact on the model.

We now have a look to the values obtained thanks to the *summary()* function applied to our new linear regression.

```
Call:
lm(formula = tab2$BRW ~ tab2$BOW)

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 ***
tab2$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
```

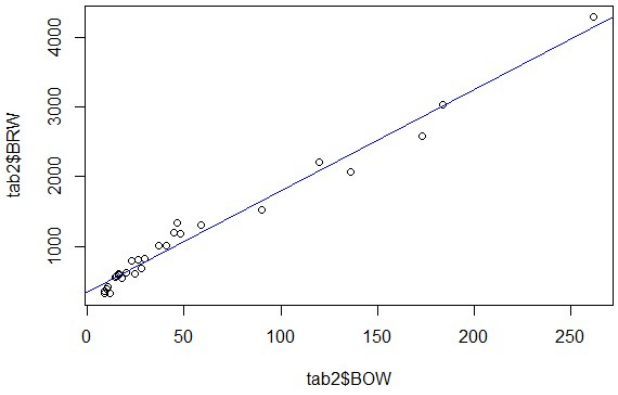
First, we notice that the interval containing the values of the residuals is narrower than before ([-269.76,322.55] versus [–628.32,1308.59] previously). This is an indicator that the variance of the residuals is lower than before. Furthermore, the median (8.73) is centered compared to the first and the third quartile

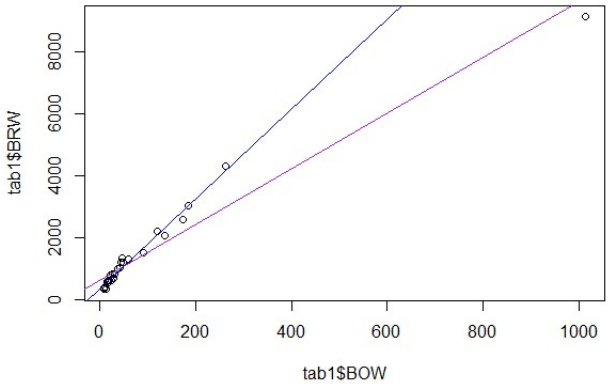
(distance between median and first quartile: 102.03 ; distance between median and third quartile: 104.2). These values are coherent with our desire to get normally distributed residuals.

The multiple R^2 coefficient is higher with a value of 0.9778 (previously it was 0.95). Same with the adjusted R^2 coefficient which has a value of 0.977 (0.948 previously). The quality of our predicted linear regression is even better than before. That shows how influent the value we removed was.

In the end, the results of this second regression are better than the first one. The homoscedasticity of the regression is more respected, the coefficient of determination is higher, the residuals now follow a normal distribution according to the Normal Q-Q graph and generally have smaller values. We can be satisfied of this result as we don't see any observation obviously misrepresenting our results.

3. Use the following commands: plot(tabBOW,tabBRW) abline(coef(mod),col="red") abline(coef(mod2),col="blue") Comment the resulting graph.





The blue line represents our new model while the purple one represents the old model.

We can see that the new model is closer to every observation than the older one, except for the value that we removed.

We also can see how a single observation can influence a model.

Why doesn't the observation of "Petropus vampyrus" fit the model? Maybe was there a mistake when they took the values for this bat. We may also suppose that this species isn't following the same standards than the other present species in our data set.

In the end, the existence of a linear relation between the mass of a bat and the mass of its brain seems probable.

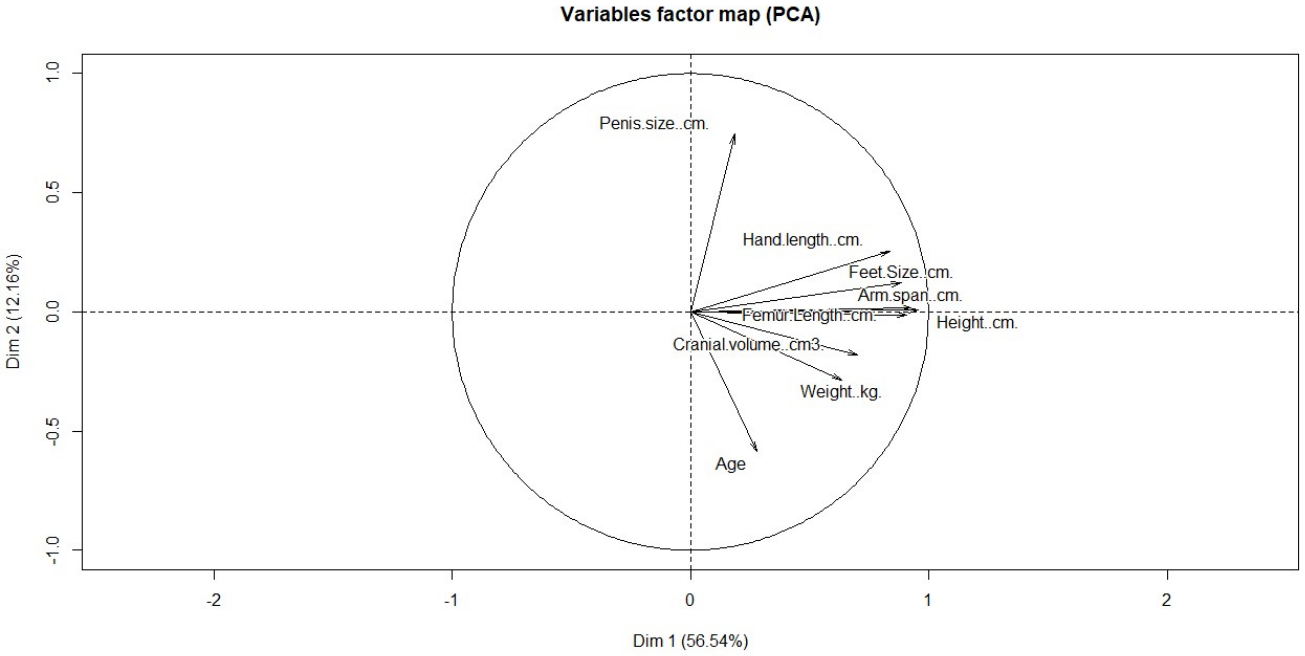
B. Application to the mansize dataset

1. Remind the correlation between the different variable and confirm them by projecting the data into the PCA plan. Use correlation circles to illustrate your results.

	Age	Height..cm.	Weight..kg.	Femur.Length..cm.	Feet.Size..cm.	Arm.span..cm.	Hand.length..cm.	Cranial.volume..cm3.
Age	1.00000000	0.1980260	0.14680238	0.2125544	0.2267084	0.2217911	0.1663874	0.1789928
Height..cm.	0.19802602	1.0000000	0.59151560	0.8905730	0.8024375	0.9032031	0.7915676	0.6246087
Weight..kg.	0.14680238	0.5915156	1.00000000	0.5170937	0.4394847	0.5605218	0.2186416	0.5999177
Femur.Length..cm.	0.21255435	0.8905730	0.51709372	1.0000000	0.7542053	0.8232579	0.7420294	0.5800319
Feet.Size..cm.	0.22670837	0.8024375	0.43948467	0.7542053	1.0000000	0.7583445	0.8710756	0.5046619
Arm.span..cm.	0.22179113	0.9032031	0.56052176	0.8232579	0.7583445	1.0000000	0.7951266	0.5599201
Hand.length..cm.	0.16638740	0.7915676	0.21864163	0.7420294	0.8710756	0.7951266	1.0000000	0.3861985
Cranial.volume..cm3.	0.17899279	0.6246087	0.59991769	0.5800319	0.5046619	0.5599201	0.3861985	1.0000000
Penis.size..cm.	-0.07167913	0.1273751	0.06840311	0.1005534	0.1763982	0.1401550	0.1827799	0.1242201

	Penis.size..cm.
Age	-0.07167913
Height..cm.	0.12737507
Weight..kg.	0.06840311
Femur.Length..cm.	0.10055335
Feet.Size..cm.	0.17639823
Arm.span..cm.	0.14015502
Hand.length..cm.	0.18277993

We've recovered the correlation matrix from TP2. From this table we can say that a correlation seems to exist between the femur length, the height, the arm span, the feet size and the hands size. The age, cranial volume, weight and the size of the penis appear to be not acceptable for determining characteristics of the human body. Let's put this in perspective with the PCA (Principal Component Analysis) of the same dataset.

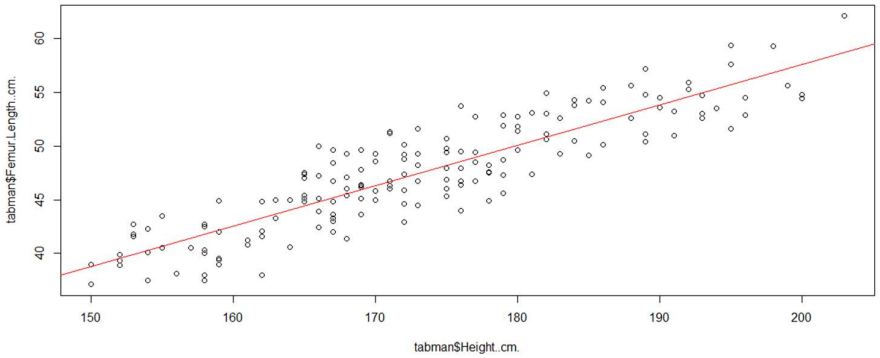


Of course, the PCA plot reflects the correlation coefficients that we say above, we can analyze it a bit further nevertheless. The Sum of Dim 1 & Dim 2 is giving us 68.7, It means we can represent approximately 69% of the variance with only two dimensions. This result is definitely not perfect as in the first exercise we had 95% covered by two dimensions.
We can also see that the age, cranial volume, weight and the size of the penis are not best represented in this PCA plot as the distance between these variables and the origin of the axe tend to be shorter than the "better" correlated values.

2. ***Run a linear regression to predict the size of an individual based on the size of his femur bone.***

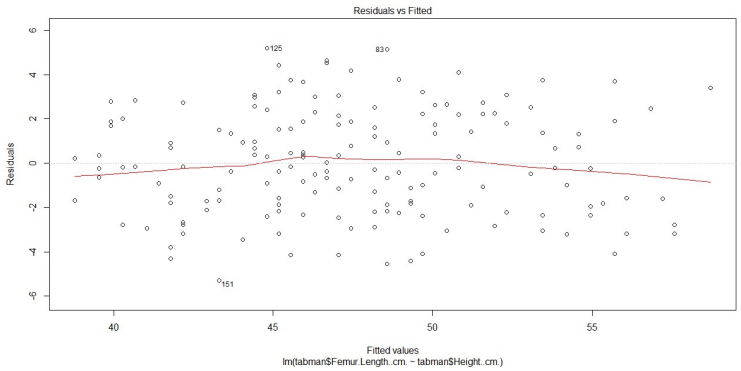
```
modman

Coefficients:
       (Intercept)  tabman$Height..cm.
          -17.5939              0.3759
```

Here are the coefficients for the graph. With $\hat{\beta}_0 = -17.59$ *and* $\widehat{\beta_1} = 0.37$ following a simple regression model $\hat{Y} = \widehat{\beta_1}X + \widehat{\beta_0}$

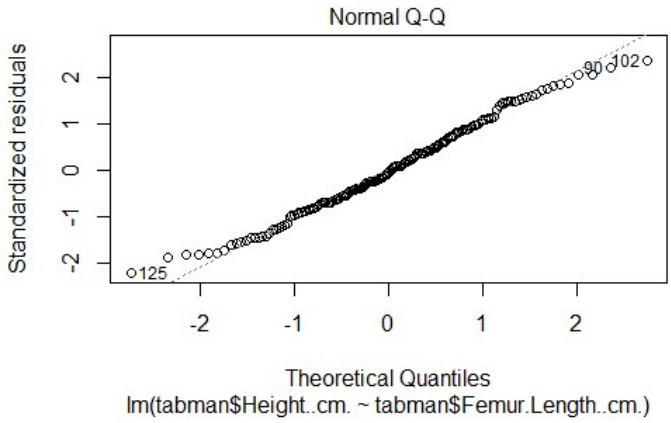


Here is the plot with the line (in red) representing the linear regression, we must now proceed with an assessment of this linear regression.

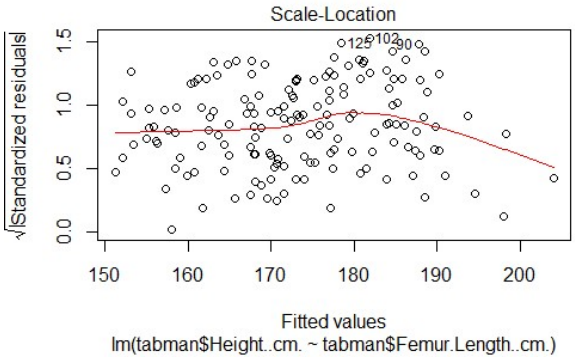
3. ***Comment the regression results by focusing on the different graphic and indexes computed by R.***



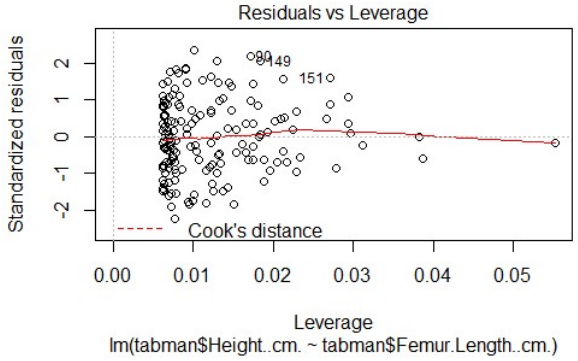
The "Residuals vs Fitted" shows us here that the residuals have a linear pattern. We can see that there's a linear relation between *Femur.Length and Height, has we got here* a horizontal line with randomly distributed points around this line.



Thanks to this Normal Q-Q plot, we can determine that here residuals are normally distributed, most points are aligned on the dashed line, only some are slightly diverging in the extremities.



As the red line is mostly horizontal across the range of fitted values, it proves the homoscedasticity of our regression. We can however see a considerable loss in residuals variance equality above the 190 value, proving again, that in the extreme values our regression is not perfect.



We notice that we don't have any observation beyond the Cook's distance. The leverage values here reside in a very tight interval, it's a low leverage thus leading to a negligible influence. No specific values influence much the model.

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.2975 -1.8597 -0.1699  1.8714  5.1990

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        -17.59389    2.64382  -6.655 4.35e-10 ***
tabman$Height..cm.   0.37587    0.01522  24.689  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.378 on 159 degrees of freedom
Multiple R-squared:  0.7931,   Adjusted R-squared:  0.7918
F-statistic: 609.6 on 1 and 159 DF,  p-value: < 2.2e-16
```

The residuals are contained between –5.3 and 5.1. This is a narrow interval, meaning the variance of the residuals are quite low. The median value is equal to –0.2. 25% of the residuals are lesser than –1.85, while 25% of them are higher than 1.87. The median value is centered compared to the first and the third quartile *(distance between median and first quartile: 1.69; distance between median and third quartile: 1.7)*, we can say that the residuals are following a normal distribution. That confirms what we saw in the Normal Q-Q graph.

The R^2 coefficient and the adjusted R^2 coefficient are close to 80%, which means that the model does fit our data, but not so well.

To conclude this analysis of our regression model. We can probably say that it is a proper model. The homoscedasticity of the regression is quite good, the coefficient of determination R^2 is high, the residuals do follow a normal distribution according to the Normal Q-Q graph.