Regression Analysis I

Andrea Huerfano

August 19, 2019

```
library(GLMsData)
data(AIS)
library(dplyr)
AIS_f<-AIS%>%filter(Sex=='F')
```

This file is one of the regression analysis papers. Here you can find:

- Plots
- Correlations
- Identifying the best model
- Influence and leverage points
- Residuals examination
- Variance examination

Physical measurements and blood measurements from high performance athletes at the AIS, this dataset contain information about 98 athletes and it is a subset of the data(AIS) in the library GLMsData. The function *summary* was used to obtain the following result

```
library(readxl)
colnames(AIS_f)
```

```
##  [1] "Sex"   "Sport" "LBM"   "Ht"    "Wt"    "BMI"   "SSF"   "RBC"
##  [9] "WBC"   "HCT"   "HGB"   "Ferr"  "PBF"
```

```
atletas<-subset(AIS_f,select = c("LBM", "Ht", "Wt", "BMI", "SSF"))
```

The summary function is an excelent option to see a brief description of each variable, finding here all of the quantiles.

```
summary(atletas)
```

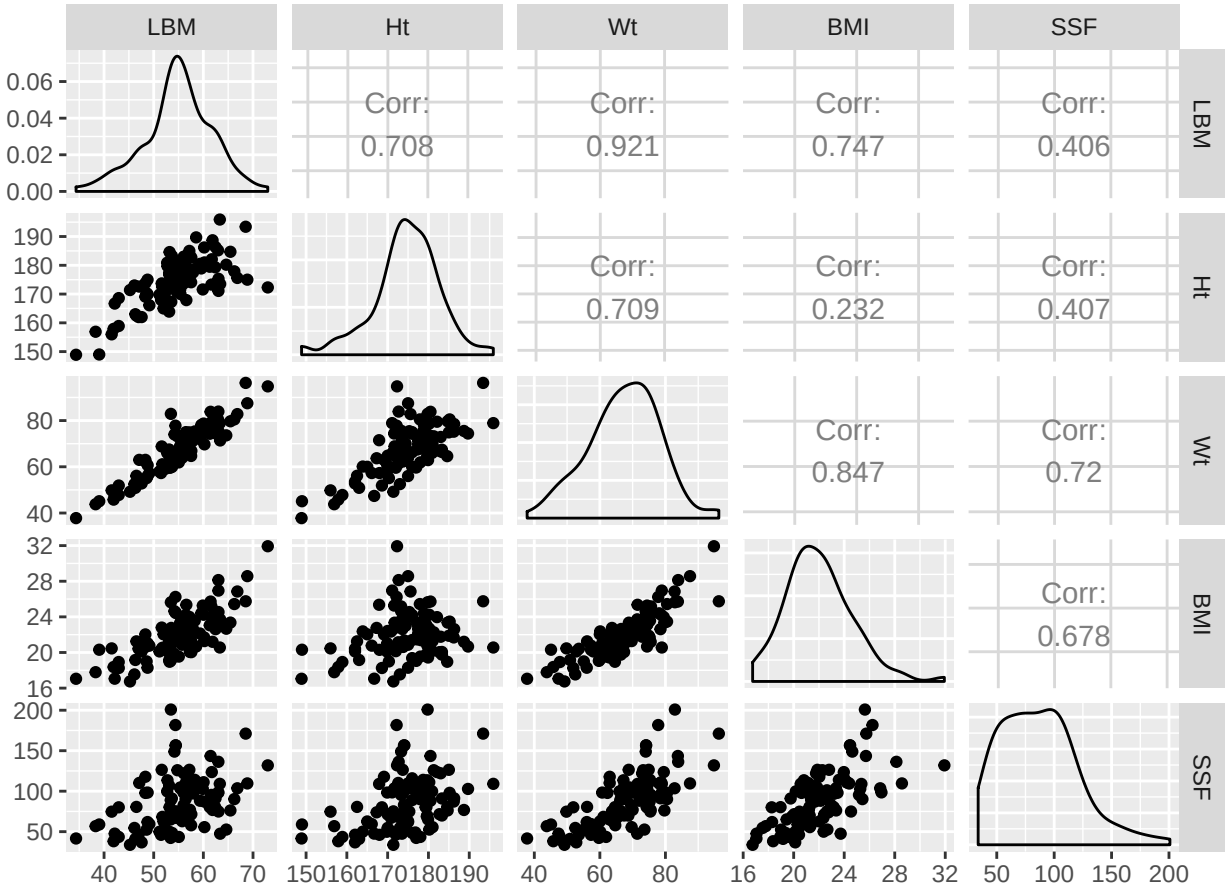
```
##       LBM              Ht              Wt              BMI
##  Min.   :34.36   Min.   :148.9   Min.   :37.80   Min.   :16.75
##  1st Qu.:51.93   1st Qu.:171.0   1st Qu.:60.08   1st Qu.:20.27
##  Median :54.92   Median :175.0   Median :68.05   Median :21.82
##  Mean   :54.89   Mean   :174.6   Mean   :67.34   Mean   :21.99
##  3rd Qu.:59.40   3rd Qu.:179.7   3rd Qu.:74.42   3rd Qu.:23.39
##  Max.   :72.98   Max.   :195.9   Max.   :96.30   Max.   :31.93
##       SSF
##  Min.   : 33.80
##  1st Qu.: 59.27
##  Median : 81.80
##  Mean   : 86.97
##  3rd Qu.:107.42
##  Max.   :200.80
```

Scatterplot matrix

Another good option is looking a matrix with all the scatterplots. For example we can see that Wt and BMI have a linear behavior with the response variable (LBM), and the other variables even when some of them

have an aproximate linear behavior is easy to see that these ones are more dispersed, for example: Ht and SSF.

```
library(GGally)
ggpairs(atletas, columns = 1:ncol(atletas), title = "",
  axisLabels = "show", columnLabels = colnames(atletas))
```

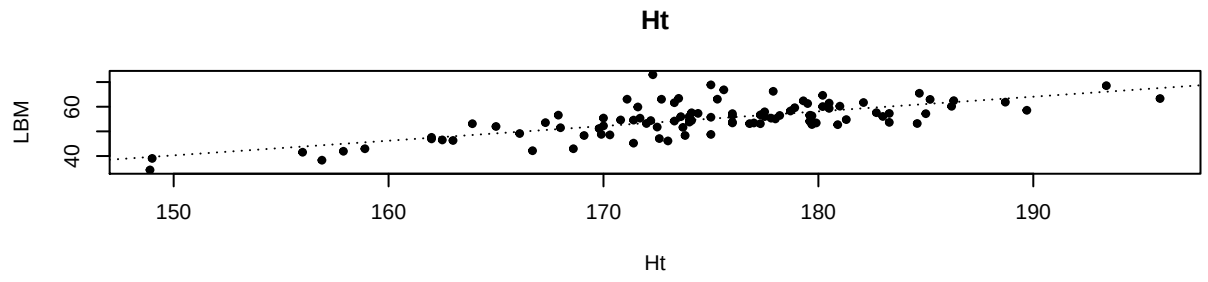


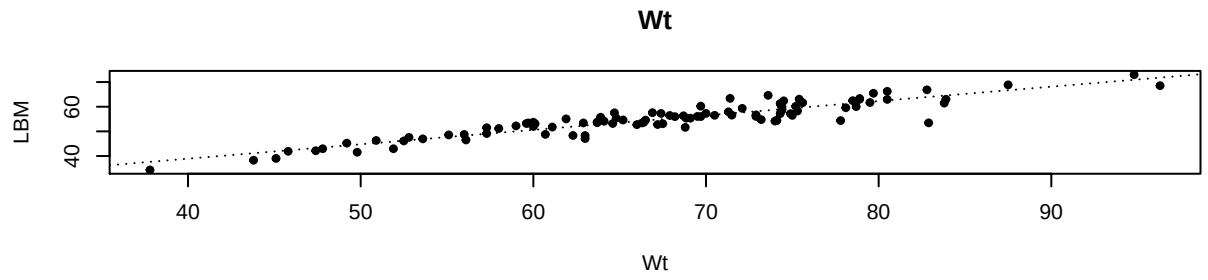
Correlations

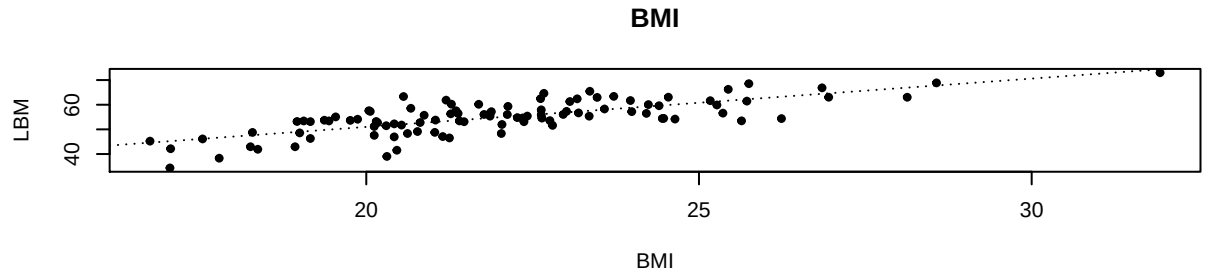
The simple correlation is a measure used to determine the strength and the direction of the relationship between two variables. In this example the independent variable is lbm and the other ones are the dependent variables.

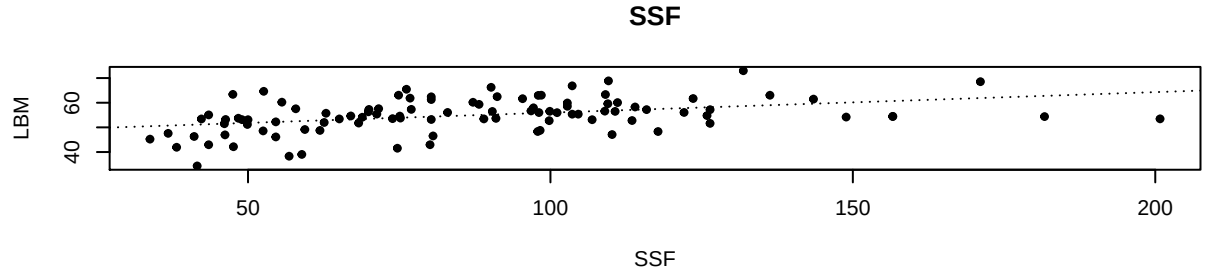
The scatterplots show a strong lineal asociation between ht and lbm that is reflected for a correlation of 0.7082, the second scatterplots shows a strong relation between wt and lbm that is respalded for a correlation of 0.9207. Between the BMI and lbm there is a correlation of 0.7474 which represent a strong relation as well. Another imporant fact is the strong relation reflected for a correlation of 0.847 between the BMI and wt, that means that some of the information in the variable BMI are in wt and viceversa, furthemore the correlation between wt and ht is really hight (0.7087).

```
fit <- lm(LBM ~ 1+Ht+Wt+BMI+SSF, data=atletas) ##This model contain all of the variables
##It is used to determine the simple and partial correlations
```









,

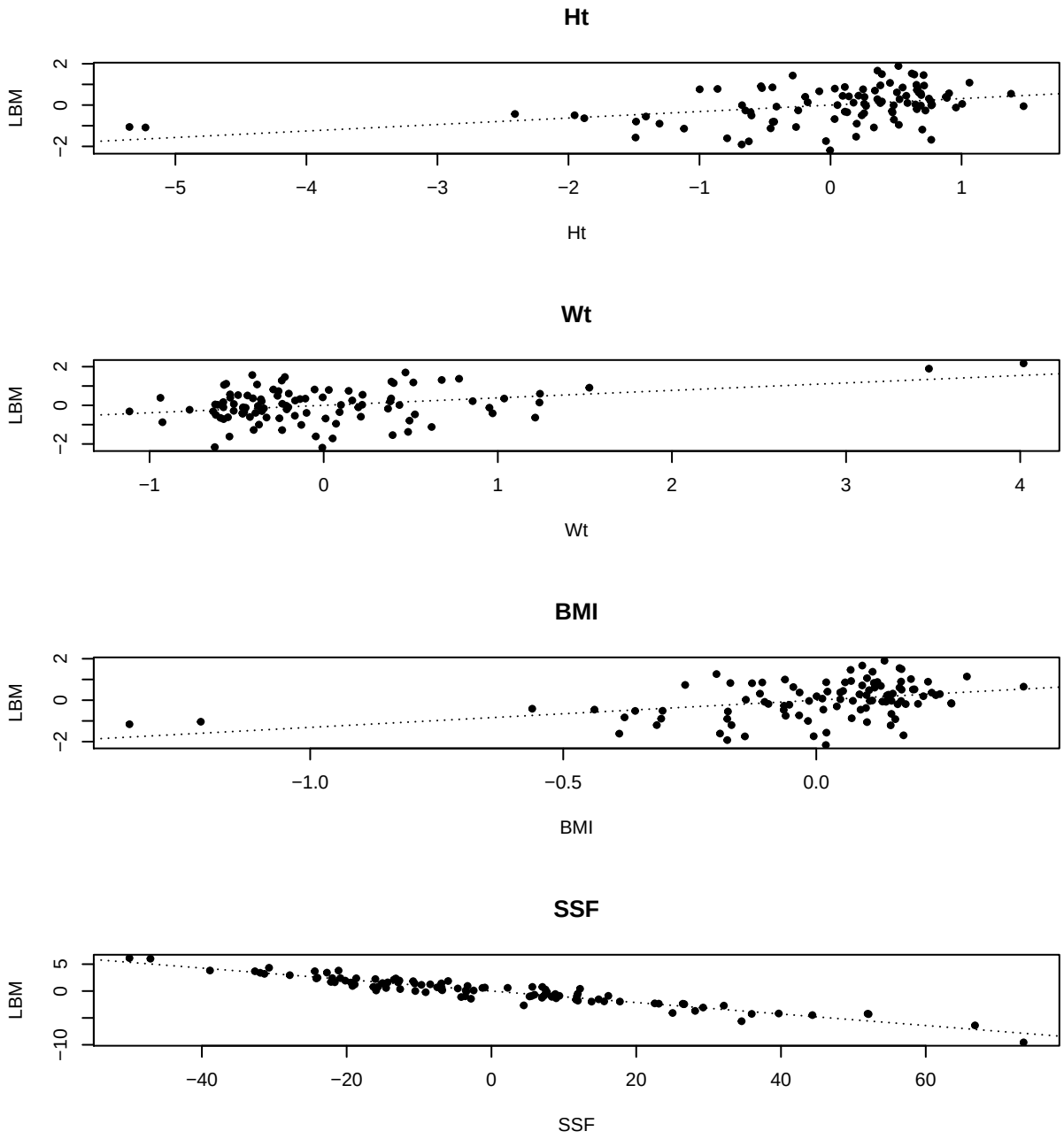
```
##           LBM        Ht        Wt       BMI       SSF
## LBM 1.0000000 0.7082934 0.9207976 0.7474915 0.4064912
## Ht  0.7082934 1.0000000 0.7087400 0.2316648 0.4065155
## Wt  0.9207976 0.7087400 1.0000000 0.8470335 0.7196649
## BMI 0.7474915 0.2316648 0.8470335 1.0000000 0.6784880
## SSF 0.4064912 0.4065155 0.7196649 0.6784880 1.0000000
```

Partial correlation

The partial correlation for ht and lbm is shorter than the obtained in the plot of simple correlation, that means that part of the information of the ht varaible over the response variable is also conteined in the other

dependent variables . The same result appears when the simple correlation coeficient is compared 0.7082 with 0.377. This situations occurs in the same way with Wt having 0.9207 vs 0.338. Finally, rcc has a partial correlation coefficient of 0.195. This correlations describe the information that has every variable and there is not conteined in ahother one.

```
Correlaciones.parcial(fit,4,1,1,"")
```




```
##      [,1]        [,2]
## [1,] "Variables" "Respuesta"
## [2,] "Ht"        "0.377"
## [3,] "Wt"        "0.338"
## [4,] "BMI"       "0.382"
```

```
## [5,] "SSF"       "-0.949"
```

```
library(MASS)
library(zoo)
library(lmtest)
```

Possible models

Models with one, two and 3 parameters are examined using the quality mesures: SCRes, R2, R2 ajust and AIC. Maximum R2 and Shorter AIC

```
ajuste.normal(fit,1)
```

```
##      [,1]          [,2]  [,3]     [,4]    [,5]       [,6]
## [1,] "Modelo"      ""    "SCRes"  "R2"    "R2 Ajust" "AIC"
## [2,] "(Intercept)" "Ht"  "2363.9" "0.502" "0.497"    "606.1"
## [3,] "(Intercept)" "Wt"  "721.7"  "0.848" "0.846"    "487.4"
## [4,] "(Intercept)" "BMI" "2093.2" "0.559" "0.554"    "593.9"
## [5,] "(Intercept)" "SSF" "3959.9" "0.165" "0.157"    "657.7"
```

```
ajuste.normal(fit,2)
```

```
##      [,1]          [,2]  [,3]  [,4]     [,5]    [,6]       [,7]
## [1,] "Modelo"      ""    ""    "SCRes"  "R2"    "R2 Ajust" "AIC"
## [2,] "(Intercept)" "Ht"  "Wt"  "692.1"  "0.854" "0.851"    "485.2"
## [3,] "(Intercept)" "Ht"  "BMI" "657.8"  "0.861" "0.858"    "480.2"
## [4,] "(Intercept)" "Ht"  "SSF" "2284"   "0.519" "0.509"    "604.6"
## [5,] "(Intercept)" "Wt"  "BMI" "704"    "0.852" "0.849"    "486.9"
## [6,] "(Intercept)" "Wt"  "SSF" "75.9"   "0.984" "0.984"    "264.2"
## [7,] "(Intercept)" "BMI" "SSF" "2004.1" "0.578" "0.569"    "591.6"
```

```
ajuste.normal(fit,3)
```

```
##      [,1]          [,2] [,3]  [,4]  [,5]    [,6]    [,7]       [,8]
## [1,] "Modelo"      ""   ""    ""    "SCRes" "R2"    "R2 Ajust" "AIC"
## [2,] "(Intercept)" "Ht" "Wt"  "BMI" "653.3" "0.862" "0.858"    "481.5"
## [3,] "(Intercept)" "Ht" "Wt"  "SSF" "75.9"  "0.984" "0.983"    "266.2"
## [4,] "(Intercept)" "Ht" "BMI" "SSF" "73.2"  "0.985" "0.984"    "262.6"
## [5,] "(Intercept)" "Wt" "BMI" "SSF" "75.6"  "0.984" "0.984"    "265.8"
```

```
ajuste.normal(fit,4)
```

```
##      [,1]          [,2] [,3] [,4]  [,5]  [,6]    [,7]    [,8]
## [1,] "Modelo"      ""   ""   ""    ""    "SCRes" "R2"    "R2 Ajust"
## [2,] "(Intercept)" "Ht" "Wt" "BMI" "SSF" "64.9"  "0.986" "0.986"
##      [,9]
## [1,] "AIC"
## [2,] "252.5"
```

Over this results the best model will be that which contain all of the variables, however the model which contains just Wt and BMI has almost the same R2, furthemore, notice that the variable SSF seems with non homogenize variance, after running some models the best results were obtained appling log over there.

```
fit <- lm(LBM ~ 1+Wt+log(SSF), data=atletas)
summary(fit)
```

```
##
## Call:
```

```
## lm(formula = LBM ~ 1 + Wt + log(SSF), data = atletas)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0308 -0.5960  0.0031  0.8787  2.4240
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.07183    1.51368   25.81   <2e-16 ***
## Wt           0.83149    0.01795   46.31   <2e-16 ***
## log(SSF)    -9.14915    0.49798  -18.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.289 on 97 degrees of freedom
## Multiple R-squared:  0.966,  Adjusted R-squared:  0.9653
## F-statistic:  1380 on 2 and 97 DF,  p-value: < 2.2e-16
```

```
AIC(fit)
```

```
## [1] 339.4675
```

Now,see this model without intercept

```
fit2 <- lm(LBM ~ -1+Wt+log(SSF), data=atletas)
summary(fit2)
```

```
##
## Call:
## lm(formula = LBM ~ -1 + Wt + log(SSF), data = atletas)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5921  -1.5619   0.5926   2.8043   6.8489
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## Wt        0.69927    0.04802  14.561   <2e-16 ***
## log(SSF)  1.71290    0.74312   2.305   0.0233 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.596 on 98 degrees of freedom
## Multiple R-squared:  0.9959, Adjusted R-squared:  0.9958
## F-statistic: 1.178e+04 on 2 and 98 DF,  p-value: < 2.2e-16
```

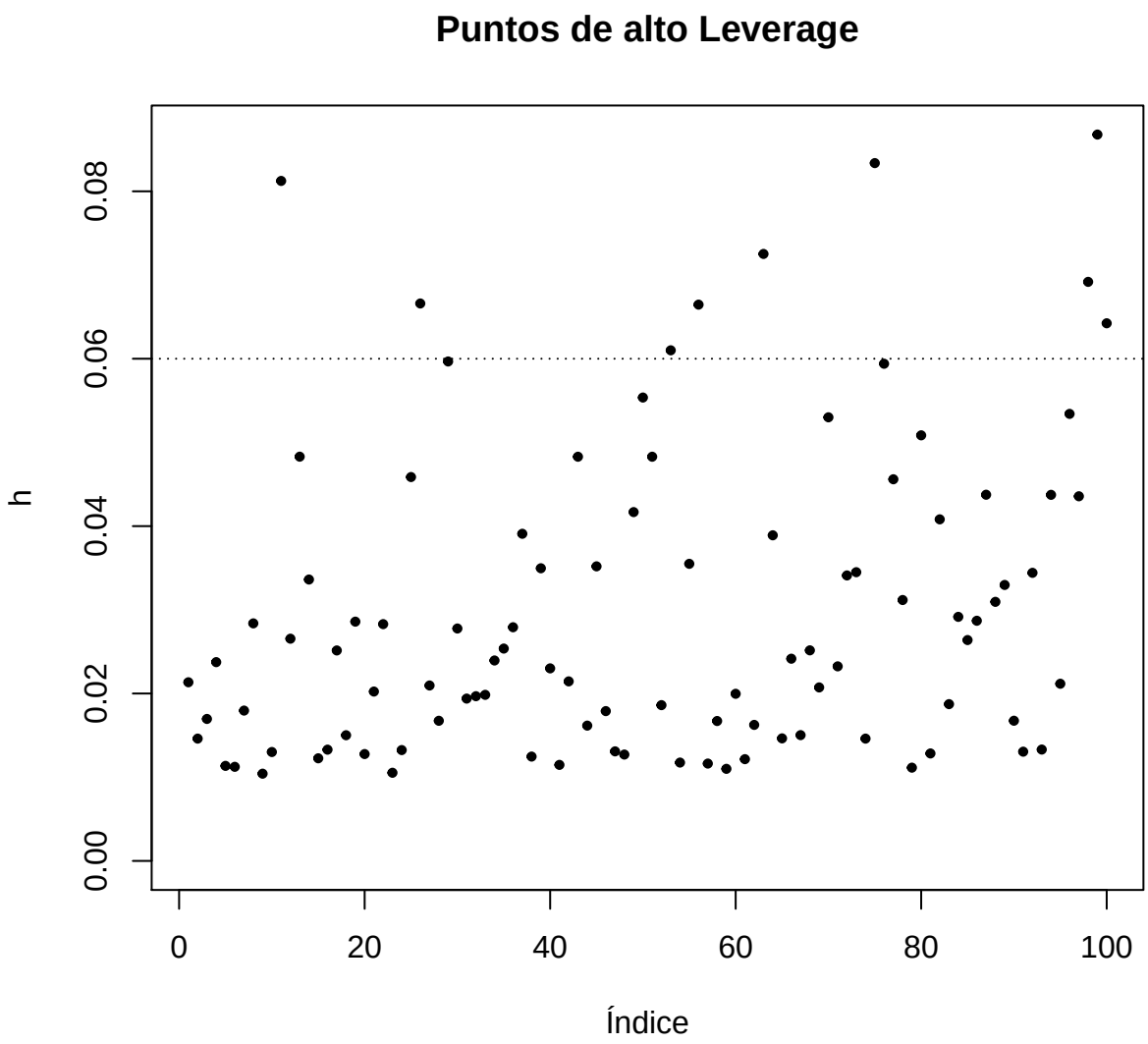
```
AIC(fit2)
```

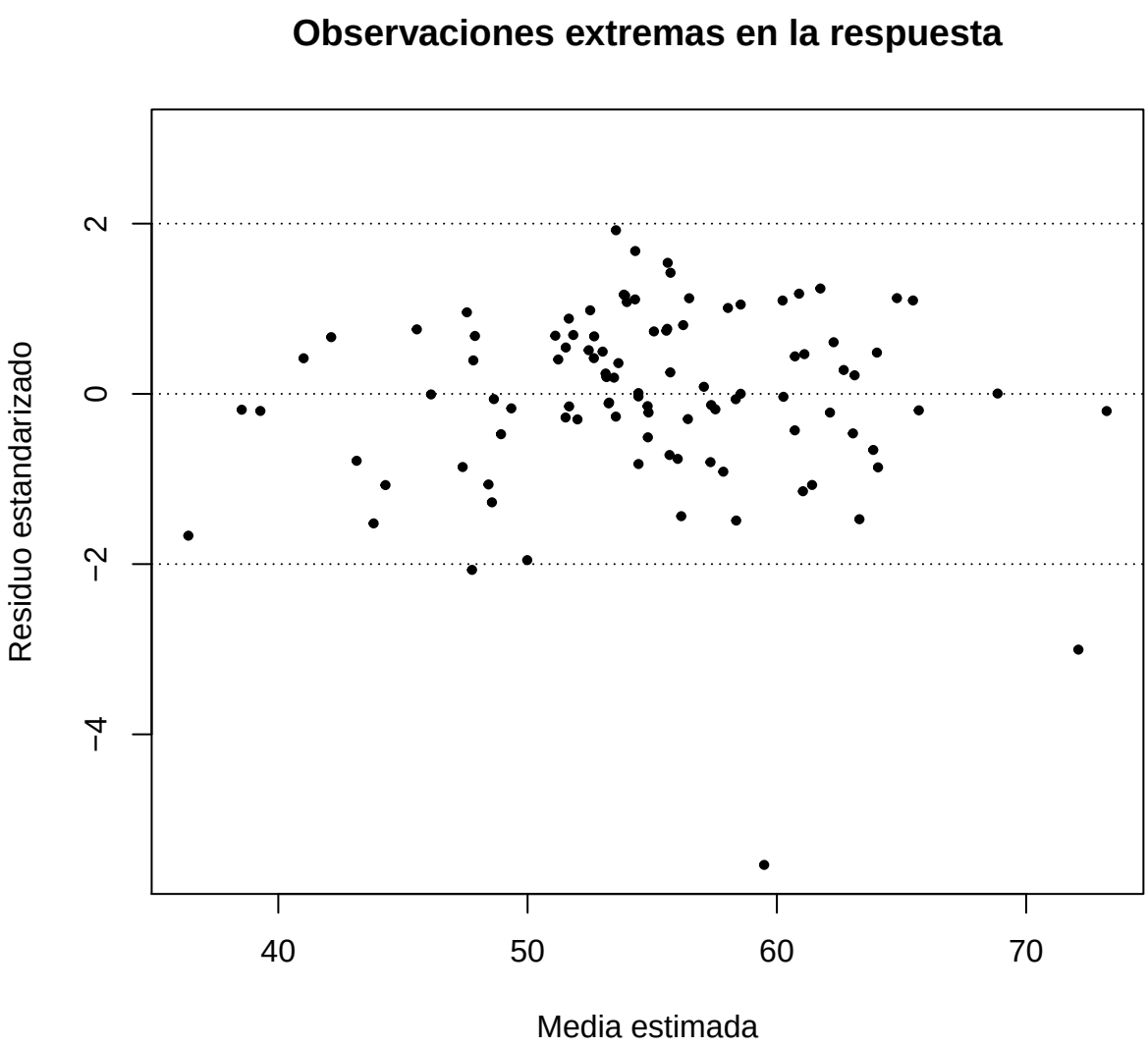
```
## [1] 543.7588
```

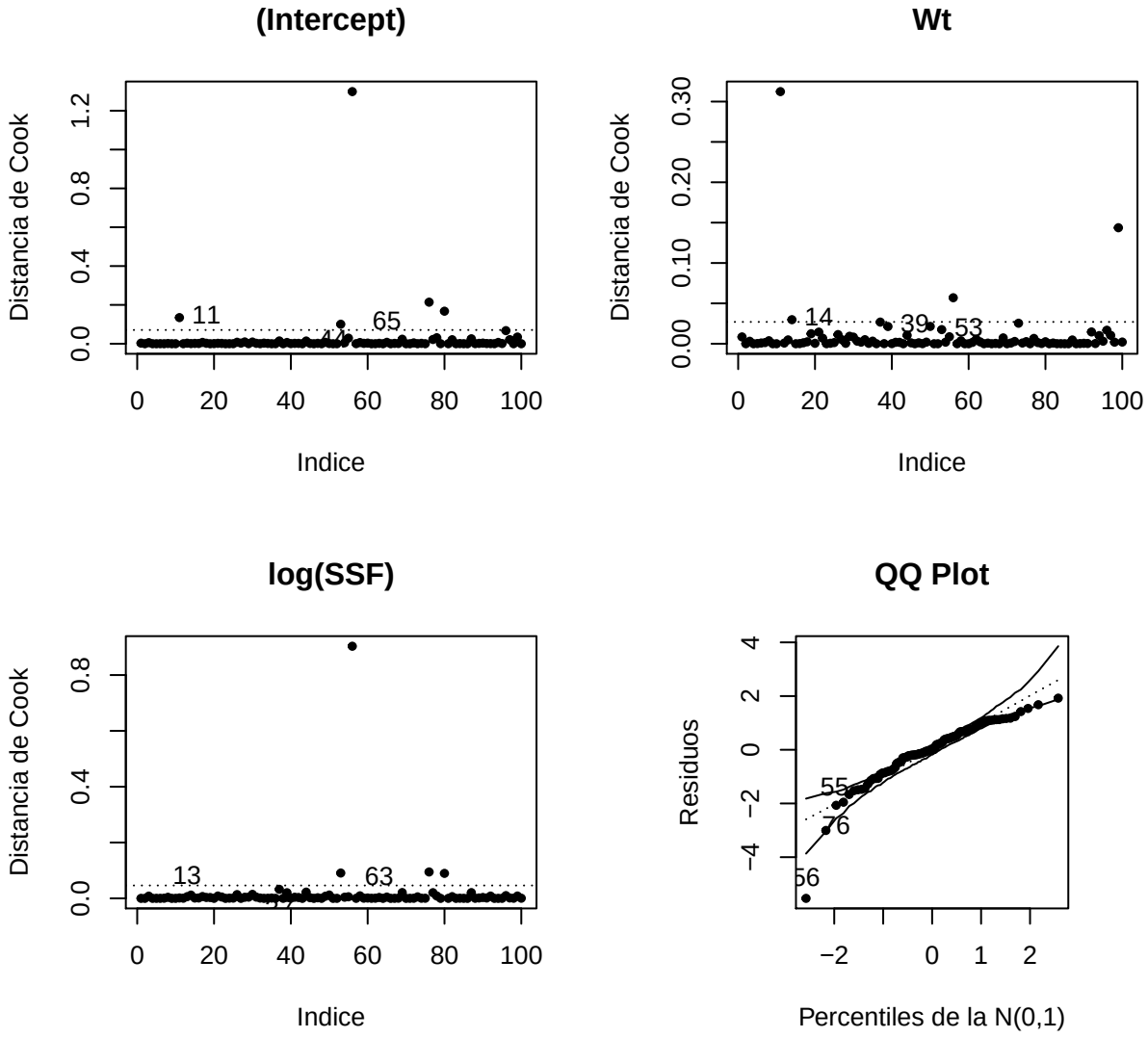
The R2 definition changes depends on the presence/ausence of the intercept, for that reason the R2 ajus there are not comparable between the two final that difers in intercept. To decide about the best model the AIC help us, we can see that the model I has shorter AIC than the model II, furthemore all of these parameter are significants, that is the reason to choose this model over the first one.

Residuals examination

It is necessary to examine the residuals distribution, points of leverage and influence.


Puntos de alto Leverage
h
0.00
0.02
0.04
0.06
0.08
0
20
40
60
80
100
Índice


Observaciones extremas en la respuesta
Residuo estandarizado
2
0
−2
−4
40
50
60
70
Media estimada




```
## [1] 55 56 76
```

Well, the points 56,71 and 76 are the outliers, it is necessary talking with the expert to decide if it is necessary removing this points of the sample.

Variance examination

```
######## Evaluando homogeneidad de la varianza ########
library(lmtest)
######## Test de Breusch-pagan ########
bptest(fit)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 5.1186, df = 2, p-value = 0.07736
```

The Breusch-pagan's test to examine the homogenicity in the variance is not rejected with a significant level of 5%

```
######## Evaluando NO correlaciÃ³n de los errores ########
######## Test de Durbin-Watson ########

dwtest(fit)
```

```
##
##  Durbin-Watson test
##
## data:  fit
## DW = 1.7363, p-value = 0.07647
## alternative hypothesis: true autocorrelation is greater than 0
```

The autocorrelation between residuals is rejected at significant level of 5%

Removing outliers

For this example, we are going to suppose that the best option is deleting the outliers 11 and 56 of out sample. Let see how the model behaviour witout this points.

```
atletas2<-atletas[-c(11,56),]
```

```
fit3<- lm(LBM ~ 1+Wt+log(SSF), data=atletas2)
summary(fit3)
```

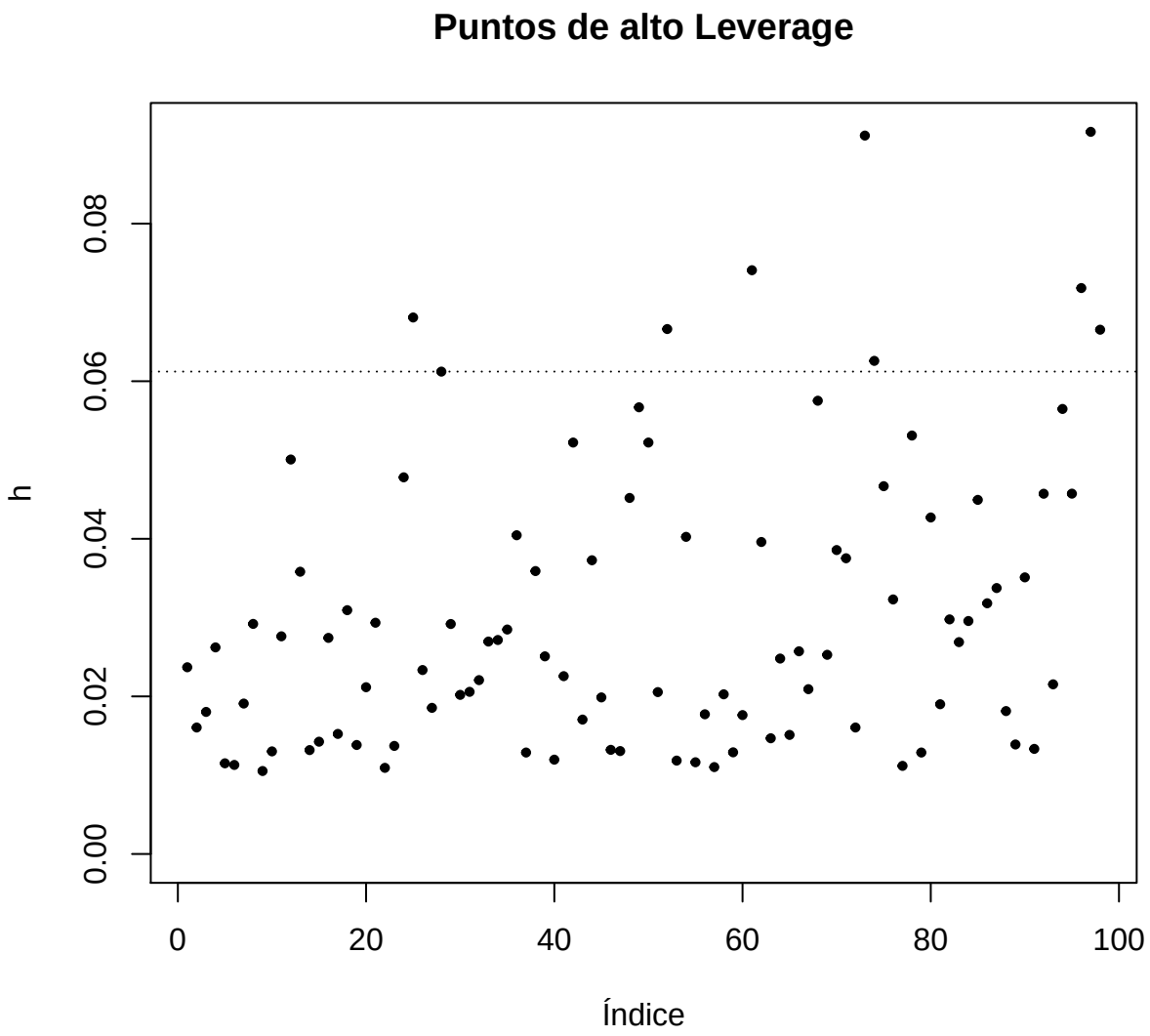
```
##
## Call:
## lm(formula = LBM ~ 1 + Wt + log(SSF), data = atletas2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34120 -0.74054  0.03627  0.82504  2.20410
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.68876    1.28526   28.55   <2e-16 ***
## Wt           0.83796    0.01499   55.90   <2e-16 ***
## log(SSF)    -8.68087    0.41566  -20.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 95 degrees of freedom
## Multiple R-squared:  0.9768, Adjusted R-squared:  0.9763
## F-statistic:  1996 on 2 and 95 DF,  p-value: < 2.2e-16
```

```
AIC(fit3)
```

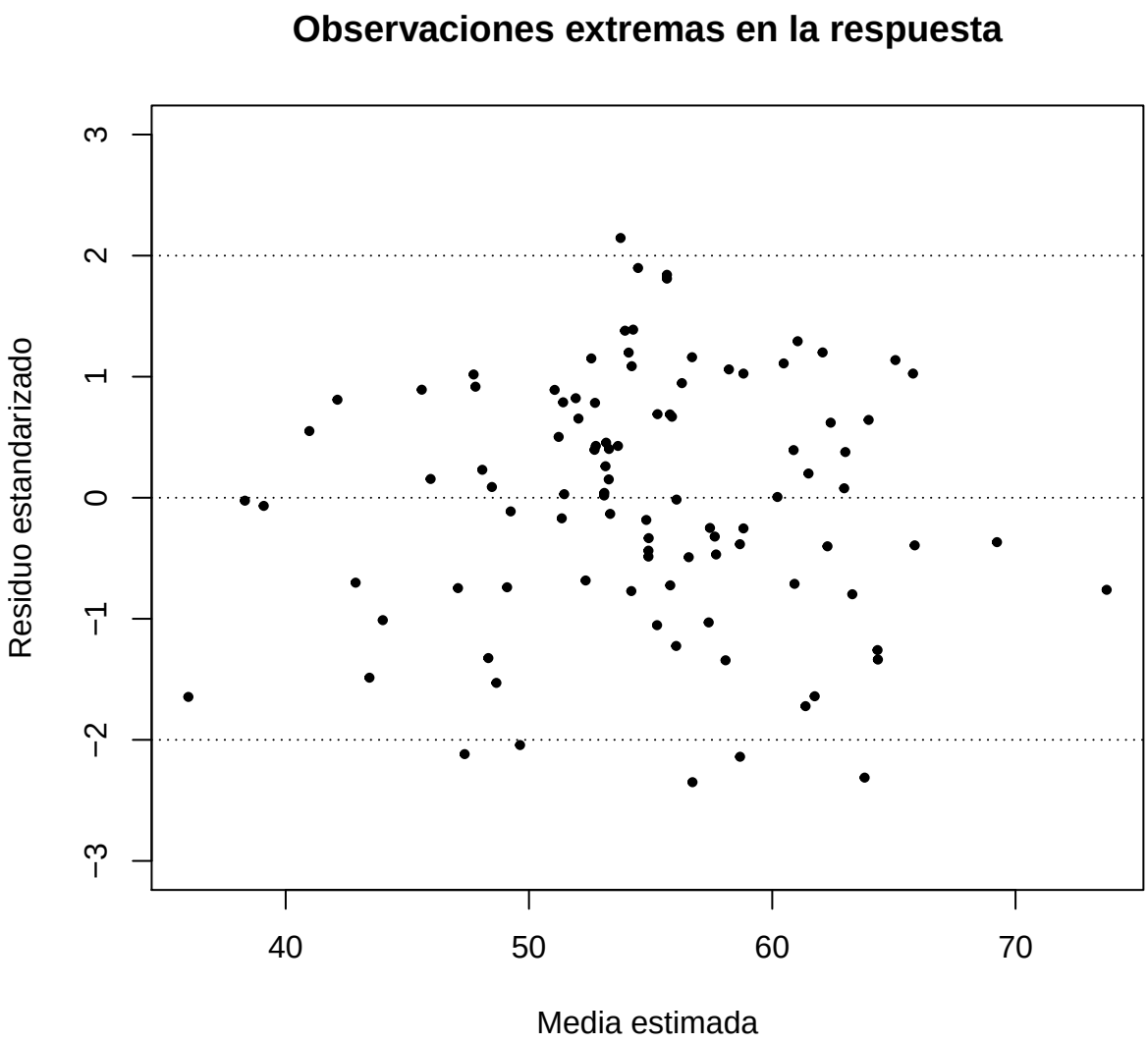
```
## [1] 293.6598
```

Well, the model seems good. Let's go to see the residuals examination

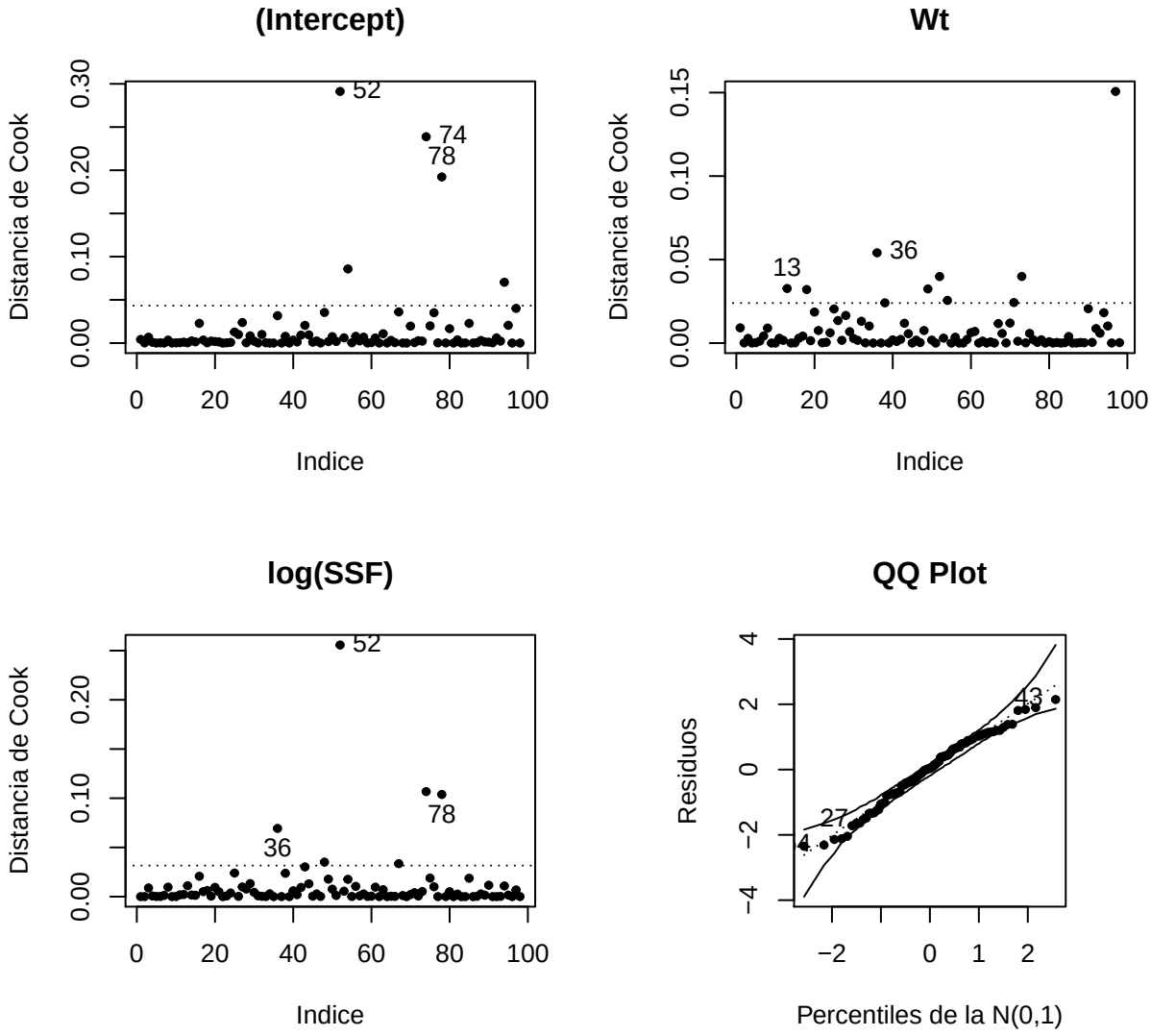
```
######## Leverage ####################################
Leverage <- Leverage.normal(fit3,3,"")
```




```
######## Residuos ######################################
residuos <- Residuos.normal(fit3,3,"")
```



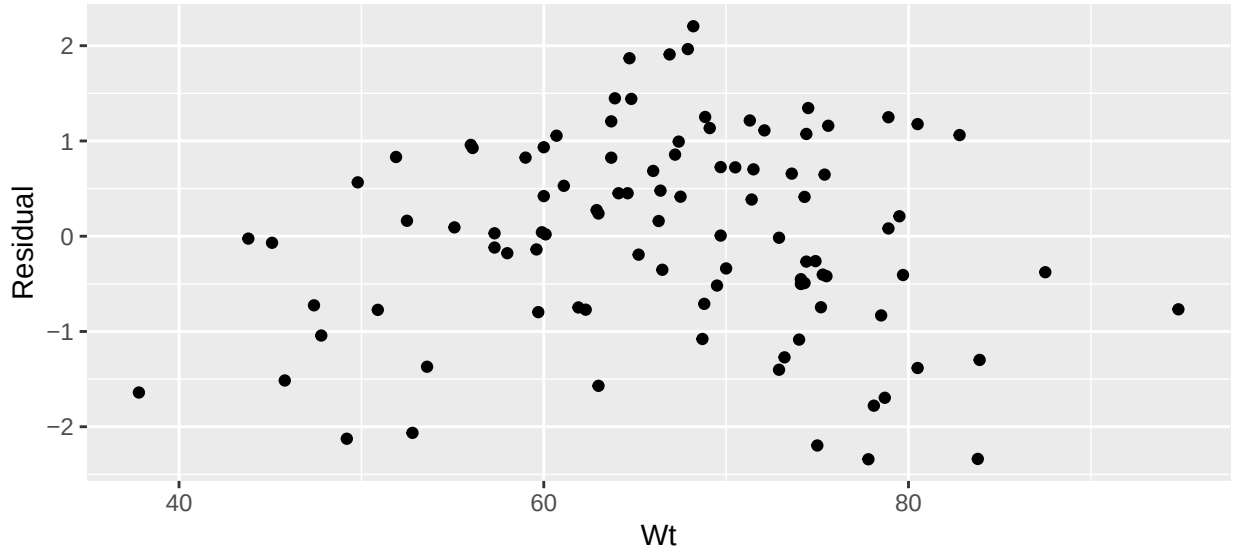

```
######## Influencia ######################################
influence <- Influence.normal(fit3,2,2,3,"")

######## QQ Plot y sus bandas de confianza ########
qqplot.normal(fit3,500,0.01,3,"")
```



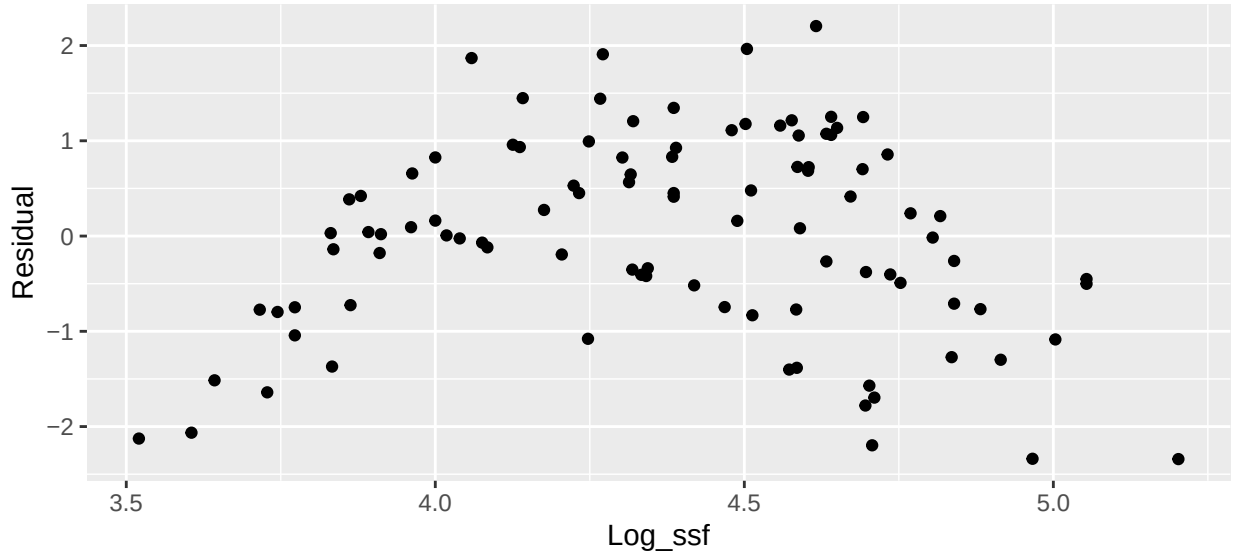

```
## [1] 27 43 54
```

We can see that there is any patter in the residuals plotted against the variables.

```
Wt<-atletas2$Wt
Log_ssf<-log(atletas2$SSF)
Residual<-fit3$residuals
df<-as.data.frame(cbind(Wt,Log_ssf, Residual))
ggplot(data=df,aes(Wt,Residual))+geom_point()
```




```
ggplot(data=df,aes(Log_ssf,Residual))+geom_point()
```




```
######## Evaluando homogeneidad de la varianza ########
library(lmtest)
######## Test de Breusch-pagan ########
bptest(fit3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit3
## BP = 1.0013, df = 2, p-value = 0.6061
```

```
######## Evaluando NO correlaciÃ³n de los errores ########
######## Test de Durbin-Watson ########

dwtest(fit3)
```

```
##
##  Durbin-Watson test
##
## data:  fit3
## DW = 1.9963, p-value = 0.4375
## alternative hypothesis: true autocorrelation is greater than 0
```

We can see that the model fits good the data and the residuals examination were good. Then our model is finished.

Moldel interpretation

The final model is: $LBM = B_0 + B_1 Wt + B_2 log(SSF)$, now, it is the time to interpretate its parameters. From the summary function we can identify just the coefficients.

```
summary(fit3)$coef
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 36.6887553  1.2852625  28.54573 2.118217e-48
## Wt           0.8379619  0.0149899  55.90176 1.725750e-74
## log(SSF)    -8.6808703  0.4156628 -20.88440 2.802111e-37
```

- B_0 for a person who has 0 in Wt and log(SSF) the LBM should be in average 36.6887
- B_1 for each adicional unit of the Wt the LBM should increase in average 0.8379 units
- B_2 for each adicional unit of log(SSF) the LBM should decrease in average -8.68087 units

References

- Peter K. Dunn and Gordon K. Smyth (2018). GLMsData: Generalized Linear Model Data Sets. R package version 1.0.0. https://CRAN.R-project.org/package=GLMsData
- Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr
- Hadley Wickham and Jennifer Bryan (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl
- Barret Schloerke, Jason Crowley, Di Cook, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg and Joseph Larmarange (2018). GGally: Extension to 'ggplot2'. R package version 1.4.0. https://CRAN.R-project.org/package=GGally
- Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0
- Achim Zeileis and Gabor Grothendieck (2005). zoo: S3 Infrastructure for Regular and Irregular Time Series. Journal of Statistical Software, 14(6), 1-27. doi:10.18637/jss.v014.i06
- Achim Zeileis, Torsten Hothorn (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL https://CRAN.R-project.org/doc/Rnews/