INTRODUCTION

Generalized Linear Models is an extension of simple and multiple regression models, as they allow other distributions for errors and a linkage function related to the mean of the response variable to the linear combination of explanatory variables.

The objective, therefore, is to analyze the influence that one or more (explanatory) variables, measured in individuals or objects, have on a variable of interest, which we call the response variable. Through the study of a regression model that relates this variable of interest with the so-called explanatory variables.

Thus allowing to define the behavior (distribution) of the response variable, the explanatory variables, the function that will link the explanatory variables to the response variable. With generalized linear models it is possible to model variables of interest that take the form of counts, symmetric and asymmetric continuous, binary and categorical.

In the case of this work, we are going to study the relationship of the dependent variable FAT (Y) with the other independent variables (X). In order to determine which variables have more correlation with the response variable and adjust a model that better explains certain behavior
We will analyze the association of body fat percentage in a sample of 252 men along with several other body measurements.

The objective is to make a linear model that allows obtaining the percentage of fat (outcome) using measurements of the body that are easy to obtain. The study variables will be ID, fat, weight, abdomen, age, adiposity, neck, hip, arm

IMPORTING DATABASE

```
rm(list = ls(all = TRUE))
setwd("~/R/tabela.xlsx")
tabela <- read_excel("tabela.xlsx")
```

CREATING VARIABLES

```
id<-c(1:252)
gordura<-c(0:45.1)
idade<-c(22:81)
peso<-c(118.5:363.15)
altura<-c(29.5:77.75)
adip<-c(18.1:48.9)
pesc<-c(31.1:51.2)
abdomem<-c(69.4:148.1)
quadril<-c(85:147.7)
braco<-c(24.8:45)
```

```
head(tabela)
```

```
ID gordura idade  peso altura  adip  pesc abdomem quadril braço
  <dbl>   <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl>    <dbl>   <dbl> <dbl>
1     1    12.6    23  154.   67.8  23.7  36.2    85.2    94.5  32
2     2     6.9    22  173.   72.2  23.4  38.5    83      98.7  30.5
3     3    24.6    22  154    66.2  24.7  34      87.9    99.2  28.8
4     4    10.9    26  185.   72.2  24.9  37.4    86.4   101.   32.4
5     5    27.8    24  184.   71.2  25.6  34.4   100     102.   32.2
6     6    20.6    24  210.   74.8  26.5  39      94.4   108.   35.7
```

Above we have a brief reading of the data set. The database has 252 observations, 9 variables being a FAT response variable and 8 explanatory variables.

LOADING PACKAGES

```
library(readxl)
library(ggplot2)
library(scales)
library(mtcars)
library(faraway)
```

DESCRIPTIVE ANALYSIS OF VARIABLES

```
> summary(tabela)
       ID            gordura          idade            peso           altura           adip            pesc      
 Min.   :  1.00   Min.   : 0.00   Min.   :22.00   Min.   :118.5   Min.   :29.50   Min.   :18.10   Min.   :31.10  
 1st Qu.: 63.75   1st Qu.:12.80   1st Qu.:35.75   1st Qu.:159.0   1st Qu.:68.25   1st Qu.:23.10   1st Qu.:36.40  
 Median :126.50   Median :19.00   Median :43.00   Median :176.5   Median :70.00   Median :25.05   Median :38.00  
 Mean   :126.50   Mean   :18.94   Mean   :44.88   Mean   :178.9   Mean   :70.15   Mean   :25.44   Mean   :37.99  
 3rd Qu.:189.25   3rd Qu.:24.60   3rd Qu.:54.00   3rd Qu.:197.0   3rd Qu.:72.25   3rd Qu.:27.32   3rd Qu.:39.42  
 Max.   :252.00   Max.   :45.10   Max.   :81.00   Max.   :363.1   Max.   :77.75   Max.   :48.90   Max.   :51.20  
    abdomem          quadril          braço      
 Min.   : 69.40   Min.   : 85.0   Min.   :24.80  
 1st Qu.: 84.58   1st Qu.: 95.5   1st Qu.:30.20  
 Median : 90.95   Median : 99.3   Median :32.05  
 Mean   : 92.56   Mean   : 99.9   Mean   :32.27  
 3rd Qu.: 99.33   3rd Qu.:103.5   3rd Qu.:34.33  
 Max.   :148.10   Max.   :147.7   Max.   :45.00  
```

Some variables presented very close mean and median data, which indicates a low data dispersion. Below is an analysis of the standard deviation and variance of the variables.

Standard deviation:

```
> sd(gordura)
[1] 7.750856
> sd(idade)
[1] 12.60204
> sd(peso)
[1] 29.38916
> sd(altura)
[1] 3.662856
> sd(adip)
[1] 3.648111
> sd(pesc)
[1] 2.430913
> sd(abdomem)
[1] 10.78308
> sd(quadril)
[1] 7.164058
> sd(braço)
[1] 3.021274
```

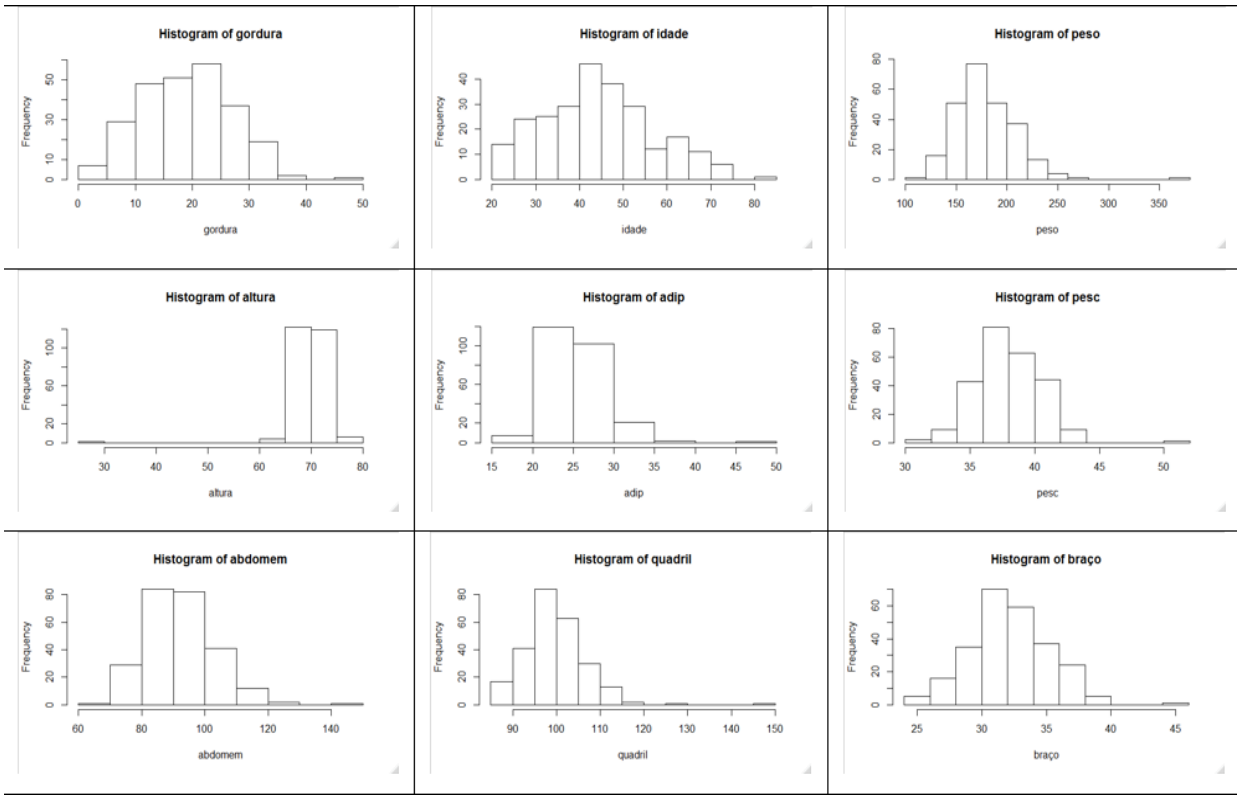
The standard deviation indicates what the “error” is if we wanted to replace one of the collected values with the mean value.

```
> var(tabela)
          gordura      idade       peso     altura       adip       pesc    abdomem    quadril      braço
gordura  60.075763  28.245483 139.671527 -2.5297548 20.5847491  9.260466  68.007997  34.743601 11.545529
idade    28.245483 158.811405  -4.720686 -7.9230459  5.4640249  3.477171  31.310050  -4.544071 -1.567215
peso    139.671527  -4.720686 863.722719 33.1856467 95.1373826 59.348441 281.410541 198.099047 71.071090
altura   -2.529755  -7.923046  33.185647 13.4165125 -0.3326053  2.259054   3.468334   4.471301  2.299789
adip     20.584749   5.464025  95.137383 -0.3326053 13.3087123  6.898222  36.343465  23.084485  8.226603
pesc      9.260466   3.477171  59.348441  2.2590543  6.8982223  5.909339  19.766422  12.799440  5.369868
abdomem  68.007997  31.310050 281.410541  3.4683338 36.3434647 19.766422 116.274745  67.522123 22.315796
quadril  34.743601  -4.544071 198.099047  4.4713005 23.0844849 12.799440  67.522123  51.323722 16.001243
braço    11.545529  -1.567215  71.071090  2.2997889  8.2266026  5.369868  22.315796  16.001243  9.128095
```

The greater the variance, the farther the values are from the mean, and the smaller the variance, the closer the values are to the mean.

Below we have the histograms of the variables.

```
hist(gordura)
hist(idade)
hist(peso)
hist(altura)
hist(adip)
hist(pesc)
hist(quadril)
hist(braço)
```



We can notice through the histograms above the dispersion of the data. The variables showed the highest concentration of data:

Fat: between 10% to 30%

Age: between 40 and 50 years.

Weight: between 150 and 200 pounds.

Height: 65 and 75 units [1]

Adiposity: 20 to 30 units [2]

Neck: 35 to 45 centimeters

Abdomen: 80 and 100 centimeters

Hip: 95 and 105 centimeters

Arm: 30 to 35 centimeters

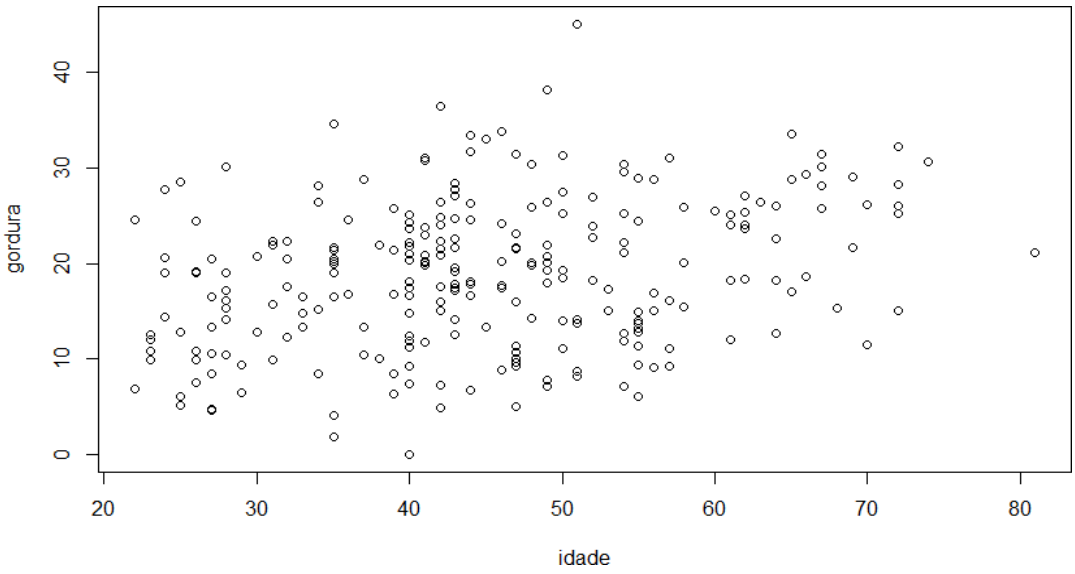
```
CORRELAÇÃO DAS VARIÁVEIS
```

```
> cor(tabela)
                ID     gordura       idade        peso      altura        adip       pesc    abdomem     quadril
ID       1.00000000  0.11095086  0.34125350  0.03372794  0.04094313  0.04771746 0.07111233 0.12171973 -0.02373697
gordura  0.11095086  1.00000000  0.28917352  0.61315611 -0.08910641  0.72799418 0.49148893 0.81370622  0.62569993
idade    0.34125350  0.28917352  1.00000000 -0.01274609 -0.17164514  0.11885126 0.11350519 0.23040942 -0.05033212
peso     0.03372794  0.61315611 -0.01274609  1.00000000  0.30827854  0.88735216 0.83071622 0.88799494  0.94088412
altura   0.04094313 -0.08910641 -0.17164514  0.30827854  1.00000000 -0.02489094 0.25370988 0.08781291  0.17039426
adip     0.04771746  0.72799418  0.11885126  0.88735216 -0.02489094  1.00000000 0.77785691 0.92388010  0.88326922
pesc     0.07111233  0.49148893  0.11350519  0.83071622  0.25370988  0.77785691 1.00000000 0.75407737  0.73495788
abdomem  0.12171973  0.81370622  0.23040942  0.88799494  0.08781291  0.92388010 0.75407737 1.00000000  0.87406618
quadril -0.02373697  0.62569993 -0.05033212  0.94088412  0.17039426  0.88326922 0.73495788 0.87406618  1.00000000
braço   -0.01567689  0.49303089 -0.04116212  0.80041593  0.20781557  0.74638418 0.73114592 0.68498272  0.73927252
              braço
ID       -0.01567689
gordura   0.49303089
idade    -0.04116212
peso      0.80041593
altura    0.20781557
adip      0.74638418
pesc      0.73114592
abdomem   0.68498272
quadril   0.73927252
braço     1.00000000
```

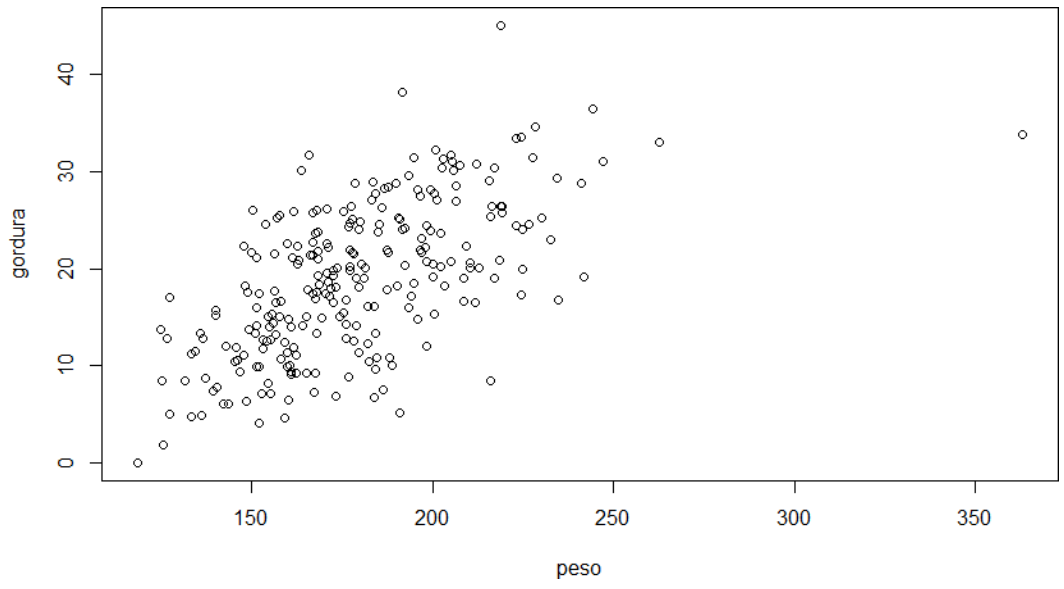
After making the correlation of the variables, we can notice that the fat variable has a strong correlation with the variables weight, adiposity, neck, abdomen, hips and arms. We verified the occurrence of multicollinearity, which can influence the model result. Below are the correlation graphs of the variables under study with the response variable FAT.

```
plot(idade, gordura)
plot(peso, gordura)
plot(altura, gordura)
plot(adip, gordura)
plot(pesc, gordura)
plot(abdomem, gordura)
plot(quadril, gordura)
plot(braço, gordura)
```

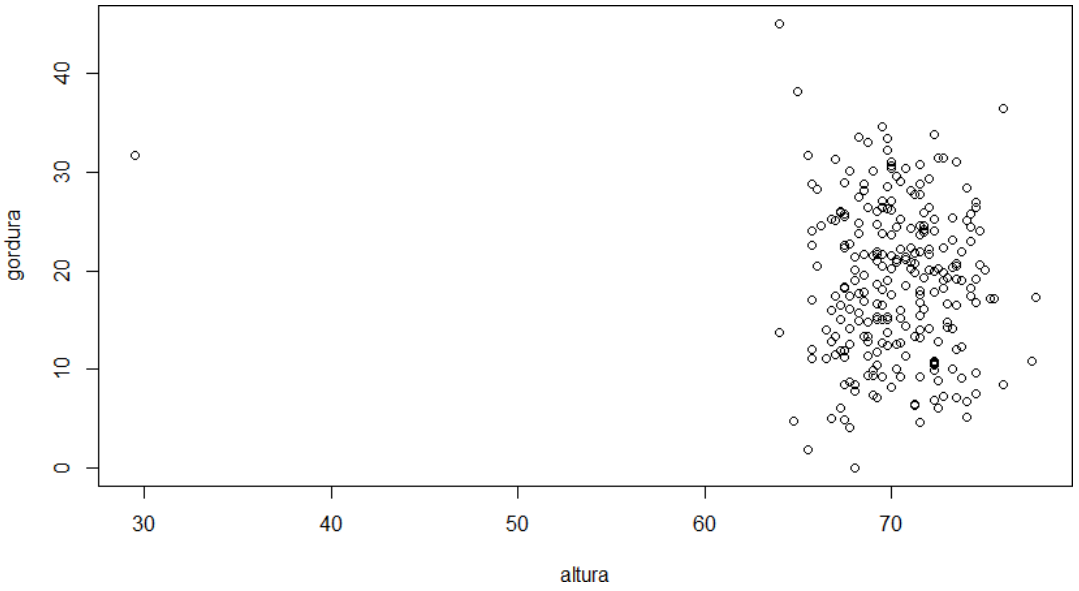
IDADE X GORDURA



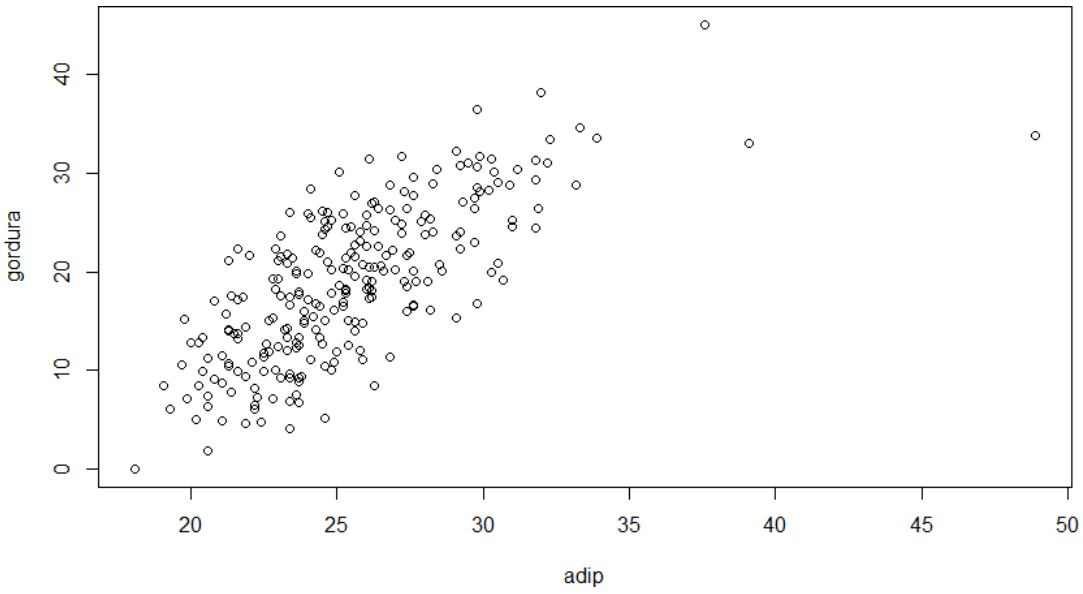
PESO X GORDURA



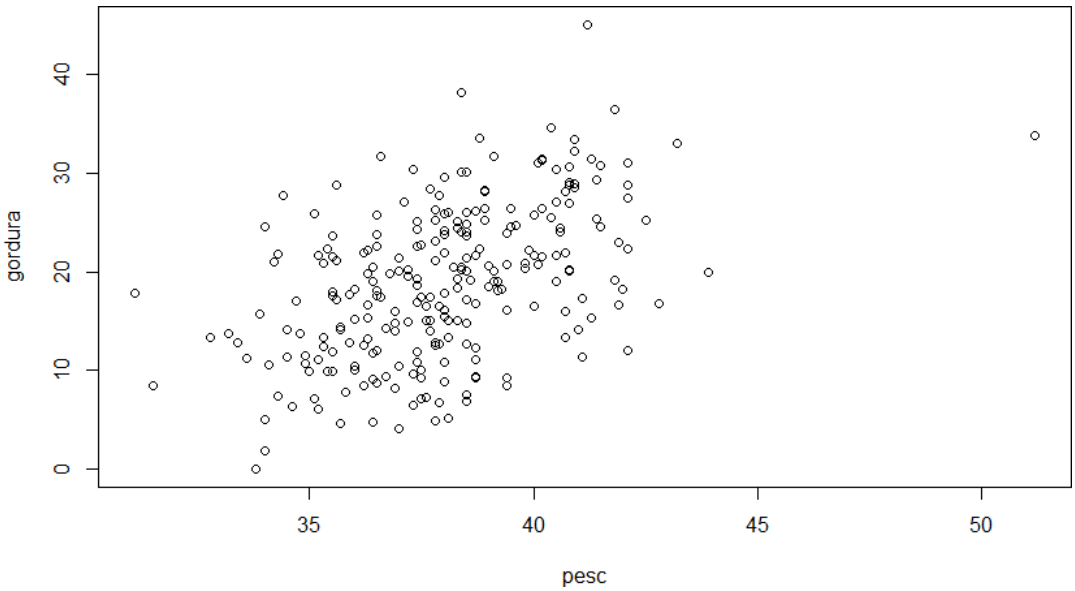
ALTURA X GORDURA



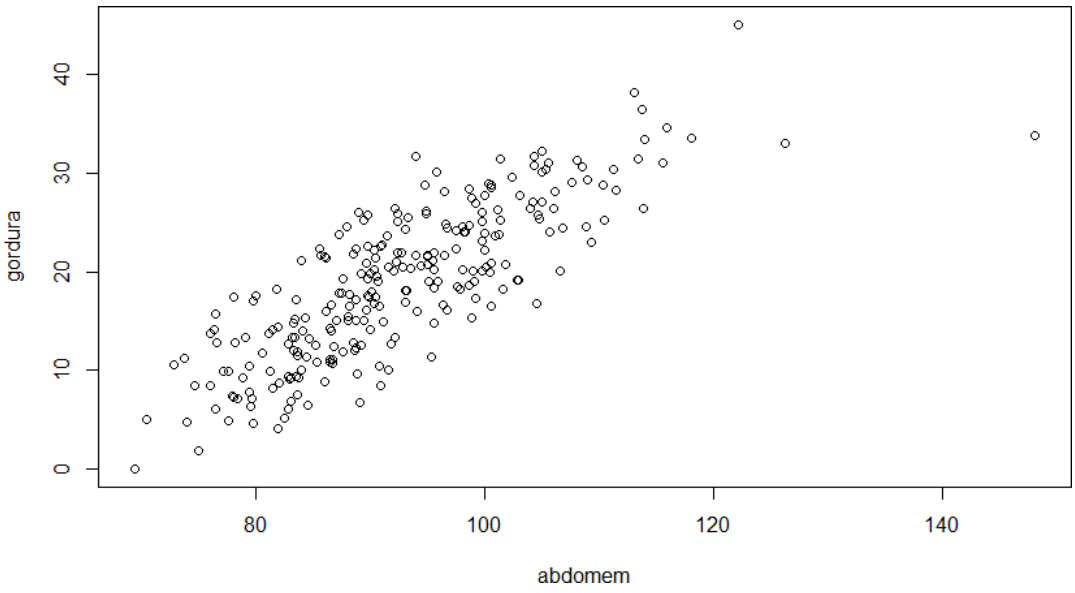
ADIP X GORDURA



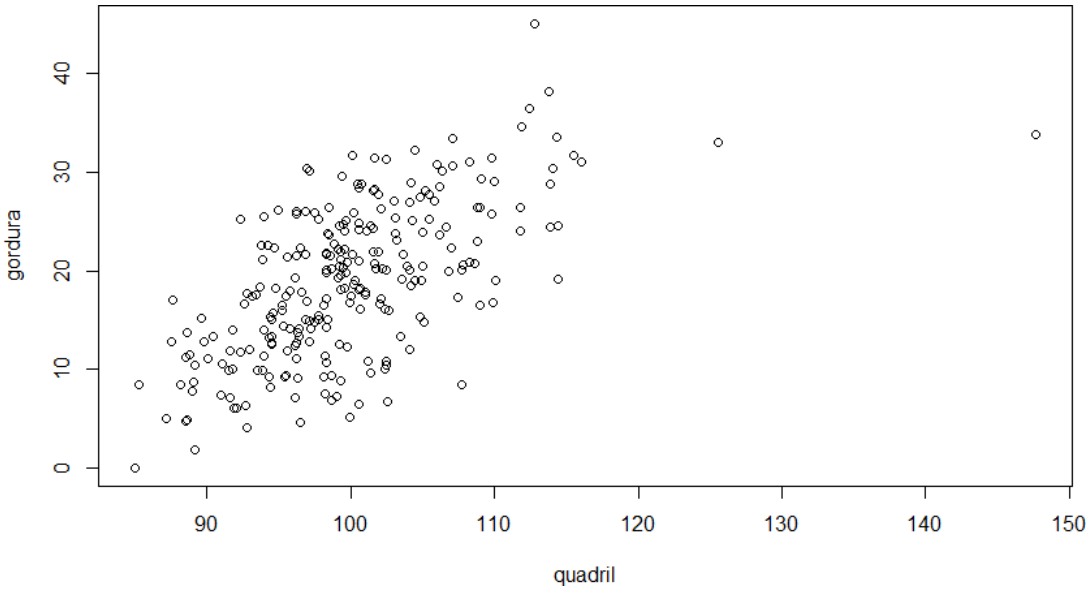
PESC X GORDURA



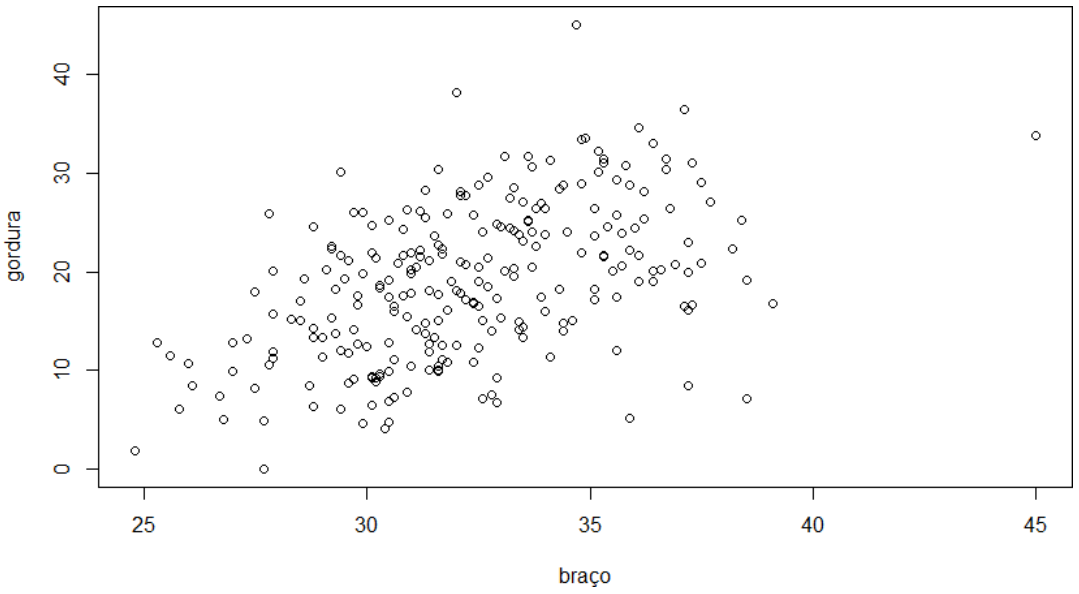
ABDOMEM X GORDURA



QUADRIL X GORDURA



BRAÇO X GORDURA



Through the analysis of the graphs, we can see that there is a significant relationship between fat and the variables Weight, Adiposity and Abdomen. The variables Age and Height showed no evidence of relationship with the response variable Fat.

CREATING THE DATA FRAME

```
tabela2 = data.frame(tabela ,produto)
attach(tabela2)
```

SCALING THE VARIABLES

```
tabela= scale(tabela)
tabela2 = scale(tabela2)
```

IDENTIFYING OCCURRENCE OF MULTICOLINEARITY

```
vif(tabela)
```

	Idade	Peso	Altura	Adip.	Circ. Pescoço	Circ. Abdômen	Circ. Quadril	Circ. Braço
VIF	1,57	***24,36***	2,21	***13,43***	4,02	***17,12***	***11,94***	3,15

Through the analysis of VIF (Variance Inflation Factors) we verified multicollinearities in some variables presented VIF greater than 10 which can cause problems in the estimation of the coefficients. Let's recalculate the VIF only with the variables that presented the highest correlation with the fat response variable, which in this case were weight, adiposity, abdomen and hip. We will create a set of values with just these variables and recalculate the VIF.

VARIABLE CREATED TO SOLVE THE MULTICOLINEARITY PROBLEM

```
produto = (peso~adip~abdomem~quadril)
```

CALCULATING VIF AGAIN:

	Idade	Altura	Circ. Pescoço	Circ. Braço	Produto
VIF	1,21	1,18	3,05	2,54	2,64

We have verified that the multicollinearity problem has been solved, we can now adjust the model to solve the problem.

3 - MODEL ADJUSTMENT

MEASURING THE SIGNIFICANCE OF VARIABLES IN THE PROPOSED MODEL.

```
summary(mod0 <-lm(gordura~ ., data=tabela))

Call:
lm(formula = gordura ~ ., data = tabela)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0503  -3.0669   0.1313   2.9528  10.3543

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.903673  13.184532  -1.358  0.17575
ID           -0.002335   0.003831  -0.610  0.54274
idade         0.004406   0.026339   0.167  0.86730
peso         -0.091798   0.042969  -2.136  0.03365 *
altura       -0.102157   0.104498  -0.978  0.32925
adip          0.062689   0.259059   0.242  0.80899
pesc         -0.548176   0.209627  -2.615  0.00948 **
abdomem       0.908096   0.080467  11.285  < 2e-16 ***
quadril      -0.137004   0.125104  -1.095  0.27455
braço         0.291270   0.150415   1.936  0.05398 .
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.085 on 242 degrees of freedom
Multiple R-squared:  0.7322,  Adjusted R-squared:  0.7223
F-statistic: 73.53 on 9 and 242 DF,  p-value: < 2.2e-16
```

MODEL 1

To create a multiple regression model, we would multiply the fat response variable plus all other independent variables as shown in the model below.

```
modgeral = lm(gordura ~ idade + peso + altura + adip + pesc + abdomem
+ quadril + braço)

summary(modgeral)

step(modgeral)
```

GORDURA = -17.90 + 0.002335 ID - 0.004406 IDADE - 0.091798 **PESO** - 0.102157 ALTURA - 0.062689 ADIP - 0.548176 PESC - 0.908096 ABDOMEM - 0.137004 QUADRIL - 0.291270 **BRAÇO**.

However, this first model is not an adequate model for the study because we verified that some variables showed little or no significance for the explanation of the dependent variable. Some variables presented values very close to 0, that is, variables with little significance to explain the dependent variable.

There is no reason to put these variables with values very close to 0 in the model. The T n Test shows us that we should accept the H0 hypothesis because the betas are very close to 0 . After removing the insignificant variables from the models, we will again estimate the model with the most significant variables.

MODEL 2

```
modproduto = lm(gordura ~ idade + altura + pesc + braço + produto)
summary(modproduto)
step(modproduto)
```

Using Stepwize to choose the variables that should compose the final model. The final model selected resulted in the following equation:

Gordura = 10.498 + Idade 0.156 - Altura 0.297 - Braço 0.689 - Produto 3.066

FINAL MODEL

```
modfinal = lm(gordura ~ idade + altura + braço + produto)
```

4- INFERENCE OF MODEL PARAMETERS

The Stepwise method for selecting variables is widely used in linear regression. Any procedure for selecting or excluding variables from a model is based on an algorithm that checks the importance of variables, including or excluding them from the model based on a decision rule. The importance of the variable is defined in terms of a measure of statistical significance of the coefficient associated with the variable for the model. This statistic depends on the model's assumptions.

A confidence interval of (95%) was used to test the regression estimates, the result shows us that the model suggested by stepwize is correct because it eliminated the Pesc variable because it contained a confidence interval equal to 0 and cannot be considered values 0 for this reason we will not use the Pesc variable in the final model.

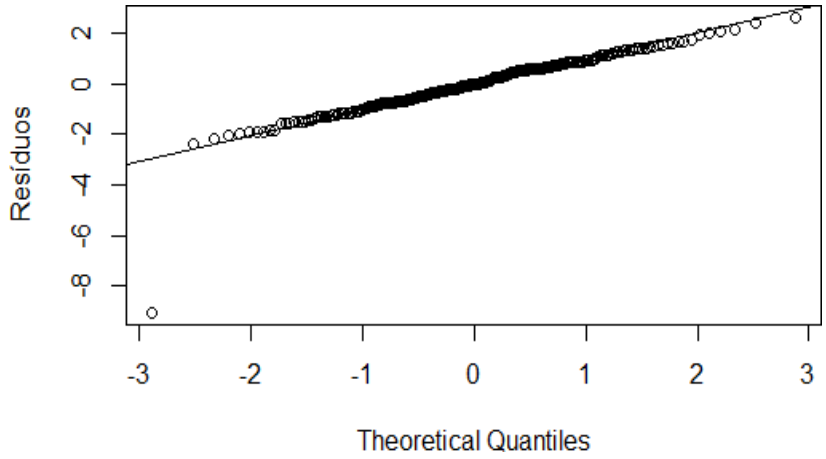
```
confint(modproduto)
confint(modfinal)
```

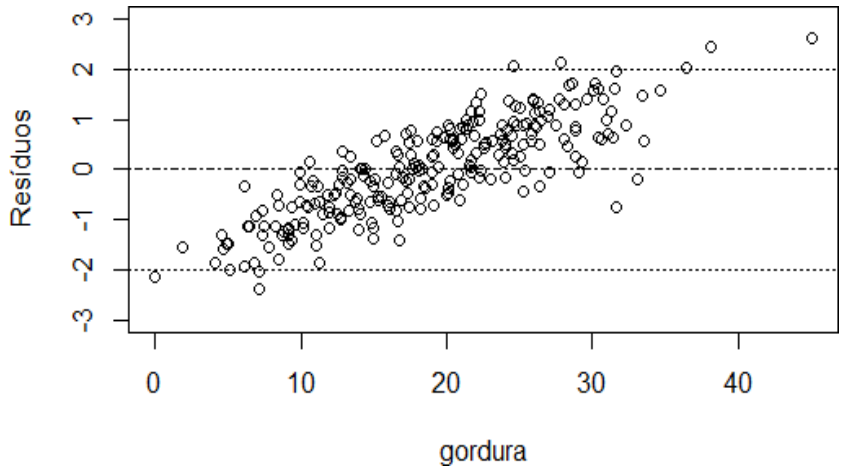
Variável	ID modelo 1		IC modelo final	
Intercepto	-11.617	27.775	-6.490	27.486
Idade	0.093	0.213	0.098	0.215
Altura	-0.521	-0.098	-0.502	-0.092
Pescoço	-0.393	0.647	---	
Braço	0.273	1.019	0.361	1.018
Produto	1.816	4.052	2.088	4.043

5 - MODEL DIAGNOSIS

We will use residual plots to analyze the regression.

```
qqnorm(rstudent(modfinal), ylab="Resíduos", main="")
qqline(rstudent(modfinal))
plot(gordura,rstudent(modfinal),ylab="Resíduos", main="",ylim=c(-3,3))
abline(h=0, lty=4)
abline(h=-2,lty=3)
abline(h=2,lty=3)
```





Through the analysis of the residuals we can see that the residuals remain within the estimated confidence interval of 95%. Then the model is valid.

5 - FINAL CONSIDERATIONS

The database showed multicollinearity between the variables, being necessary for the adjustment of the model to make a set of values in it containing the variables weight, adiposity, abdomen and hip. The variable created product together with the other variables explained the indicated model.

After adjusting the variables, the model indicated by stepwise

the ideal model would be composed of the variables Age, Height, Arm Circumference and Product.

It was possible to evaluate the validity of the proposed model because the residual plots used for the tests showed that the data were concentrated within the confidence interval used in the work, which was 95%. In other words, the adjusted model is indeed capable of explaining the response variable of the study in question.