PROGRAMA DE PÓS-GRADUAÇÃO

ESPECIALIZAÇÃO EM ESTATÍSTICA APLICADA

MODELOS LINEARES

Salvador

2019

ANA PAULA DE SOUZA VANDERLEY

MODELOS LINEARES

Trabalho apresentado ao Curso de Especialização em Estatística Aplicada do Centro Universitário Jorge Amado, como requisito para a conclusão da disciplina Modelos lineares, sob a orientação do Profº Jackson Conceição..

.

Salvador

2019

SUMÁRIO

1 INTRODUÇÃO

2 ANÁLISE DESCRITIVA

3 AJUSTE DO MODELO

4 INFERÊNCIA DOS PARÂMETROS DO MODELO

5 DIAGNÓSTICO DO MODELO PROPOSTO

6 CONSIDERAÇÕES FINAIS

1- INTRODUÇÃO

Modelos Lineares Generalizados é uma extensão dos modelos de regressão simples e múltipla, pois possibilitam outras distribuições parra os erros e uma função de ligação relacionada a média da variável resposta à combinação linear das variáveis explicativas.

O Objetivo portanto é analisar a influência que uma ou mais variáveis (explicativas), medidas em indivíduos ou objetos, tem sobre uma variável de interesse a que damos o nome de variável resposta. Através do estudo de um modelo de regressão que relacione essa variável de interesse com as variáveis ditas explicativas.

Permitindo assim definir o comportamento (distribuição) da variável resposta, as variáveis explicativas, a função que irá ligar as variáveis explicativas a variável resposta. Com os modelos lineares generalizados é possível modelar variáveis de interesse que assumem a forma de contagem, contínuas simétricas e assimétricas, binárias e categóricas.

No caso desse Trabalho vamos estudar a relação da variável dependente GORDURA (Y) com as demais variáveis independentes (X) . Afim de determinar quais variáveis tem mais correlação com a variável resposta e ajustar um modelo que explique melhor determinado comportamento

Analisaremos a associação do percentual de gordura corporal em uma amostra de 252 homens junto com diversas outras medidas corporais. O objetivo é fazer um modelo linear que permita obter o percentual de gordura (desfecho) usando medidas do corpo fáceis de serem obtidas. As variáveis do estudo serão ID , gordura, peso, abdomem ,idade, adiposidade, pescoço, quadril, braço.

TODOS OS COMANDOS DO R ESTÃO EM AZUL.

IMPORTANDO BANCO DE DADOS

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

CRIANDO VARIAVEIS

```
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
	<db1>	*<db1>*	*<db1>*	*<db1>*	*<db1>*	*<db1>*	*<db1>*	*<db1>*	*<db1>*
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

Acima temos uma breve leitura do conjunto de dados.O banco de dados possui 252 observações, 9 variáveis sendo uma variável resposta GORDURA e 8 vari áveis explicativas.

CARREGANDO PACOTES

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

2 - ANÁLISE DESCRITIVA DAS VARIÁVEIS

```
> 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  
```

Algumas variáveis apresentaram dados de média e mediana muito próximos, o que indica uma baixa dispersão dos dados. Abaixo uma análise do desvio padrão e variância das variáveis.

Desvio Padrão:

```
> 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
```

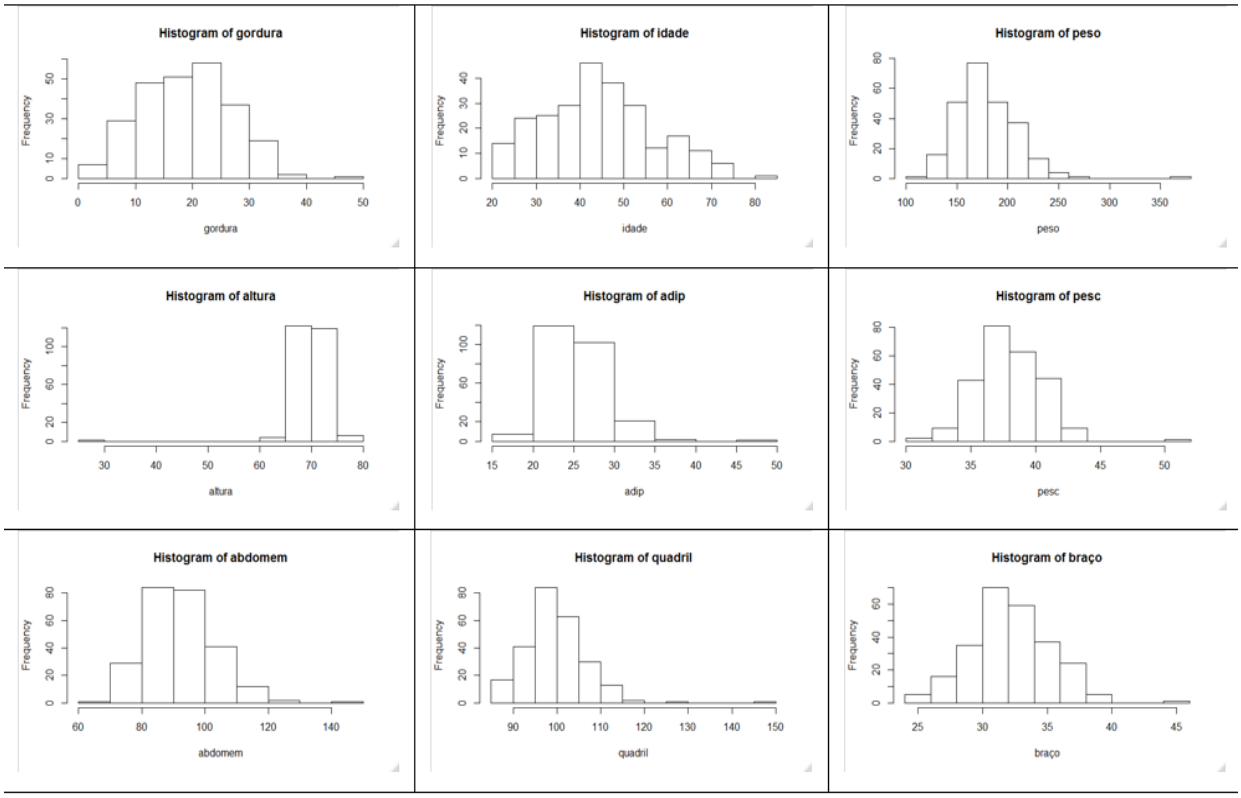
o desvio padrão indica qual é o “erro” se quiséssemos substituir um dos valores coletados pelo valor da média.

```
> 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
```

Quanto maior for a variância, mais distantes da média estarão os valores, e qu anto menor for a variância, mais próximos os valores estarão da média.

A seguir temos os histogramas das variáveis.

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



Podemos notar através dos histogramas acima a dispersão dos dados. As variáveis apresentaram maior concentração de dados :

Gordura: entre 10% a 30%

Idade: entre 40 e 50 anos.

Peso: entre 150 e 200 libras.

Altura: 65 e 75 unidades [1]

Adiposidade: 20 a 30 unidades [2]

Pescoço: 35 a 45 centímetros

Abdômen: 80 e 100 centímetros

Quadril: 95 e 105 centímetros

Braço: 30 a 35 centímetros

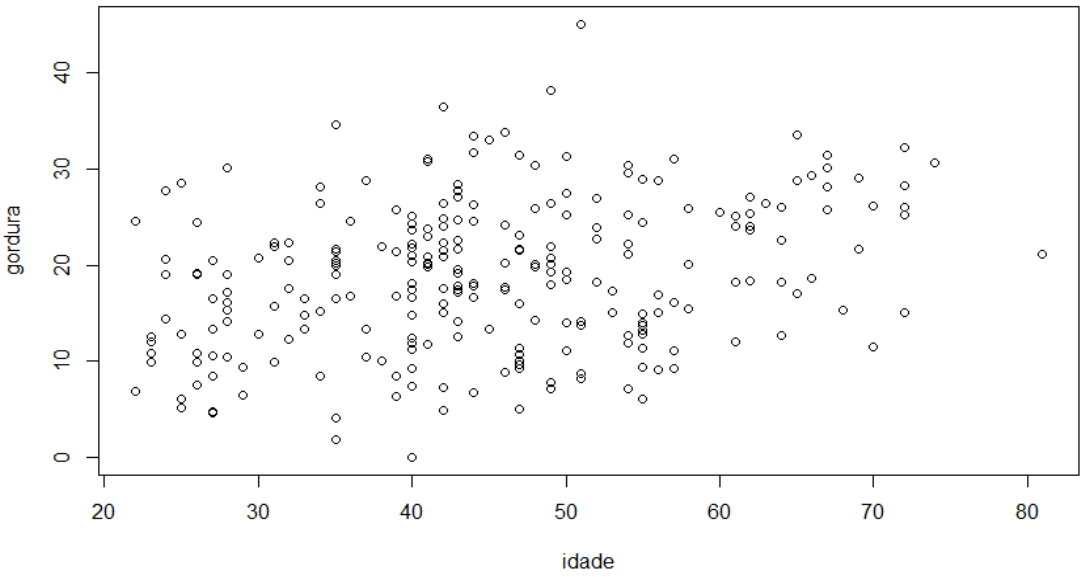
CORRELAÇÃO DAS VARIÁVEIS

```
> 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
```

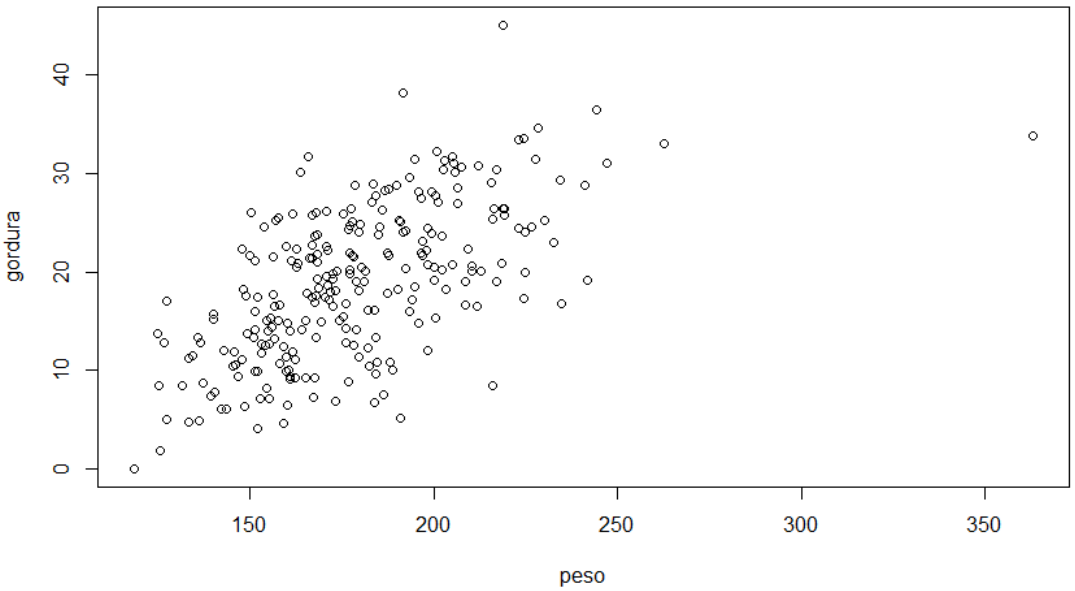
Depois de feita a correlação das variáveis podemos notar que a variável **gordura** tem uma forte correlação com as variáveis **peso, adip , pescoço, abdomem , quadril e braço.** Verificamos ocorrência de multicolinearidade, que pode influenciar o resultado do modelo. A Seguir os Gráficos de correlação das variáveis em estudo com a variável resposta GORDURA.

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

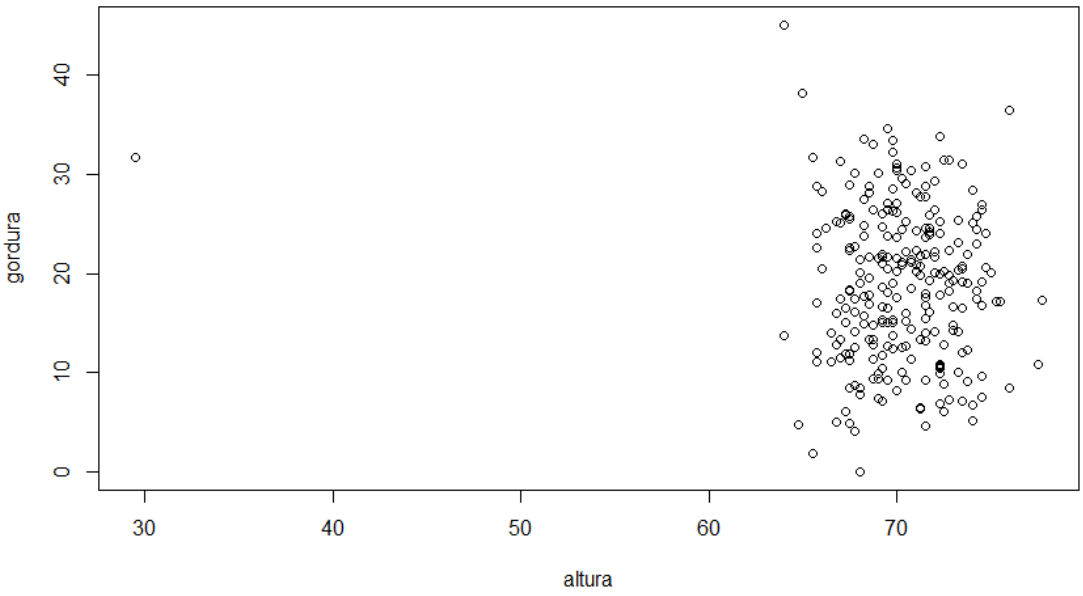
IDADE X GORDURA



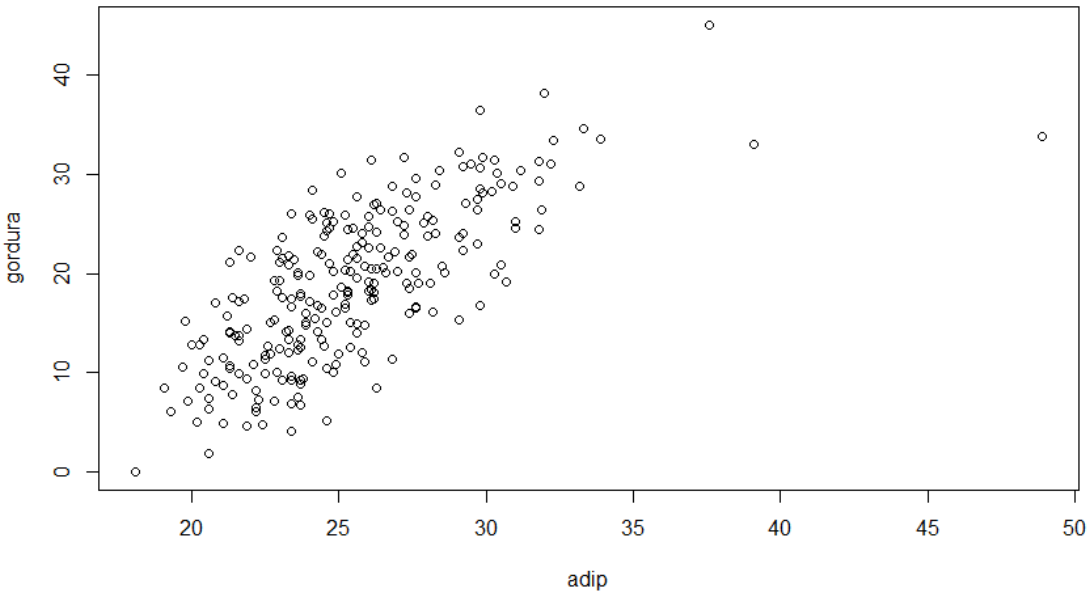
PESO X GORDURA



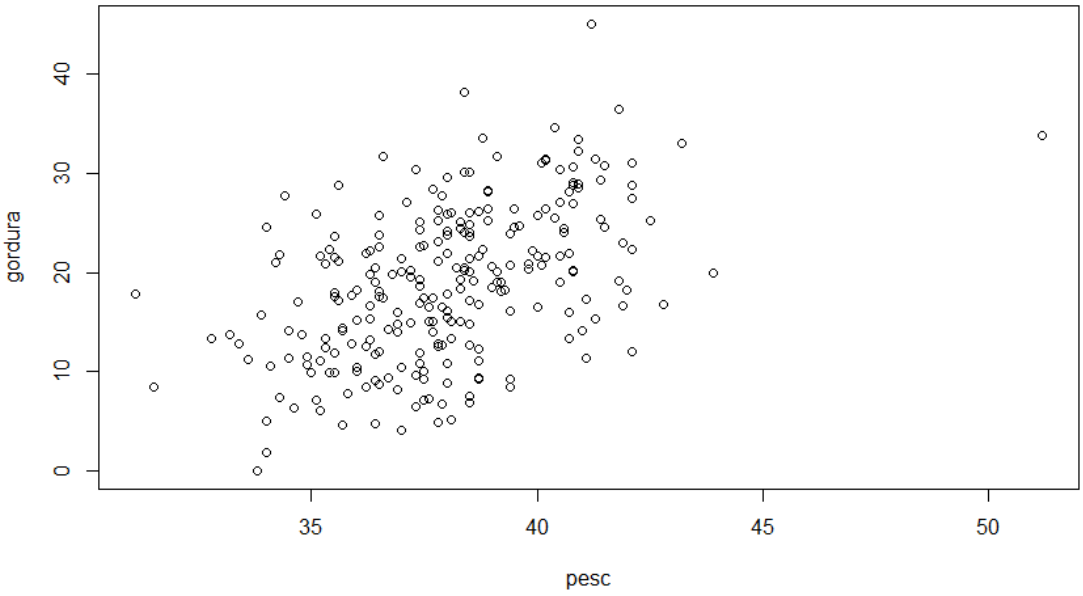
ALTURA X GORDURA



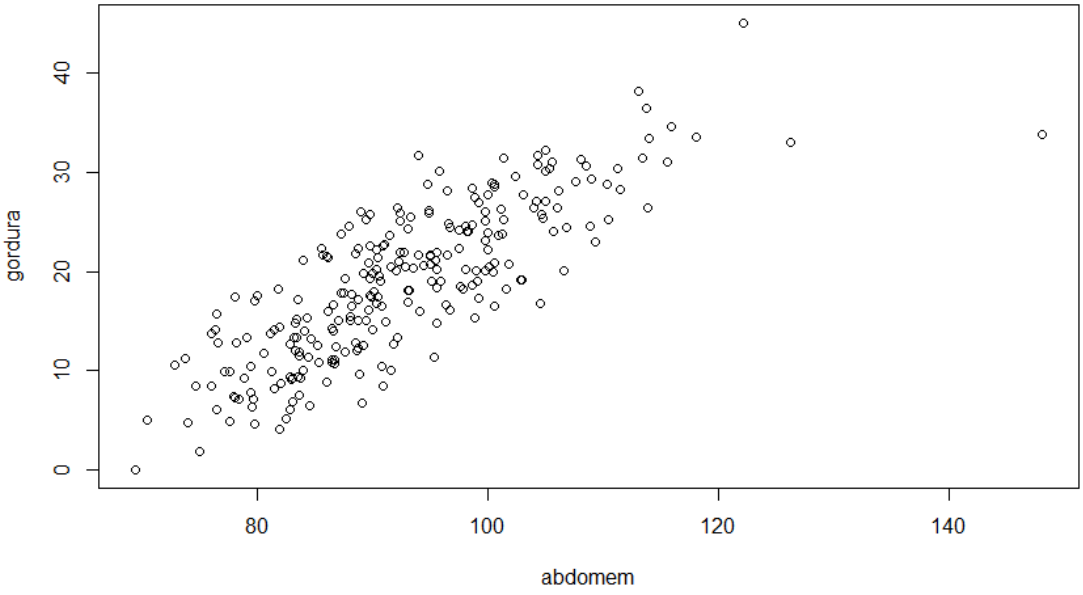
ADIP X GORDURA



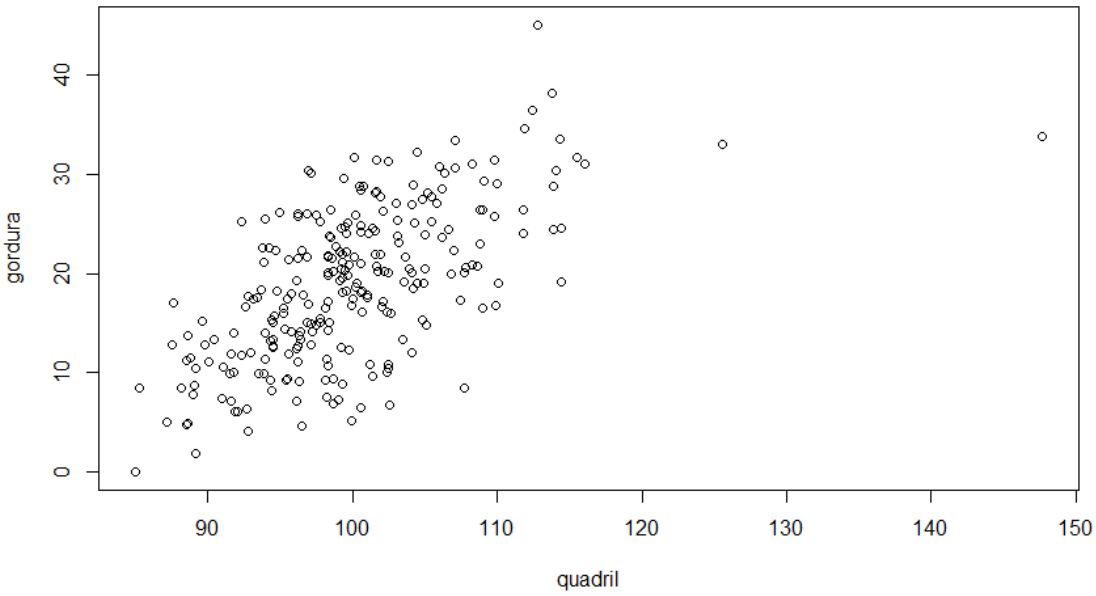
PESC X GORDURA



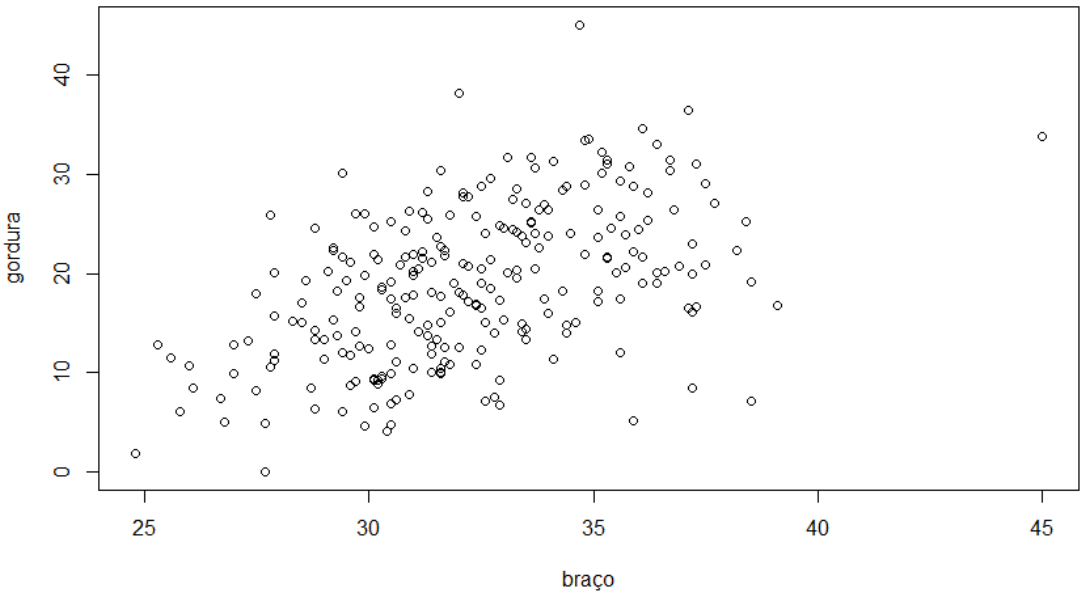
ABDOMEM X GORDURA



QUADRIL X GORDURA



BRAÇO X GORDURA



Através da análise dos gráficos podemos notar que existe uma significativa relação da **gordura** com as variáveis **Peso, Adiposidade e Abdômen.** As variáveis **Idade e Altura** não apresentaram nenhuma evidência de relação com a variável resposta **Gordura**

CRIANDO O DATA FRAME

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

ESCALONAMENTO DAS VARIÁVEIS

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

IDENTIFICANDO OCORRÊNCIA DE MULTICOLINEARIDADE

```
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

Através da análise de **VIF (Variance Inflation Factors)** verificamos multicolinearidades em algumas variáveis apresentaram VIF maior que 10 o que pode causar problemas na estimação dos coeficientes. Vamos calcular novamente o VIF apenas com as variáveis que apresentaram maior correlação com a variável resposta gordura, que nesse caso foram peso, adip, abdomem e quadril. Criaremos um conjunto de valores apenas com essas variáveis e calcularemos novamente o VIF.

VARIÁVEL CRIADA PARA RESOLVER O PROBLEMA DE MULTICOLINEARIDADE

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

CALCULANDO O VIF NOVAMENTE:

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

Verificamos que o problema de multicolinearidade foi solucionado, podemos agora ajustar o modelo para a solução do problema.

3 - .AJUSTE DO MODELO

MEDINDO A SIGNIFICÂNCIA DAS VARIAVÉIS NO MODELO PROPOSTO.

```
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
```

MODELO 1

Para criarmos um modelo de regressão múltipla multiplicaríamos a variável resposta gordura mais todas as outras variáveis independentes como segue exemplo do modelo abaixo.

```
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**.

Porém esse primeiro modelo não é um modelo adequado para o estudo pois verificamos que algumas variáveis apresentaram pouca ou nenhuma significância para a explicação da variável dependente.

Algumas variáveis apresentaram valores muitos próximos de 0 ou seja variáveis com pouca significância para explicar a variável dependente. Não há porque colocar essas variáveis com valores muito perto de 0 no modelo. O Teste T nos mostra que devemos aceitar a hipótese H0 pois os betas são muito próximos de 0 . Após tirarmos as variáveis pouco significativas dos modelos vamos estimar novamente o modelo com as variáveis mais significativas .

MODELO 2

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

Usando o Stepwize para escolha das variáveis que devem compor o modelo final. O modelo final selecionado resultou na seguinte equação:

Gordura = 10,498 + Idade 0,156 - Altura 0,297 - Braço 0,689 - Produto 3,066

MODELO FINAL

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

4– INFERÊNCIA DOS PARAMÊTROS DO MODELO

O método Stepwise para a seleção de variáveis é muito usado em regressão linear. Qualquer procedimento para seleção ou exclusão de variáveis de um modelo é baseado em um algoritmo que checa a importância das variáveis, incluindo ou excluindo-as do modelo se baseando em uma regra de decisão. A importância da variável é definida em termos de uma medida de significância estatística do coeficiente associado à variável para o modelo. Essa estatística depende das suposições do modelo.

Foi usado um intervalo de confiança de (95%) para testar as estimativas da regressão, o resultado nos mostra que o modelo sugerido pelo stepwize está certo pois ele eliminou a variável Pesc pois continha um intervalo de confiança iguais a 0 e não se pode considerar valores 0 por esse motivo não utilizaremos a variável Pesc No modelo final.

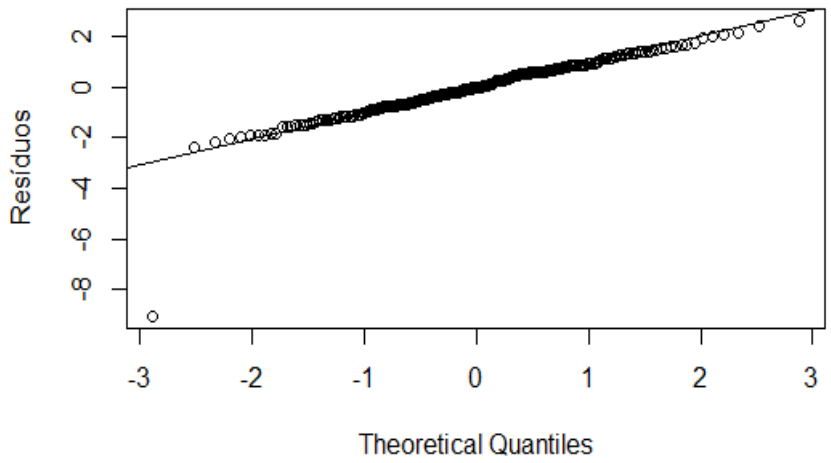
```
confint(modproduto)
confint(modfinal)
```

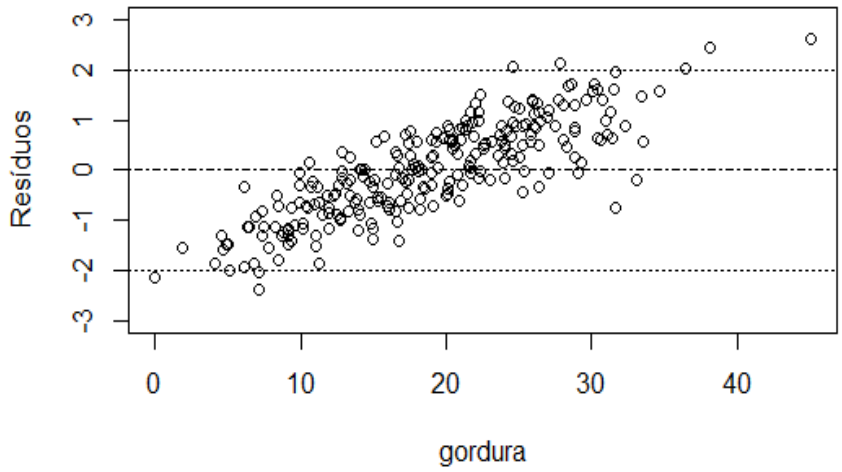
Variável	ID modelo 1		IC modelo final	
Intercepto	-11.617	27.775	-6.490	27.486
Idade	0.093	0.213	0.098	0.215
Altura	-0.521	-0.098	-0.502	-0.092
Pescoço	-0.393	0.647	---	
Braço	0.273	1.019	0.361	1.018
Produto	1.816	4.052	2.088	4.043

5 - DIAGNOSTICO DO MODELO

Usaremos os gráficos de resíduos para analisar a regressão.

```
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)
```





Através das análises dos resíduos podemos notar que os resíduos permanecem dentro do intervalo de confiança estimado de 95%. Logo o modelo é valido.

6 CONSIDERAÇÕPES FINAIS

O banco de dados apresentou multicolinearidade entre as variáveis sendo necessária para a ajuste do modelo que se fizesse um conjunto de valores nele contendo as variáveis peso, adiposidade, abdomem e quadril. A variável criada produto junto com as outras variáveis explicou assim o modelo indicado.

Depois de feito o ajuste das variáveis o modelo indicado pelo *stepwise* o modelo ideal seria composto pelas variáveis Idade, Altura, Circunferência do Braço e Produto.

Foi possível avaliar a validade do modelo proposto pois os gráficos de resíduos usados para os testes mostraram que os dados se concentravam dentro do intervalo de confiança usado no trabalho que foi de 95%. Ou seja o modelo ajustado é sim capaz de explicar a variável resposta do estudo em questão.