

Curso de Estatística no R

Análise de Dados

Prof. Enivaldo Carvalho da Rocha

Monitores: Antônio Fernandes e Anderson Henrique

Mestrado Profissional em Políticas Públicas - MPPp
Departamento de Ciência Política (DCP/UFPE)

24 de novembro de 2018

Sumário

- 1 Noções do R
- 2 Estatística descritiva
Gráficos
- 3 Testes para média e proporção
- 4 Correlação
- 5 Regressão linear

Operações Aritméticas

O R é uma ferramenta para realizar cálculos estatísticos e gráficos de alta qualidade. As operações básicas de soma, subtração, multiplicação e potência podem ser obtidas utilizando os seguintes operadores: $+ - */ ** .$

- Multiplicar $\rightarrow 18 * 2$ [1]36
- Primeiro se calcula a multiplicação $\rightarrow 2 + 3 * 4$ [1]14
- Calcula-se a operação entre parênteses e depois a multiplicação $\rightarrow (2 + 3) * 4$ [1]20
- Primeiro se calcula a potência $\rightarrow 4 * 3^2$ [1]36
- Primeiro resolve o parêntese $\rightarrow (4 * 3)^2$ [1]144

Operador Lógico

Tabela: Verdadeiro ou Falso

Operador	Descrição
<	menor que
>	maior que
<=	menor ou igual
>=	maior ou igual
==	igual a
!=	diferente
	ou
!	não

Exemplos

- $(4 * 3)^2 == 4^2 * 3^2$
TRUE
- $TRUE == 1$
TRUE
- $FALSE == 0$
TRUE
- $x < -21$
x
21
- $rm(x)$
- x
Error: object 'x' not found
<- 21

Vetores

Minúscula e Maiúscula

R faz diferença entre letras maiúsculas e minúsculas, x e X por exemplo. Pode-se atribuir um conjunto de valores a uma variável usando o comando `c()`, neste caso denominaremos essa variável de vetor.

- `vetor <- c(1, 2, 3, 4)`
vetor
1 2 3 4

Posição de um elemento num vetor

O número dentro dos colchetes representa a posição do elemento cujo valor é apresentado ao realizarmos o comando

- `vetor[2]`
2

Vetores

- Valores entre a posição 2 e 4
`vetor[2:4]`
2 3 4
- Valores na posição exata
`vetor[c(1,3)]`
1 3

Extração de elementos de um vetor

Subtraindo os valores nas posições indicadas entre os parentes do comando `c()`

- `vetor[-c(2,3)]`
1 4

Operadores Lógicos com Vetores

- 1 `vetor <- c(100,18,41,21,53)`
`vetor > 50`
TRUE FALSE FALSE TRUE
- 2 `vetor[vetor > 50]`
100 53
- 3 `vetor[vetor > 20 & vetor < 80]`
41 21 53

multiplicar um vetor por uma constante

- 4 `vetor * 3`
300 123 63 159

Operações com Vetores

- Soma de dois vetores
`vetor2 <- c(27,28,29,30,31)`
`vetor + vetor2`
127 46 70 51 84
- Soma dos elementos de um vetor
`sum(vetor)`
233
- Soma acumulada dos elementos de um vetor
`cumsum(vetor)`
100 118 159 180 233

Matriz

Criando um vetor coluna:

```
> matrix(1:6)
```

1

2

3

4

5

6

Cria uma matriz com valores de 1 a 6 em 2 linhas

```
> matrix(1:6,nrow=2)
```

1 3 5

2 4 6

Cria um vetor x com valores entre 3 e 8:

```
> x<-3:8  
> x  
3 4 5 6 7 8
```

Transforma o vetor x em uma matriz com dimensão 3×2 , 3 linhas e 2 colunas:

```
> matrix(x,3,2)  
3 6  
4 7  
5 8
```

Matriz $\times 3 \times 2$:

```
> matrix(x,ncol=2)
```

```
3 6
```

```
4 7
```

```
5 8
```

Cria uma matriz $\times 2 \times 3$ organizada por coluna:

```
> matrix(x,ncol=3)
```

```
3 5 7
```

```
4 6 8
```

Cria uma matriz $\times 2 \times 3$ organizada por linha

```
> matrix(x,ncol=3,byrow=TRUE)
```

```
3 4 5
```

```
6 7 8
```

Exemplo de uma Matriz de dados

O vetor abaixo contém os dados sobre investimento em pesquisa e desenvolvimento pelos setores público e privado em porcentagem do PIB, usando o comando `c()` vamos criar o vetor denominado `dados` com as seguintes variáveis: população em milhões de habitantes, pib em trilhões de dolares, PD privado, PD público e total

- ```
dados <- c(126,4.9,2.4,0.6,3.0,50, 3.6,
2.1,0.7,2.8,325,18.6,1.6,0.8,2.4,205,2.05,0.59,0.65,1.24)
> dados
126,4.9,2.4,0.6,3.0,50, 3.6,
2.1,0.7,2.8,325,18.6,1.6,0.8,2.4,205,2.05,0.59,0.65,1.24
```

### Exemplo: Vetor de nomes - regioao

```
> região.dados <- matrix(dados,nrow=4,byrow=TRUE)
```

```
> região.dados
```

```
126 4.90 2.40 0.60 3.00
```

```
50 3.60 2.10 0.70 2.80
```

```
325 18.60 1.60 0.80 2.40
```

```
205 2.05 0.59 0.65 1.24
```

## Nomes

```
> dimnames(região.dados)
```

```
NULL
```

```
> dim(região.dados)
```

```
4 5
```

```
> região <- c("Japão", "Corea", "Estados Unidos", "Brasil")
```

```
> região
```

```
"Japão", "Corea", "Estados Unidos", "Brasil"
```

## Variáveis

```
> Variaveis <- c("Pop", "Pib", "PDpri", "PDpub", "TotalPD")
```

```
> Variaveis
```

```
"Pop", "Pib", "PDpri", "PDpub", "TotalPD"
```

```
> dimnames(região.dados) <- list(região, NULL)
> região.dados
Japão 126 4.90 2.40 0.60 3.00
Corea 50 3.60 2.10 0.70 2.80
Estados Unidos 325 18.60 1.60 0.80 2.40
Brasil 205 2.05 0.59 0.65 1.24
```

```
> dimnames(região.dados)<-list(NULL,Variaveis)
> região.dados
Pop Pib PDprPDpubTotalPD
126 4.90 2.40 0.60 3.00
50 3.60 2.10 0.70 2.80
325 18.60 1.60 0.80 2.40
205 2.05 0.59 0.65 1.24
```

```
> dimnames(região.dados) <- list(região, Variaveis)
> região.dados
Pop Pib PDpri PDpub TotalPD
Japão 126 4.90 2.40 0.60 3.00
Coreia 50 3.60 2.10 0.70 2.80
Estados Unidos 325 18.60 1.60 0.80 2.40
Brasil 205 2.05 0.59 0.65 1.24
```

## nomes das linhas e colunas

```
dimnames(região.dados)
```

```
[1]
```

```
"JapãoCoreaEstados UnidosBrasil"
```

```
[2]
```

```
"PopPibPD_priPD_pubTotal_PD"
```

## Acessando elementos na matriz

- região.dados[1,2]  
4.9
- região.dados[2,1:5]
- Pop Pib PD\_pri PD\_pub Total\_PD  
50.0 3.6 2.1 0.7 2.8
- região.dados[2,]  
Pop Pib PD\_pri PD\_pub Total\_PD  
50.0 3.6 2.1 0.7 2.8
- região.dados["Brasil", "PD\_pri"]  
0.59
- região.dados["Corea", "PD\_pri"]  
2.1

## Lendo CSV

- `variavel=read.csv("dados.csv", header=T, dec=",")`

### Exemplo:

```
h <- read.csv("homi_4.csv", sep=";", dec=",", header=TRUE)
head(h)
```

```
class(h)
```

```
"data.frame"
```

|       | RO | PA | PE  | SP   | Brasil |
|-------|----|----|-----|------|--------|
| Jan 1 | 25 | 49 | 261 | 1059 | 3334   |
| Feb 1 | 24 | 54 | 301 | 1068 | 3342   |
| Mar 1 | 30 | 49 | 269 | 1120 | 3533   |
| Apr 1 | 22 | 61 | 247 | 1040 | 3224   |
| May 1 | 29 | 40 | 247 | 1028 | 3184   |
| Jun 1 | 30 | 53 | 275 | 998  | 3195   |

## Escrevendo CSV

### Write CSV in R

- `write.csv(MyData, file = "MyData.csv")`

### Exemplo:

```
write.csv(h,file = "h.csv")
```

## Lendo TXT

- `variavel=read.table("mydata.txt", header=T, dec=",")`

### Exemplo:

```
t <- read.table("homi_4.txt", sep=";", dec=",", header=TRUE)
head(t)
```

```
class(t)
```

```
"data.frame"
```

|       | RO | PA | PE  | SP   | Brasil |
|-------|----|----|-----|------|--------|
| Jan 1 | 25 | 49 | 261 | 1059 | 3334   |
| Feb 1 | 24 | 54 | 301 | 1068 | 3342   |
| Mar 1 | 30 | 49 | 269 | 1120 | 3533   |
| Apr 1 | 22 | 61 | 247 | 1040 | 3224   |
| May 1 | 29 | 40 | 247 | 1028 | 3184   |
| Jun 1 | 30 | 53 | 275 | 998  | 3195   |

## Escrevendo TXT

### Write TXT in R

- `write.table(MyData, file = "MyData.csv")`

### Exemplo:

```
write.table(t,file = "t.txt")
```

Exemplo:

## Número de Homicídios

- O arquivo homi\_4.csv contém os dados mensais do número de homicídios dos estados de Rondônia, Pará, Pernambuco, São Paulo e Brasil, de janeiro de 1996 a dezembro de 2016. Após ajustar um modelo autorregressivo integrado de médias móveis ARIMA para série temporal de cada estado do Brasil, com base nos registros dos homicídios passados, as previsões dos homicídios nessas regiões foram obtidas considerando um horizonte de previsão de 4 anos, ou 48 meses. Isto é de janeiro de 2017 a dezembro de 2020.

## A matriz das séries temporais

```
Read homi_4 setwd("/Curso de Estatística no R")
h <- read.csv("homi_4.csv", sep=";", dec=".", header=TRUE)
head(h)
class(h)
"data.frame"
transformar o dataframe na classe ts (série temporal)
```

```
h <- ts(homi_4, start=c(1996,1), end=c(2016,12), frequency=12)
class(h)
"mts" "ts" "matrix"
```

```
>head(h[,3:5])
```

|       | PE  | SP   | Brasil |
|-------|-----|------|--------|
| Jan 1 | 261 | 1059 | 3334   |
| Feb 1 | 301 | 1068 | 3342   |
| Mar 1 | 269 | 1120 | 3533   |
| Apr 1 | 247 | 1040 | 3224   |
| May 1 | 247 | 1028 | 3184   |
| Jun 1 | 275 | 998  | 3195   |

## Extraíndo as séries da matriz

```
ro <- h[,1]
pa <- h[,2]
pe <- h[,3]
sp <- h[,4]
br <- h[,5]
```

## Estimando os modelos

```
ro_ <- auto.arima(ro)
pa_ <- auto.arima(pa)
pe_ <- auto.arima(pe)
sp_ <- auto.arima(sp)
br_ <- auto.arima(br)
```

## Calculando as previsões até 2020

```
ro_arima <- forecast(ro_, level=c(80, 95, 99), h=48)
pa_arima <- forecast(pa_, level=c(80, 95, 99), h=48)
pe_arima <- forecast(pe_, level=c(80, 95, 99), h=48)
sp_arima <- forecast(sp_, level=c(80, 95, 99), h=48)
br_arima <- forecast(br_, level=c(80, 95, 99), h=48)
```

## Gráficos das previsões - Comandos

```
Plot ARIMA Models
par(mfrow=c(2,2))
```

```
plot(forecast(ro_arima)
,main = "Rondônia"
,type = "l"
,lwd = 2.5,ylim=c(0,700), shadecols="oldstyle")
```

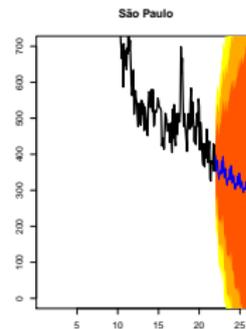
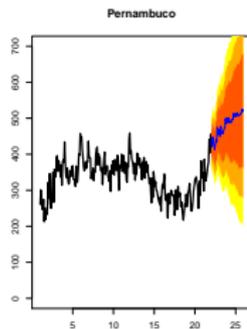
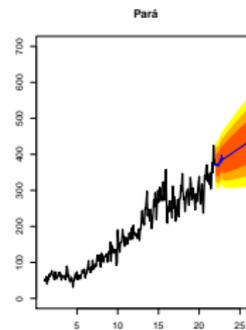
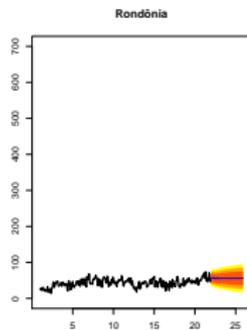
```
plot(forecast(pa_arima)
,main = "Pará"
,type = "l"
,lwd = 2.5,ylim=c(0,700), shadecols="oldstyle")
```

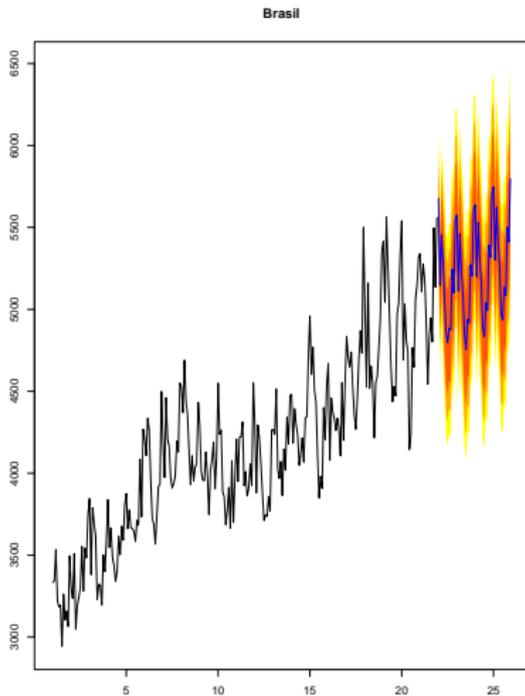
```
plot(forecast(pe_arima)
,main = "Pernambuco"
,type = "l"
,lwd = 2.5,ylim=c(0,700), shadecols="oldstyle")
```

```
plot(forecast(sp_arima)
,main = "São Paulo"
,type = "l"
,lwd = 2.5, ylim=c(0,700), shadecols="oldstyle")
```

```
par(mfrow=c(1,1))
plot(forecast(br_arima)
```

## Gráficos das previsões - Estados





## Extraíndo as previsões

```
rom <- as.matrix(summary(ro_arima))
pam <- as.matrix(summary(pa_arima))
pem <- as.matrix(summary(pe_arima))
spm <- as.matrix(summary(sp_arima))
brm <- as.matrix(summary(br_arima))
```

## Montando a matriz dos estados e Brasil

```
previsão <- cbind(
rom[,1]
,pam [,1]
,pem [,1]
,spm [,1]
,brm [,1]
)
```

## identificação das colunas

```
estado <- c(
"ro"
,"pa "
,"pe "
,"sp "
,"br "
)
```

```
nomes <- rownames(previsão)
dimnames(previsão) <- list(nomes, estado)
pre_homi_4 <- as.data.frame(previsão)
write.csv(pre_homi_4, file = "pre_homi_4.csv")
head(pre_homi_4)
```

|        | ro       | pa       | pe       | sp       | br       |
|--------|----------|----------|----------|----------|----------|
| Jan 22 | 54.80717 | 372.4143 | 435.6678 | 397.8348 | 5677.726 |
| Feb 22 | 57.12986 | 369.1166 | 421.0979 | 353.0196 | 5146.784 |
| Mar 22 | 55.76695 | 374.0775 | 448.3554 | 384.0494 | 5455.604 |
| Apr 22 | 56.11193 | 368.6771 | 432.4629 | 380.6745 | 5284.609 |
| May 22 | 57.99066 | 375.0971 | 424.8627 | 345.8581 | 5061.442 |
| Jun 22 | 55.04819 | 367.5340 | 413.6192 | 330.8261 | 4858.719 |

# Estatística Descritiva

banco de dados: Bussab e Morettin

lendo o arquivo milsa.csv

```
milsa <- read.csv("milsa.csv", sep=";",
dec="," ,header=TRUE)
head(milsa)
class(milsa)
```

Extraindo as variáveis do dataframe

Variáveis

```
civil <- milsaestciv
edu <- milsaeducacao
filho <- milsafilhos
ano <- milsaano
sal <- milsasalarario
origem <- milsaorigem
```

# Estatística Descritiva

## Variáveis Qualitativas

### Tabelas

```
table(origem)
```

```
table(origem,civil)
```

```
table(origem,civil,edu)
```

### Tabelas de proporção

```
prop.table(table(edu))
```

## Varáveis Qualitativas

### Resumo das estatísticas

```
summary(sal)
```

```
summary(sal[civil=="casado"])
```

```
summary(sal[civil=="solteiro"])
```

```
summary(origem)
```

# Média e Variância

## Média

```
mean(filho)
```

```
mean(filho, na.rm=TRUE)
```

## Variância

```
var|(filho)
```

```
var(filho, na.rm=TRUE)
```

## Desvio Padrão

```
sd(filho)
```

```
sd(filho, na.rm=TRUE)
```

ou a raiz quadrada da variância

```
sqrt(var(filho, na.rm=TRUE))
```

# Mediana

## Médiana

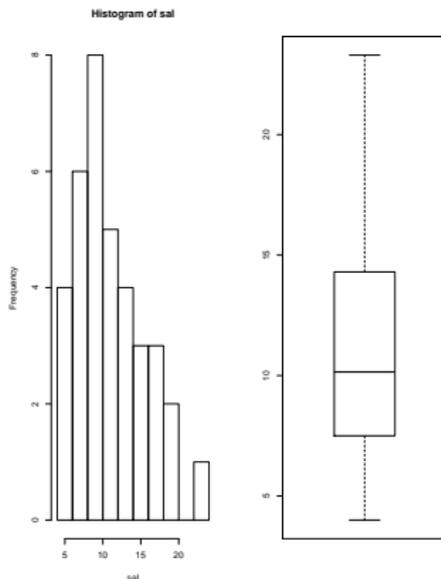
```
median(filho)
median(filho, na.rm = "TRUE")
sort(filho)
```

## Exemplo

```
sal
sort(sal)
length(sal)
mean(sal)
median(sal)
```

# Distribuição da variável salário

```
par(mfrow=c(1,2))
hist(sal)
boxplot(sal)
```



# Resumo dos comandos utilizados

read.csv

head

class

sum

cumsum

length

summary

table

prop.table(table())

corr

mean

var

sd

median

tapply

barplot

plot

hist

boxplot

pie

# Média

- $sum(vetor)/length(vetor)$   
46.6

## O comando `mean( )`

Cálculo da média usando o comando direto

- `mean(vetor)`  
[1] 46.6

## Definição

Se as observações numa amostra de tamanho  $n$  são  $x_1, x_2, \dots, x_n$ , então a média aritmética é:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

# Mediana dos Elementos de um Vetor

## O comando sort

### Ordenando os elementos do vetor

- `svetor <- sort(vetor)`  
svetor  
18 21 41 53 100
- `svetor[5/2+1]`  
41

## Definição da Mediana

Seja  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$  denotar uma amostra em ordem crescente, então a mediana será:

$$\tilde{x} = x_{\frac{n+1}{2}}, \text{ se } n \text{ é ímpar}$$

$$\tilde{x} = \frac{x_{\frac{n}{2}} + x_{(\frac{n}{2}+1)}}{2}, \text{ se } n \text{ é par}$$

- `median(vetor)`  
41

# Separatrizes

## Percentil

Representa o elemento que divide a amostra e ordena em percentis, por exemplo: o percentil 50 é igual a mediana.

## Quantil

O comando `quantile` apresenta o valor que divide os dados em percentis 25, 50 e 75.

- `quantile(vetor, probs=0.5)`  
50  
41
- `quantile(vetor, probs=c(0.25,0.75))`  
25 75  
21 53
- `diff(quantile(vetor,probs=c(0.25,0.75)))`  
75

# O Comando Summary

O comando summary fornece as estatísticas mínimo, q1, mediana, média, q3 e o máximo de uma variável quantitativa.

- Exemplo: Considere X o peso do cérebro de uma amostra de 19 animais e Y a sua massa corporal em gramas.

$X = c(1176, 273, 151, 123, 110, 289, 165, 119, 95, 32, 700, 166, 118, 115, 41, 28, 5.2, 2.6, 0.5)$

$Y = c(78000, 60000, 16000, 37000, 11000, 780000, 230000, 72000, 25000, 4500, 272000, 35000, 50000, 50000, 22000, 60000, 2000, 23, 19)$

summary(X)

Min. 1st Qu. Median Mean 3rd Qu. Max.  
0.5 36.5 118.0 195.2 165.5 1176.0

summary(Y)

Min. 1st Qu. Median Mean 3rd Qu. Max.  
19 13500 37000 94976 66000 780000

# Medidas de Variabilidade

## Range

É a medida de variação mais simples, range amostral, definido como a diferença entre a maior e a menor das observações na amostra

$$r = \max(x_i) - \min(x_i)$$

## Intervalo interquartil

Definido como a diferença entre o 3o quartil e o primeiro quartil

$$IQR = q3 - q1$$

- Exemplo: Considere a amostra (1, 5, 5, 5, 7, 7, 9)

```
n <- c(1,5,5,5,7,7,9)
```

```
r <- 9 - 1
```

```
IQR <- quantile(n, probs = 0.75) - quantile(n, probs = 0.25)
```

ou

```
IQR <- diff(quantile(n, probs=c(0.25,0.75)))
```

# Variância e Desvio Padrão

## Variância

Se as observações numa amostra de tamanho  $n$  são  $x_1, x_2, \dots, x_n$ , então a variância amostral é:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

e o desvio padrão amostral é a raiz quadrada positiva de  $s^2$

- Exemplo: Considere a amostra (1, 5, 5, 5, 7, 7, 9)  
`n <- c(1,5,5,5,7,7,9)`  
`var(n) [1] 6.285714 sqrt(var(n)) [1] 2.507133 mean(n)`

```
...1.....2.....3.....4.....5.....6.....7.....8.....9...
.....|.....-2.5.....5.6.....+2.5.....|.....
```

O coeficiente de variação é uma medida adimensional muito útil para avaliação de amostras de diferentes dimensões e tamanhos.

$$CV_1 = \frac{s_1}{\bar{X}_1}$$

Exemplo: Considere a amostra do peso do cérebro (X) e a massa corporal (Y) dos 19 animais, e as notas dos alunos da disciplina análise de dados do mppp 2017.2 (nota).

```
> notas <- read.csv("notas.csv", sep=";", dec=".",header=TRUE) > nota <- notas$notas
```

```
s1 <- sqrt(var(X))
> s2 <- sqrt(var(Y))
> s3 <- sqrt(var(nota))
> m1 <- mean(X)
> m2 <- mean(Y)
> m3 <- mean(nota)
> CVx <- s1/m1
> CVy <- s2/m2
> CVnota <- s3/m3
> CVx
1.461587
> CVy
1.908605
> CVnota
0.305887
```

# Distribuição de Frequência

```
notas <- read.csv("notas.csv", sep=";", dec=".",header=TRUE)
```

## Ordenação das Notas

```
sort(notas$notas)
```

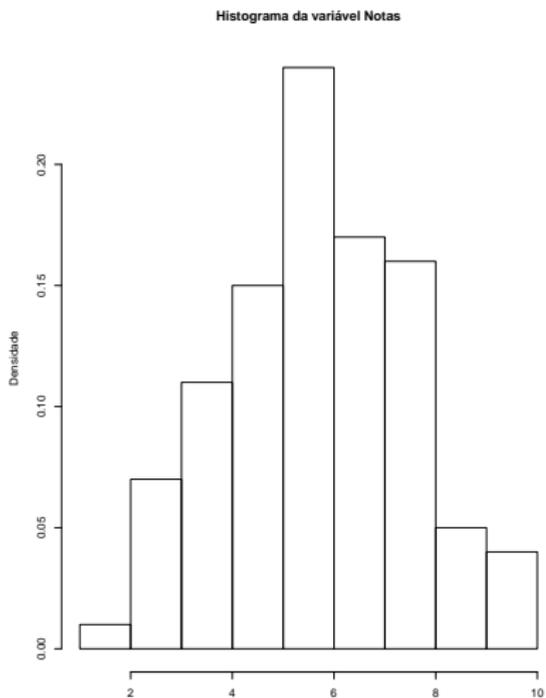
```
1.52.52.52.53.03.03.03.03.53.53.53.53.54.04.04.04.04.04.04.54.54.54.5
4.5 4.5 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.5 5.5 5.5 5.5 5.5 5.5 5.5 5.5 5.5 5.5 6.0
6.0 6.0 6.0 6.0 6.0 6.0 6.0 6.0 6.0 6.0 6.0 6.0 6.5 6.5 6.5 6.5 6.5 6.5 6.5 6.5 6.5 6.5 6.5
7.0 7.0 7.0 7.0 7.0 7.0 7.5 7.5 7.5 7.5 7.5 7.5 7.5 7.5 7.5 7.5 7.5 7.5 8.0 8.0 8.0 8.0 8.0 8.5
8.5 8.5 9.0 9.0 9.5 10.0 10.0 10.0
```

**Tabela:** Distribuição de Frequências das Notas na  
Disciplina Análise de Dados: MPPP - 2017.2

| Classes | Frequência | Porcentagem |
|---------|------------|-------------|
| [1,2)   | 1          | 1%          |
| [2,3)   | 3          | 3%          |
| [4,5)   | 12         | 12%         |
| [5,6)   | 20         | 20%         |
| [6,7)   | 24         | 24%         |
| [7,8)   | 16         | 16%         |
| [8,9)   | 9          | 9%          |
| [9,10)  | 6          | 6%          |
| Total   | 100        | 100%        |

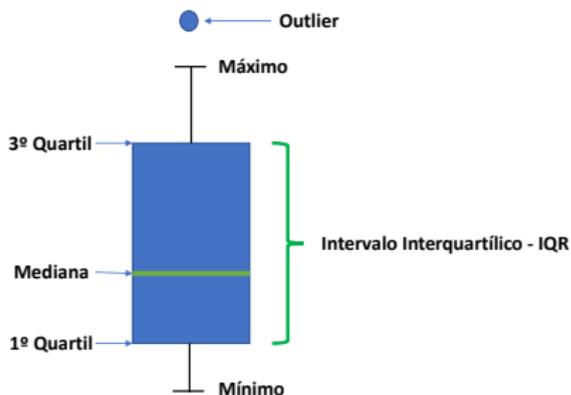
# Histograma

```
hist(notas$notas, main="Histograma da variável Notas",prob=T, xlab="Notas", ylab="Densidade")
```



# BoxPlot

- O boxplot representa um excelente método para detectar outlier presentes na distribuição da variável, o fato de usar a mediana como uma medida de centralidade permite que a distribuição fique livre da influência de pontos extremos.

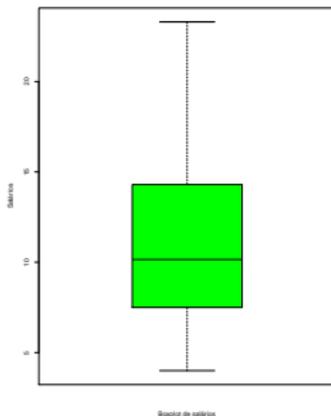


- Onde:  $\text{Mínimo} = Q1 - 1,5 * \text{IQR}$  e  $\text{Máximo} = Q3 + 1,5 * \text{IQR}$

# BoxPlot - Exemplos

Exemplo 1 - Considere as notas da disciplina análise de dados e verifique se tem algum outlier presente na distribuição.

```
boxplot(notas$notas, main = "Notas - Análise de Dados",
ylab="Notas", col="green")
```



## Dados: Bussab

```
m <- read.csv("milsa.csv", sep=";", dec=",", header=TRUE)
```

O R possui uma enorme capacidade para gerar diversos tipos de gráficos de alta qualidade totalmente configuráveis, desde cores e tipos de linhas, até legendas e textos adicionais.

Opções :// xlim: (início,fim) dupla contendo os limites do eixo X.

ylim: (início,fim) dupla contendo os limites do eixo Y.

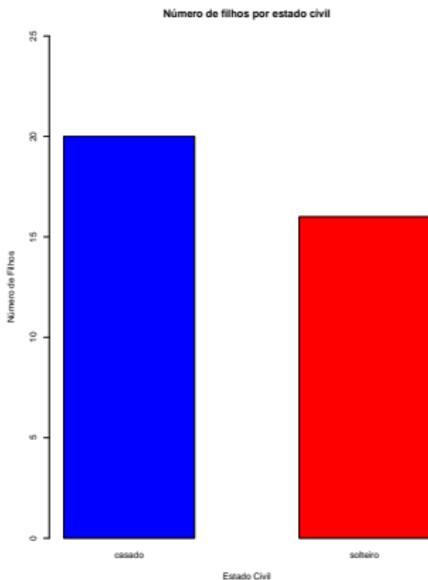
xlab: rótulo para o eixo X.

ylab: rótulo para o eixo Y.

main: título principal do gráfico.

col: cor de preenchimento do gráfico, podendo ser um vetor. A lista das cores disponíveis pode ser obtida através do comando colors().

```
barplot(table(mestciv), col = c("blue", "red"), ylim = c(0, 25), space = .8, width = c(.2, .2), main = "Número de filhos por estado civil", xlab = "Estado Civil", ylab = "Número de Filhos")
```



## Notas na disciplina análise de dados

Comando :

```
hist(dados, opções)
```

opções:

prob: T plota a densidade.

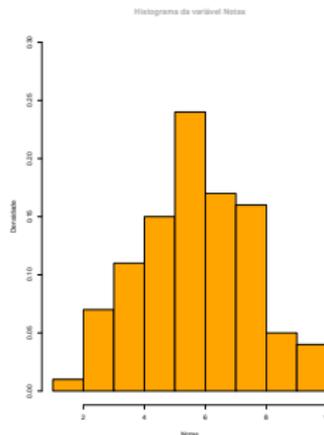
F plota a frequência absoluta.

breaks: vetor contendo os pontos de definição das larguras das barra do histograma.

```
notas <- read.csv("notas.csv", sep=";", dec=".",header=TRUE)
```

```
head(notas)
```

```
hist(notas$notas, main="Histograma da variável Notas", prob=T, xlab="Notas", ylab="Densidade",
col=c("orange"), ylim=c(0,0.3), col.main="darkgray")
```



## Boxplot - Dados Bussab

Comando :

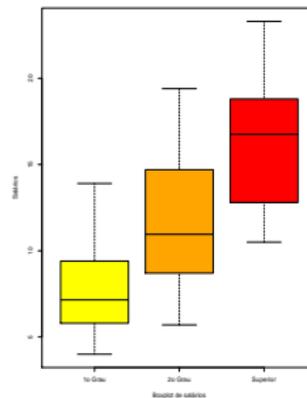
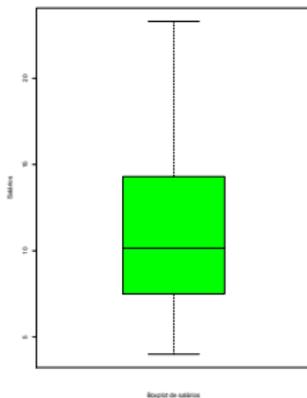
```
boxplot(dados, opções)
```

opções:

outline: T plota os outliers.

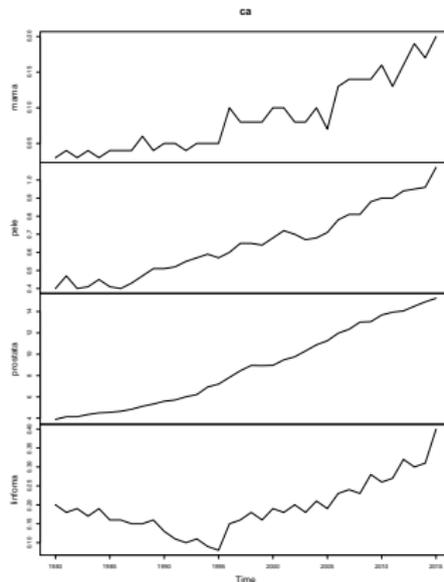
```
boxplot(m$salario, xlab="Boxplot de Salários", ylab="Salários", col="green")
```

```
boxplot(m$salario ~ m$educacao, xlab="Boxplot de salários", ylab="Salários", col=c("yellow", "orange", "red"))
```



## Datasus

```
require(graphics)
setwd("/Curso de Estatística no R/Dados da Saúde")
tx <- read.csv("tx_c_a4.csv", sep = ";", dec = ",", header = TRUE)
ca <- ts(tx, frequency = 1, start = c(1980, 1))
class(ca)
plot(ca)
```



## Tipos de câncer

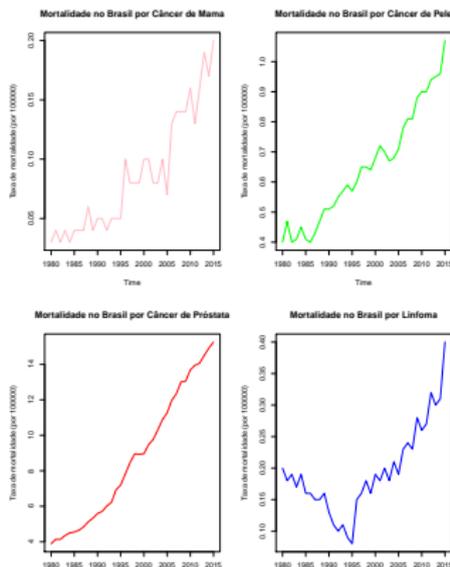
```
par(mfrow=c(2,2))
```

```
plot(ca[,1], ylab="Taxa de mortalidade (por 100000)", main="Mortalidade no Brasil por Câncer de Mama",
col="pink")
```

```
plot(ca[,2], ylab="Taxa de mortalidade (por 100000)", main="Mortalidade no Brasil por Câncer de Pele",
col="green")
```

```
plot(ca[,3], ylab="Taxa de mortalidade (por 100000)", main="Mortalidade no Brasil por Câncer de Próstata",
col="red")
```

```
plot(ca[,4], ylab="Taxa de mortalidade (por 100000)", main="Mortalidade no Brasil por Linfoma", col="blue")
```

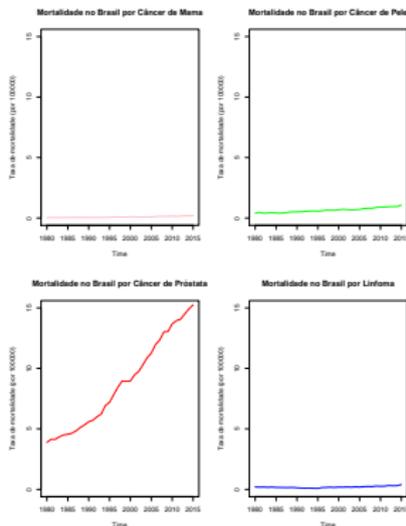


## Tipos de câncer

```

par(mfrow=c(2,2))
plot(ca[,1], ylab="Taxa de mortalidade (por 100000)", main="Mortalidade no Brasil por Câncer de Mama",
col="pink",ylim=c(0,15))
plot(ca[,2], ylab="Taxa de mortalidade (por 100000)", main="Mortalidade no Brasil por Câncer de Pele",
col="green",ylim=c(0,15))
plot(ca[,3], ylab="Taxa de mortalidade (por 100000)", main="Mortalidade no Brasil por Câncer de Próstata",
col="red",ylim=c(0,15))
plot(ca[,4], ylab="Taxa de mortalidade (por 100000)", main="Mortalidade no Brasil por Linfoma", col="blue",
ylim=c(0,15))

```



## Gráfico de Dispersão - Dados Bussab e Morettin

Comando:

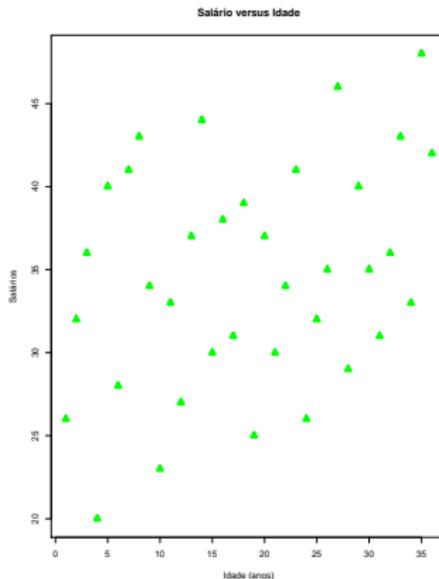
```
plot(dados1, dados2, opções)
```

opções:

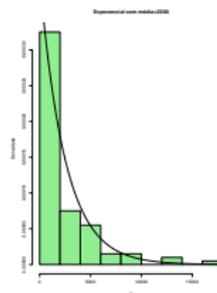
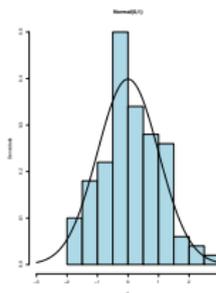
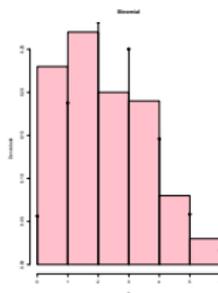
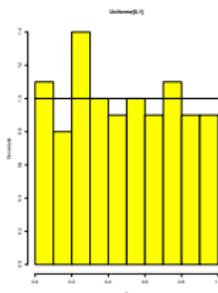
pch: Escolhe o tipo de caractere.

lwd: Espessura do caractere a ser plotado

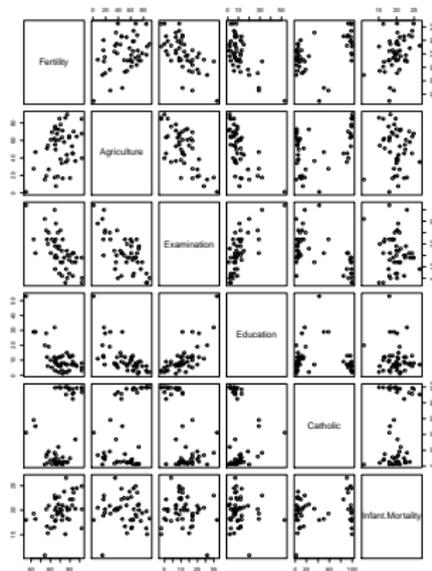
```
plot(mAno, msalario, pch=2, lwd=5, main="Salário versus Idade", xlab="Idade (anos)", ylab="Salários")
```



- Uniforme  
`hist(x,probability=TRUE,main="Uniforme[0,1]",ylab="Densidade",col="yellow")`  
`curve(dunif(x,0,1),add=T)`
- Binomial  
`n <- 10 hspace.2cm p <- 0.25 hspace,2cm x <- rbinom(100,n,p)`  
`hist(x,probability=TRUE,ylab="Densidade",col="pink",main="Binomial", ym=c(0,0.30))`  
`xvalores <- 0:n`  
`points(xvalores,dbinom(xvalores,n,p),type="h",lwd=3)`  
`points(xvalores,dbinom(xvalores,n,p),type="p", lwd=3)`
- Normal  
`x <- rnorm(100)`  
`hist(x,probability=TRUE,col="lightblue",main="Normal(0,1)",ylab="Densidade",ylim=c(0,0.5),xlim=c(-3,3))`  
`curve(dnorm(x),add=T)`
- Exponencial `x <- rexp(100,1/2500)`  
`hist(x,probability=TRUE,col="lightgreen",main="Exponencial com média=2500",ylab="Densidade")`  
`curve(dexp(x,1/2500),add=T)`



```
install.packages("datasets")
head(iris)
head(swiss)
pairs(swiss)
```



|   | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
|---|--------------|-------------|--------------|-------------|---------|
| 1 | 5.1          | 3.5         | 1.4          | 0.2         | setosa  |
| 2 | 4.9          | 3.0         | 1.4          | 0.2         | setosa  |
| 3 | 4.7          | 3.2         | 1.3          | 0.2         | setosa  |
| 4 | 4.6          | 3.1         | 1.5          | 0.2         | setosa  |
| 5 | 5.0          | 3.6         | 1.4          | 0.2         | setosa  |
| 6 | 5.4          | 3.9         | 1.7          | 0.4         | setosa  |

begintable[]

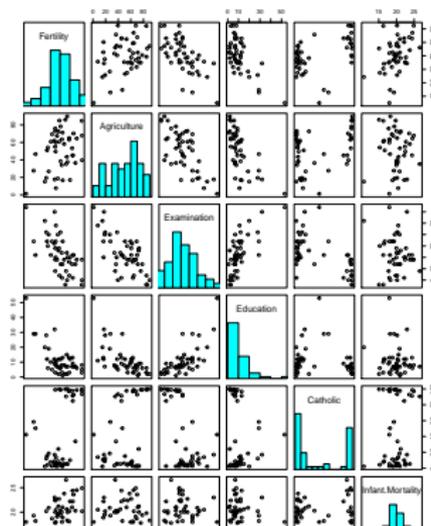
|              | Fertility | Agriculture | Examination | Education | Catholic | Infant.Mortality |
|--------------|-----------|-------------|-------------|-----------|----------|------------------|
| Courtelary   | 80.2      | 17.0        | 15          | 12        | 9.96     | 22.2             |
| Delemont     | 83.1      | 45.1        | 6           | 9         | 84.84    | 22.2             |
| Franches-Mnt | 92.5      | 39.7        | 5           | 5         | 93.40    | 20.2             |
| Moutier      | 85.8      | 36.5        | 12          | 7         | 33.77    | 20.3             |
| Neuveville   | 76.9      | 43.5        | 17          | 15        | 5.16     | 20.6             |
| Porrentruy   | 76.1      | 35.3        | 9           | 7         | 90.57    | 26.6             |

## Matrizes de Gráficos

```

panel.hist <- function(x, ...)
{
 usr <- par("usr"); on.exit(par(usr))
 par(usr = c(usr[1:2], 0, 1.5))
 h <- hist(x, plot = FALSE)
 breaks <- h$breaks; nB <- length(breaks)
 y <- h$count; y <- y/max(y)
 rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
}
pairs(swiss, diag.panel = panel.hist)

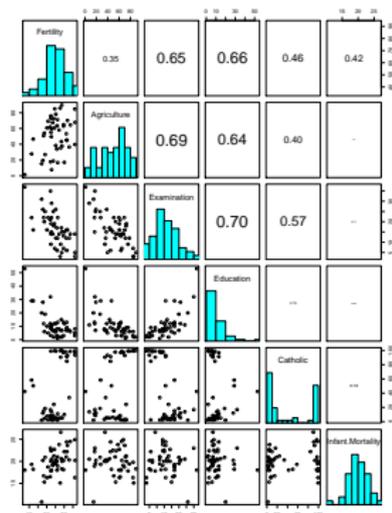
```



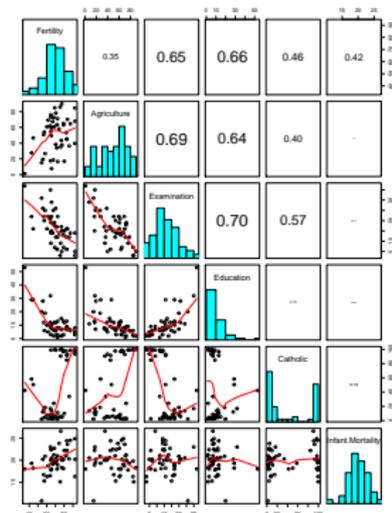
```

panel.cor <- function(x, y, digits = 2, prefix = , cex.cor, ...)
usr <- par("usr"); on.exit(par(usr))
par(usr = c(0, 1, 0, 1))
r <- abs(cor(x, y))
txt <- format(c(r, 0.123456789), digits = digits)[1]
txt <- paste0(prefix, txt)
if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
text(0.5, 0.5, txt, cex = cex.cor * r)
panel.cor <- function(x, y, digits = 2, prefix = , cex.cor, ...)
pairs(swiss, diag.panel = panel.hist, upper.panel = panel.cor)

```



```
pairs(swiss, diag.panel = panel.hist, upper.panel = panel.cor,
lower.panel = panel.smooth)
```

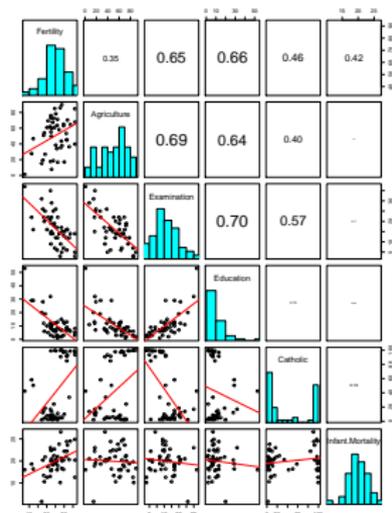


## Matrizes de Gráficos

```

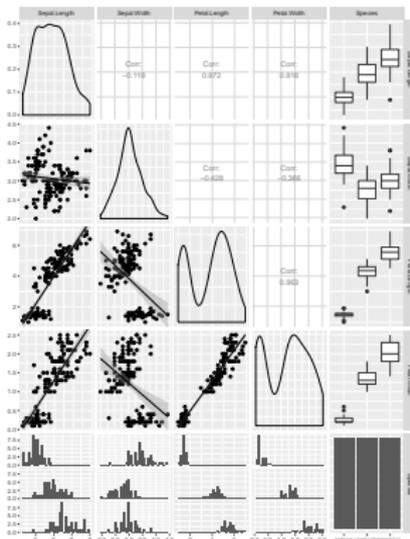
panel.lm <- function(x, y, col = par("col"), bg = NA, pch = par("pch"),
 cex = 1, col.line="red")
points(x, y, pch = pch, col = col, bg = bg, cex = cex)
ok <- is.finite(x) is.finite(y)
if (any(ok))
abline(lm(y[ok] ~ x[ok]), col = col.line)
pairs(swiss, diag.panel = panel.hist, upper.panel = panel.cor,
lower.panel = panel.lm)

```



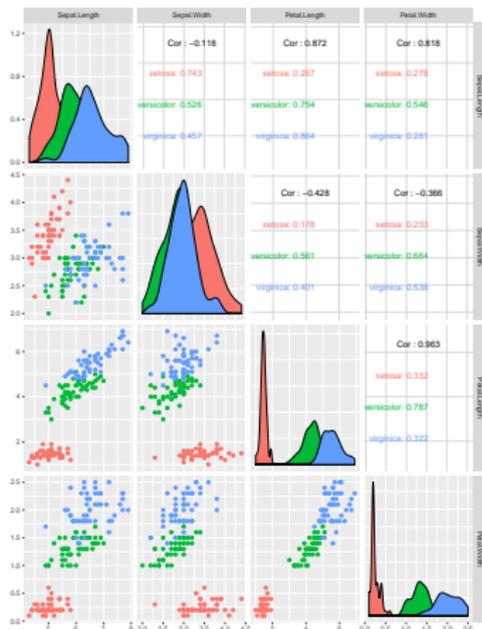
## Carregando o Pacote GGally

```
require(GGally)
library(GGally)
ggpairs(iris, lower = list(continuous = "smooth"))
```



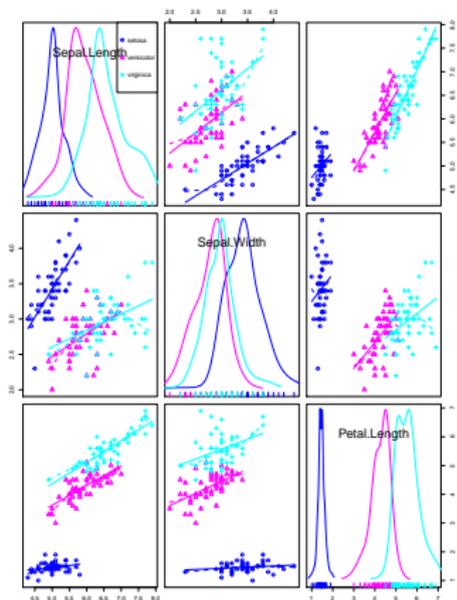
## Carregando o Pacote GGally

```
ggpairs(iris, columns = 1:4, ggplot2::aes(colour=Species))
```



## Carregando o Pacote GGally

```
car::spm(Sepal.Length + Sepal.Width + Petal.Length|Species, data = iris, by.group=TRUE)
```

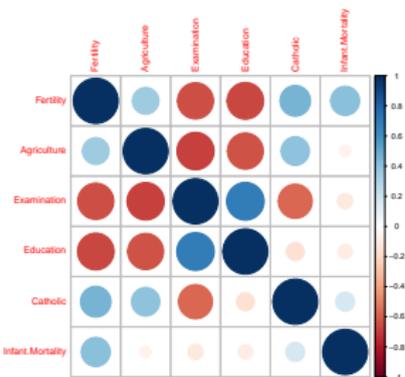




```
ggcorr(swiss, label=T)
```

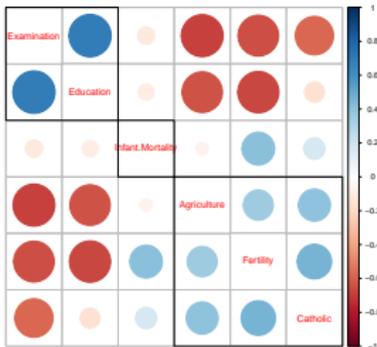


```
library(corrplot)
primeiro fazemos a matriz de correlação
M <- cor(swiss)
corrplot(M, method = "circle")
```



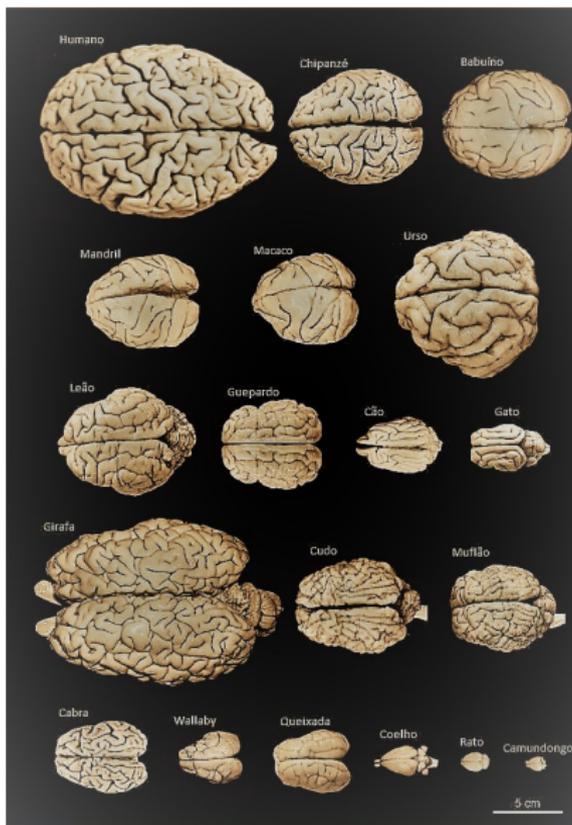
## Agrupando pela Correlação

```
corrplot(M, order = "hclust", addrect = 3, tl.pos="d")
```



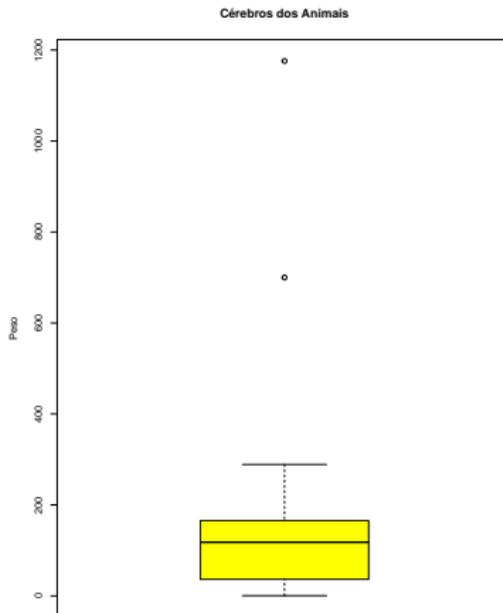


## Exemplo 2: Cérebro e Peso do Animal



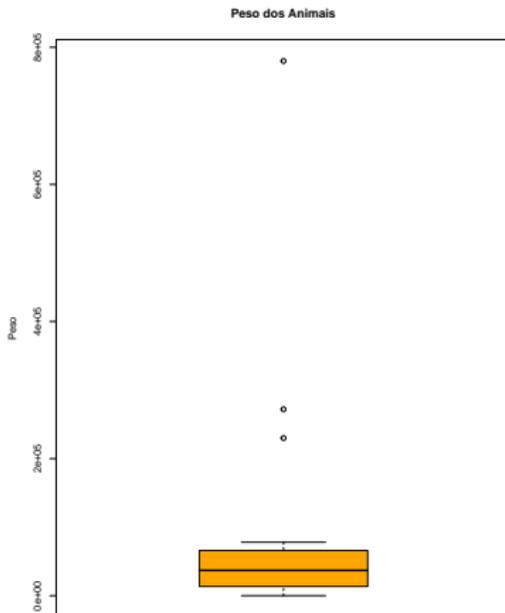
# Cérebro

```
boxplot(cerebro$cerebro, main="Cérebro dos Animais",
ylab="Peso", col="yellow")
```



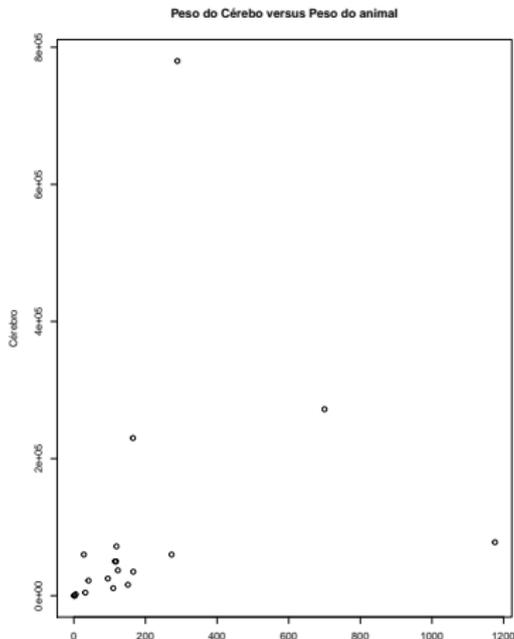
# Peso

```
boxplot(cerebro$peso, main="Peso dos Animais", ylab="Peso",
col="orange")
```



# Tamanho do Cérebro e Peso do Animal

```
plot(cerebro$cerebro,cerebro$peso,main=paste("Peso do Cérebro
versus Peso do animal"),ylab="Cérebro",xlab="Peso")
```



# Tamanho do Cérebro e Peso do Animal

```
cepe <- data.frame(cerebro$animal,100*ce_pe)
cepe
```

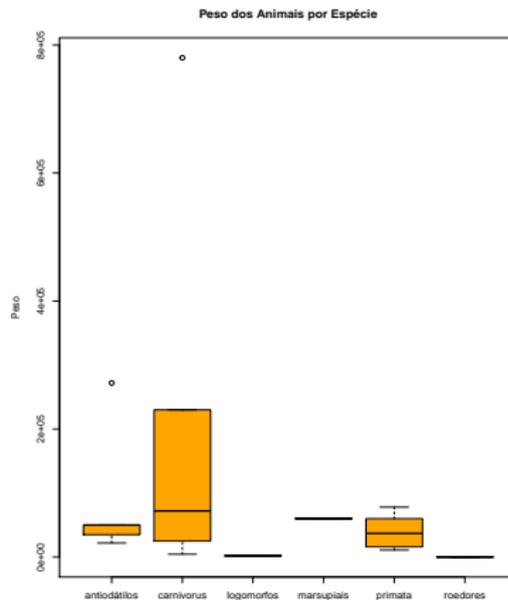
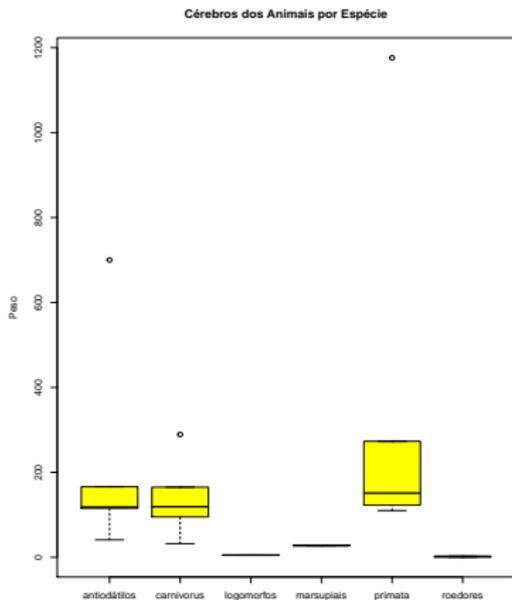
| Animal     | Cérebro (gr) | Peso (gr) | Relação (%) |
|------------|--------------|-----------|-------------|
| Beija-flor | 1            | 7         | 14,29%      |
| Rato       | 2,6          | 23        | 11,30%      |
| Camundongo | 0,5          | 19        | 2,63%       |
| Humano     | 1176         | 78000     | 1,51%       |
| Macaco     | 110          | 11000     | 1,00%       |
| babuíno    | 151          | 16000     | 0,94%       |
| Vaca       | 5600         | 720000    | 0,78%       |
| Gato       | 32           | 4500      | 0,71%       |
| Cudo       | 166          | 35000     | 0,47%       |
| Chipanzé   | 273          | 60000     | 0,46%       |
| Cão        | 95           | 25000     | 0,38%       |
| Mandrill   | 123          | 37000     | 0,33%       |
| Coelho     | 5,2          | 2000      | 0,26%       |
| Girafa     | 700          | 272000    | 0,26%       |
| Muflão     | 118          | 50000     | 0,24%       |
| Cabra      | 115          | 50000     | 0,23%       |
| Queixada   | 41           | 22000     | 0,19%       |
| Guepardo   | 119          | 72000     | 0,17%       |
| Elefante   | 5000         | 5000000   | 0,10%       |
| Leão       | 165          | 230000    | 0,07%       |
| Urso       | 289          | 780000    | 0,04%       |

# Usando o Fator

```

boxplot(cerebro$cerebro ~ cerebro$especie, main="Cérebro dos Animais por Espécie", ylab="Peso",
col="yellow")
boxplot(cerebro$peso ~ cerebro$especie, main="Peso dos Animais por Espécie", ylab="Peso", col="orange")

```



## Dados utilizados da tabela 2.1 de Bussab e Morettin (2003).

```
dados <- read.csv("milsa.csv", sep=";", dec=".", header=TRUE)
head(dados)
```

```
Funcionário estciv educacao Filhos Salario Ano Mês origem
1 1 solteiro 1o Grau 0 4.00 26 3 interior
2 2 casado 1o Grau 1 4.56 32 10 capital
```

- `table(dados$origem)`  
capital interior outro  
11 12 13
- `table(dados$origem,dados$estciv)`  
casado solteiro  
capital 7 4  
interior 8 4  
outro 5 8
- `table(dados$origem,dados$estciv,dados$educa)` , , = 1o Grau  
casado solteiro  
capital 2 2  
interior 1 2  
outro 2 3  
 , , = 2o Grau  
casado solteiro  
capital 4 1  
interior 6 1  
outro 2 4  
 , , = Superior  
casado solteiro  
capital 1 1  
interior 1 1  
outro 1 1

# Tabela de proporções

- ```
prop.table(table(dados$educacao))
1o Grau 2o Grau Superior
0.3333333 0.5000000 0.1666667
```
- ```
prop.table(table(dados$estciv, dados$origem))
capital interior outro
casado 0.1944444 0.2222222 0.1388889
solteiro 0.1111111 0.1111111 0.2222222
```
- ```
prop.table(table(dados$origem,dados$estciv,dados$educa))
, , = 1o Grau
casado solteiro
capital 0.05555556 0.05555556
interior 0.02777778 0.05555556
outro 0.05555556 0.08333333
, , = 2o Grau
casado solteiro
capital 0.11111111 0.02777778
interior 0.16666667 0.02777778
outro 0.05555556 0.11111111
, , = Superior
casado solteiro
capital 0.02777778 0.02777778
interior 0.02777778 0.02777778
outro 0.02777778 0.02777778
```

Summary

Variável quantitativa

- `summary(dados$Salario)`
Min. 1st Qu. Median Mean 3rd Qu. Max.
4.00 7.55 10.15 11.13 14.10 23.30
- `summary(dados$Salario[dados$estciv=="solteiro"])`
Min. 1st Qu. Median Mean 3rd Qu. Max.
4.000 7.275 9.050 9.881 11.700 18.800
- `summary(dados$Salario[dados$estciv=="casado"])`
Min. 1st Qu. Median Mean 3rd Qu. Max.
4.600 8.775 11.950 12.135 15.025 23.300

Variável categórica

- `summary(dados$origem)`
capital interior outro
11 12 13

O Dilema do Prisioneiro

Considere H_0 a hipótese nula e H_a a hipótese alternativa, o objetivo é decidir se rejeita ou aceita H_0 .
Hipóteses:

- 1 H_0 : O réu é culpado
- 2 H_a : O réu é inocente

	Decisão	Realidade	
		Inocente	Culpado
julgamento	Rejeita H_0	Correto	Erro I
	Aceita H_0	Erro II	Correto

Controle de erro: Colocar um criminoso em liberdade

- 1 H_0 : O réu é inocente
- 2 H_a : O réu é culpado

	Decisão	Realidade	
		Inocente	Culpado
julgamento	Rejeita H_0	Erro I	Correto
	Aceita H_0	Correto	Erro II

Controle de erro: Colocar um inocente na prisão

Testes para a média populacional

t.test()

Realiza o teste t-Student para uma ou duas amostras.

sintaxe: t.test(amostra1, amostra2, opções)

Opções:

- 1 alternative: string indicando a hipótese alternativa desejada.
Valores possíveis: "two-sided", "less" ou "greater".
- 2 mu: valor indicando o verdadeiro valor da média populacional para o caso de uma amostra,
ou a diferença entre as médias para o caso de duas amostras.
- 3 paired: TRUE – realiza o teste t pareado.
FALSE – realiza o teste t não pareado.
- 4 var.equal: TRUE – indica que a variância populacional é a igual nas duas amostras.
FALSE – indica que a variância populacional de cada amostra é diferente.
- 5 conf.level: coeficiente de confiança do intervalo.

Considere a seguinte amostra:

amostra1 = c(14.9,13.4,14.5,13.5,15.0,13.9,14.9,16.4,14.6,15.4)

Testar

$H_0 : \mu = 15$

$H_1 : \mu \neq 15$

t.test(amostra1,mu=15)

One Sample t-test

data: amostra1

t = -1.2252, df = 9, p-value = 0.2516

alternative hypothesis: true mean is not equal to 15

95 percent confidence interval:

14.00375 15.29625

sample estimates:

mean of x

14.65

Considere as seguintes amostras:

amostra1 = c(16.6,13.4,14.6,15.1,12.9,15.2,14.0,16.6,15.4,13.0)

amostra2 = c(15.8,17.9,18.2,20.2,18.1,17.8,18.3,18.6,17.0,18.4)

Testar

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Two Sample t-test

data: amostra1 and amostra2

t = -6.0257, df = 18, p-value = 1.069e-05

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-4.518003 -2.181997

sample estimates:

mean of x mean of y

14.68 18.03

Considere as seguintes amostras pareadas:

antes = c(16.6,13.4,14.6,15.1,12.9,15.2,14.0,16.6,15.4,13.0)

depois = c(15.8,17.9,18.2,20.2,18.1,17.8,18.3,18.6,17.0,18.4)

Testar

$$H_0 : \mu_{antes} = \mu_{depois}$$

$$H_1 : \mu_{antes} \neq \mu_{depois}$$

Paired t-test

data: antes and depois

t = -5.3231, df = 9, p-value = 0.000479

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-4.773642 -1.926358

sample estimates:

mean of the differences

-3.35

Teste para proporção

`prop.test()`

Realiza o teste de proporções para uma ou duas amostras.

sintaxe: `prop.test(x, n, p, opções)`

Parâmetros

x: Vetor contendo o número de sucessos em cada amostra.

n: Vetor contendo o número de realizações de cada amostra.

p: Vetor contendo as probabilidades de sucesso de cada amostra.

Opções:

- `alternative`: string indicando a hipótese alternativa desejada.
Valores possíveis: "two-sided", "less" ou "greater".
- `conf.level`: coeficiente de confiança do intervalo.
`correct`: TRUE – indica que a correção de continuidade de Yates será aplicada.
FALSE – indica que a correção de continuidade não será aplicada.

Teste para uma proporção populacional

Testar

$$H_0 : P = P_0$$

$$H_1 : P \neq P_0$$

```
prop.test(104,200,0.6,correct=F)
```

1-sample proportions test without continuity correction

data: 104 out of 200, null probability 0.6

X-squared = 5.3333, df = 1, p-value = 0.02092

alternative hypothesis: true p is not equal to 0.6

95 percent confidence interval:

0.4510379 0.5882083

sample estimates:

p

0.52

Teste para comparação de duas proporções

Testar

$$H_0 : P_1 = P_2$$

$$H_1 : P_1 \neq P_2$$

```
prop.test(c(104,50),c(200,95),correct=F)
```

2-sample test for equality of proportions without continuity correction

data: c(104, 50) out of c(200, 95)

X-squared = 0.010297, df = 1, p-value = 0.9192

alternative hypothesis: two.sided

95 percent confidence interval:

-0.1282799 0.1156483

sample estimates:

prop 1 prop 2

0.5200000 0.5263158

Teste de Médias - Banco Iris

Ronald Fisher em seu artigo de 1936

O conjunto de dados consiste em 50 amostras de cada uma das três espécies de Iris (Iris setosa , Iris virginica e Iris versicolor). Quatro características foram medidas a partir de cada amostra: o comprimento e a largura das sépalas e pétalas , em centímetros. Com base na combinação dessas quatro características, Fisher desenvolveu um modelo discriminante linear para distinguir as espécies umas das outras.

title

```
datasets::iris  
comp_sepala <- iris$Sepal.Length  
comp_petala <- iris$Petal.Length  
larg_sepala <- iris$Sepal.Width  
larg_petala <- iris$Petal.Width
```

Teste de Médias - Banco Iris

Teste t.test

```
t.test(comp_sepala, mu=5.6)
t.test(comp_sepala, mu=5.7)
t.test(comp_sepala, mu=5.8)
t.test(comp_petala, mu=3)
t.test(larg_sepala, mu=3)
```

Cria o Fator - as.factor

```
tipo <- as.factor(iris$Species)
is.factor(tipo)
```

```
boxplot(comp_sepala tipo)
boxplot(larg_sepala tipo)
```

Cálculo das Médias dos Comprimentos e Larguras das Sepalas - mean

```
mean(comp_sepala[tipo=="setosa"])
mean(comp_sepala[tipo=="versicolor"])
mean(comp_sepala[tipo=="virginica"])
mean(larg_sepala[tipo=="setosa"])
mean(larg_sepala[tipo=="versicolor"])
mean(larg_sepala[tipo=="virginica"])
```

Teste de Médias - Banco Iris

Cálculo das Variâncias dos Comprimentos e Larguras das Sepalas - var

```
var(comp_sepala[tipo=="setosa"])
var(comp_sepala[tipo=="versicolor"])
var(comp_sepala[tipo=="virginica"])
var(larg_sepala[tipo=="setosa"])
var(larg_sepala[tipo=="versicolor"])
var(larg_sepala[tipo=="virginica"])
```

Testa a Igualdade de Variâncias para Duas Amostras - var.test

```
var.test(comp_sepala[tipo=="setosa"],comp_sepala[tipo=="versicolor"])
var.test(comp_sepala[tipo=="setosa"],comp_sepala[tipo=="virginica"])
var.test(comp_sepala[tipo=="virginica"],comp_sepala[tipo=="versicolor"])
var.test(larg_sepala[tipo=="setosa"],larg_sepala[tipo=="versicolor"])
var.test(larg_sepala[tipo=="setosa"],larg_sepala[tipo=="virginica"])
var.test(larg_sepala[tipo=="virginica"],larg_sepala[tipo=="versicolor"])
```

Testa a Igualdade de Médias para Duas Amostras - t.test (variâncias desiguais)

```
t.test(comp_sepala[tipo=="setosa"],comp_sepala[tipo=="versicolor"],var.equal=FALSE)
t.test(comp_sepala[tipo=="setosa"],comp_sepala[tipo=="virginica"],var.equal=FALSE)
```

Teste de Médias - Banco Iris

Testa a Igualdade de Médias para Duas Amostras - t.test (variâncias iguais)

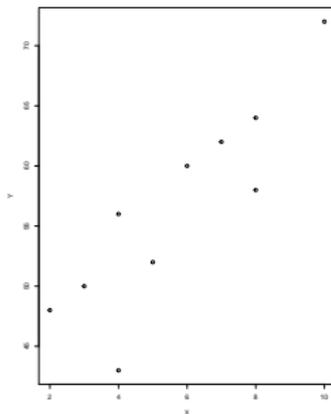
```
t.test(comp_sepala[tipo=="virginica"],comp_sepala[tipo=="versicolor"],var.equal=TRUE)
t.test(larg_sepala[tipo=="setosa"],larg_sepala[tipo=="versicolor"],var.equal=TRUE)
t.test(larg_sepala[tipo=="setosa"],larg_sepala[tipo=="virginica"],var.equal=TRUE)
t.test(larg_sepala[tipo=="virginica"],larg_sepala[tipo=="versicolor"],var.equal=TRUE)
```

Variável	Tipos de Flores Iris			Teste de Variâncias	Teste de Médias
	setosa	versicolor	virginica	Decisão	Decisão
comp_sepala	X	X		rejeita	rejeita
	X		X	rejeita	rejeita
		X	X	aceita	rejeita
larg_sepala	X	X		aceita	rejeita
	X		X	aceita	rejeita
		X	X	aceita	rejeita

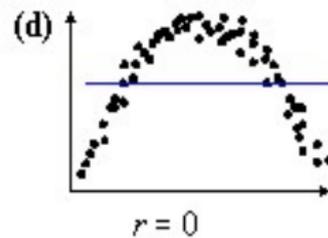
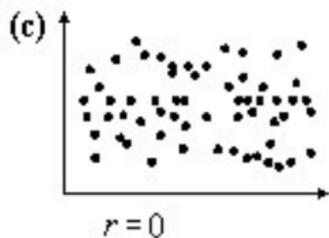
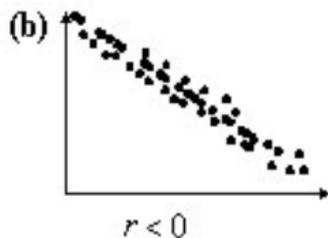
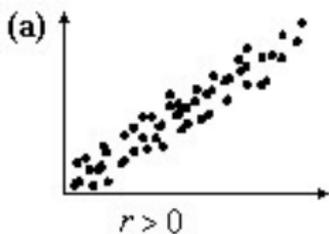
Exemplo:

Considere a amostra do tempo de serviço, em anos, de 10 funcionários de uma companhia de seguros e o número de clientes que cada um conquistou. Será que existe uma relação entre a variável número de clientes e o tempo de serviço do corretor?

X	2	3	4	5	4	6	7	8	8	10
Y	48	50	56	52	43	60	62	58	64	72



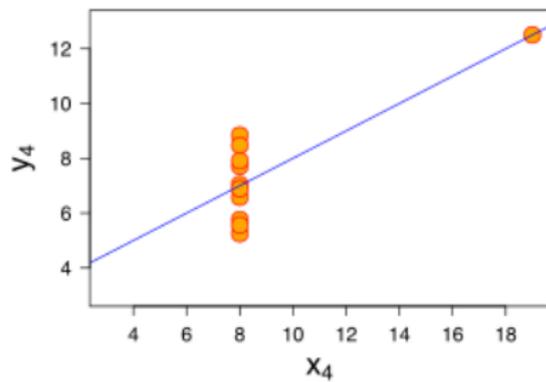
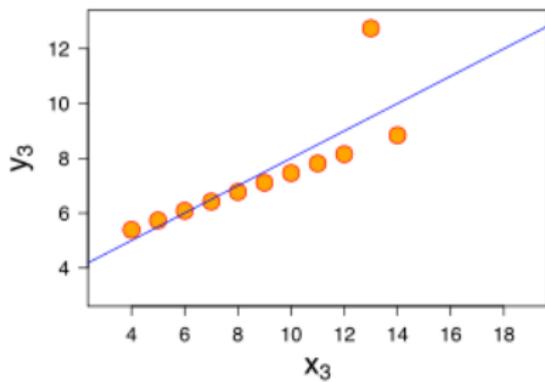
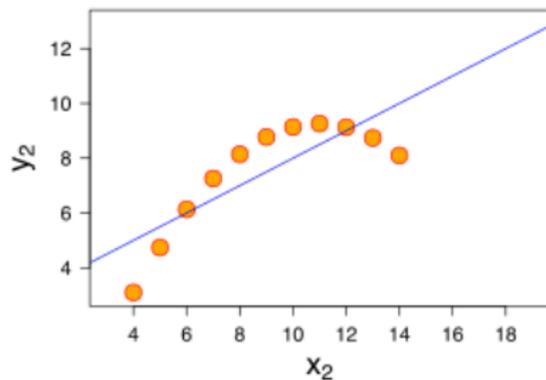
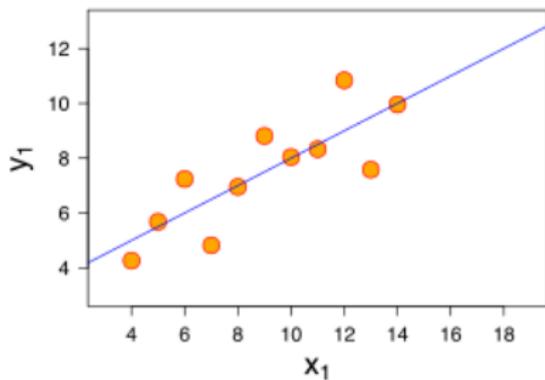
Tipos de correlação



Exemplo: dados anscombe

```
ans <- read.csv("anscombe.csv", sep=";", dec=".", header=TRUE)
head(ans)
x <- ans$Xabc
a <- ans$Ya
b <- ans$Yb
c <- ans$Yc
xd <- ans$Xd
```

```
cor(x,a)
0.8164205
cor(x,b)
0.8162867
cor(x,c)
0.8162867
cor(xd,d)
0.8165214
```



O coeficiente de correlação (r): Mede o grau da relação linear entre os pares de valores (x,y) .

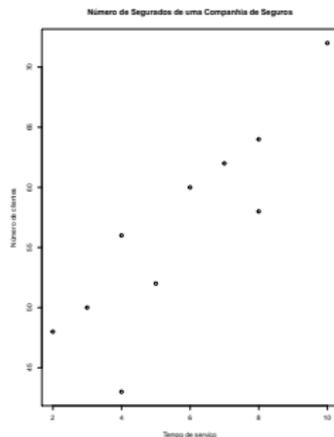
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Comandos R:

- `mean(X)` 5.7
- `mean(Y)` 56.5
- `var(X)` 6.455556
- `var(Y)` 73.16667
- `summary(X)`
- `Min. 1st Qu. Median Mean 3rd Qu. Max.` 2.00 4.00 5.50 5.70 7.75 10.00
- `summary(Y)`
- `Min. 1st Qu. Median Mean 3rd Qu. Max.` 43.0 50.5 57.0 56.5 61.5 72.0
- `cor(X,Y)` 0.8767952

Diagrama de dispersão

```
plot(X, Y, main = "CientesdeumaCompanhiadeSeguros", xlab = "Tempodeserviço(X)", ylab = "Númerodeclientes(Y)")
```



Propriedades da correlação

- 1 O valor de r é limitado entre -1 e 1.
- 2 $r(x, y) = r(y, x)$
- 3 r não mede causalidade.

Teste de hipótese para correlação ρ

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

Estatística do teste:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

```
cor.test(x, y,
alternative = c("two.sided", "less", "greater"), method = c("pearson", "kendall", "spearman"),
exact = NULL, conf.level = 0.95, continuity = FALSE, ...)
```

```
cor.test(X,Y)
```

Pearson's product-moment correlation

data: X and Y

t = 5.5989, df = 5, p-value = 0.00251

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.5841215 0.9896355

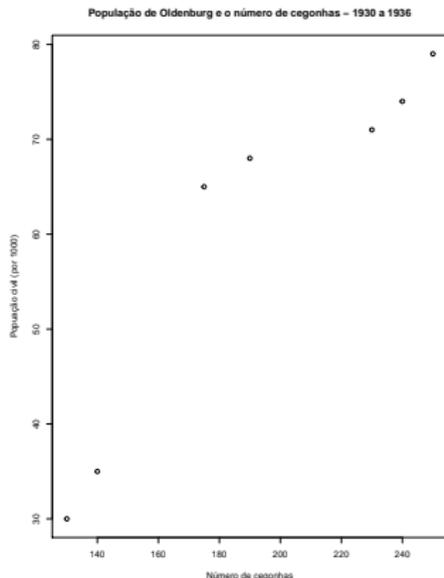
sample estimates:

cor

0.928676

Exemplo: Box, Hunter and Hunter. Statistics for experiments. News York, 1978.

O diagrama de dispersão abaixo apresenta a população (Y) da cidade de Oldenburg, na Alemanha, e o número de cegonhas (X) no final de cada ano, durante o período de sete anos entre 1930 e 1936.



Observando o gráfico acima, podemos ser induzido a concluir que o aumento no número de cegonhas causa um crescimento na população da cidade de Oldenburg. Neste caso, a correlação entre Y e X ocorre devido a um terceiro fator, W. Tanto Y como X cresce sobre o período de 7-anos, o fator comum W é o tempo.

```
cegonha <- read.csv("cegonha.csv", sep=";",  
dec=",", header=TRUE)
```

```
cegonha  
ano cegonha população  
1930 130 30  
1931 140 35
```

```
ano <- cegonha$ano
```

```
X <- cegonha$cegonha
```

```
Y <- cegonha$população
```

```
plot(X,Y, main="População de Oldenburg e o número de cegonhas  
- 1930 a 1936", xlab="Número de cegonhas", ylab="População  
civil (por 1000)")
```

```
cor(ano,cega)
```

```
0.9845357
```

Exercício:

- Obtenha os gráficos (ano,X) e (ano,Y)
- Calcule $r(\text{ano},X)$, $r(X,\text{ano})$, $r(\text{ano},Y)$ e $r(Y,\text{ano})$

Modelo de Regressão

Regressão Linear Simples

$$y = \beta_0 + \beta_1 x + \epsilon$$

Resposta = Parte explicada pelos dados + Resíduos

Onde,

y -> Variável dependente

x -> Variável independente

ϵ -> *Erro*

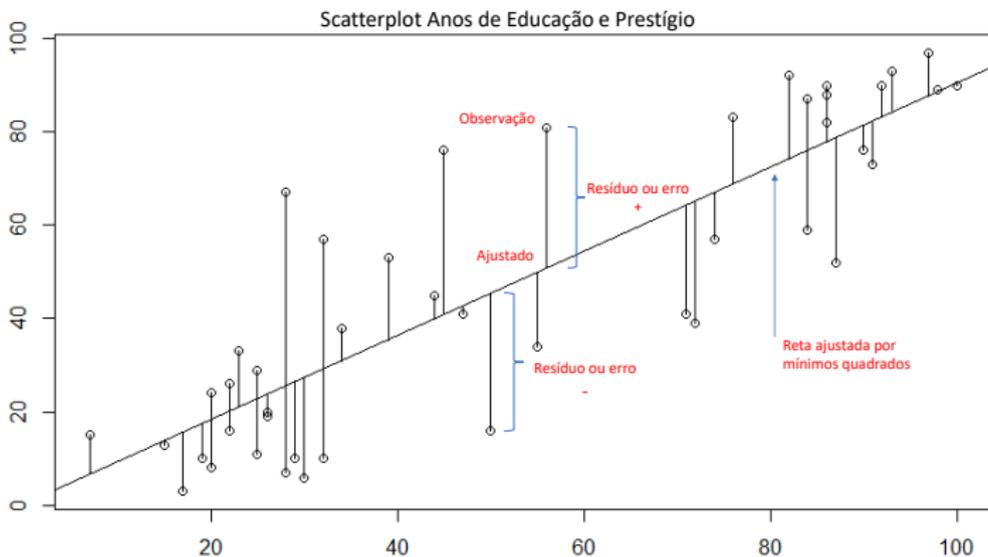
β_0 -> *Constante*

β_1 -> *Coeficiente*

Obs: neste modelo assumimos que y tem uma relação linear com x .

Exemplo: Considere os dados de prestígio e educação, John Fox, 2017.

Modelo de Regressão



Modelo de Regressão

```
pres <- read.csv("prestigio.csv", sep=";", dec=".",header=TRUE)
head(pres)
```

profissao	tipo	renda	educacao	prestigio
Accountant	prof	62	86	82
Pilot	prof	72	76	83
Architect	prof	75	92	90
Author	prof	55	90	76
Chemist	prof	64	86	90
Minister	prof	21	84	87

```
lm.pres <- lm(pres$prestigio ~ pres$educacao)
summary(lm.pres)
```

Call:

```
lm(formula = pres$prestigio ~ pres$educacao)
```

Residuals:

Min	1Q	Median	3Q	Max
-29.384	-11.834	-0.484	9.222	41.460

Coefficients:

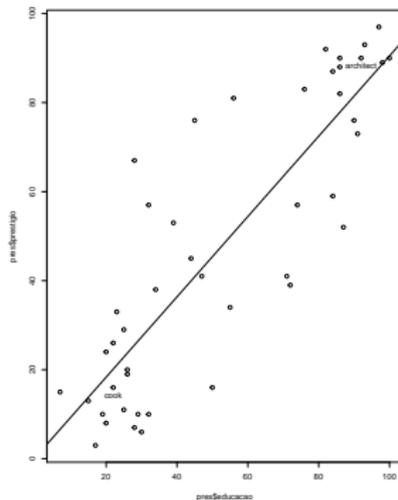
	Estimate	Std. Error	t value	value
(Intercept)	0.284000	5.09306	0.056	0.956
pres\$educacao	0.90200	0.08455	10.668	1.17e-13

Residual standard error: 16.69 on 43 degrees of freedom, Multiple R-squared: 0.7258, Adjusted R-squared: 0.7194

Modelo de Regressão

Obtendo a reta ajustada

```
plot(pres$educacao,pres$prestigio)
abline(lm(pres$prestigio ~ pres$educacao))
identify(pres$educacao,pres$prestigio, labels=pres$profissao)
```



Suposições sobre os erros:

- 1 Os erros tem média zero.
- 2 Os erros são não correlacionados.
 $\rho(\epsilon_i, \epsilon_j) = 0$, para $i \neq j$
- 3 Os erros são não correlacionado com a variável dependente.
 $\rho(\epsilon_i, y) = 0$

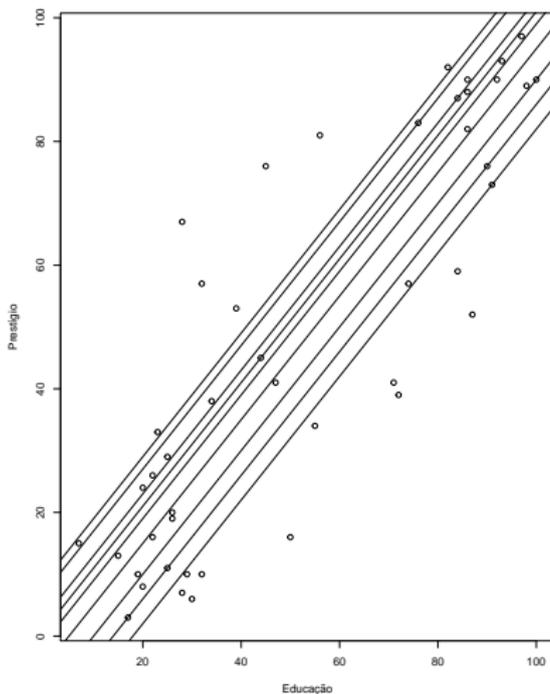
Obs: y é uma variável aleatória porém x não é uma variável aleatória.

Mínimos Quadrados

Determina os coeficientes (β_0, β_1) os quais minimizam a soma dos erros ao quadrado. Ou seja,

$$\min \sum_{i=1}^n \epsilon_i^2 = \min \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x)^2$$

O Método de mínimos quadrados irá garantir que após estimar os coeficiente teremos apenas uma reta passando entre os dados.



Equações Normais

$$\begin{cases} \frac{\partial \sum_{i=1}^n \epsilon_i^2}{\partial \beta_0} = 0 \\ \frac{\partial \sum_{i=1}^n \epsilon_i^2}{\partial \beta_1} = 0 \end{cases}$$

Estimadores de β_0 e β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

e $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Onde \bar{x} é a média das observações e \bar{y} é a média das respostas

Equação final após a estimação dos coeficientes via MQ

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Observe que a equação não tem mais a componente do erro..

Exemplo: Considere os dados de prestígio e educação

```
lm.pres <- lm(formula = pres$prestigio ~ pres$educacao)
```

Coefficients:

	Estimate	Std. Error	t value	value
(Intercept)	0.284000	5.09306	0.056	0.956
pres\$educacao	0.90200	0.08455	10.668	1.17e-13

Residual standard error: 16.69 on 43 degrees of freedom, Multiple R-squared: 0.7258, Adjusted R-squared: 0.7194

Equação de predição ou previsão

$$\hat{y}(\text{prestigio}) = 0.284 + 0.902 * x(\text{anos de estudo})$$

Significa que para uma unidade de x equivale a um acréscimo 0.902 na resposta y . Ou seja, para cada ano de estudo a mais o indivíduo tem um crescimento de aproximadamente 1% no seu prestígio.

Resíduos

$$\hat{\epsilon} = y - \hat{y}$$

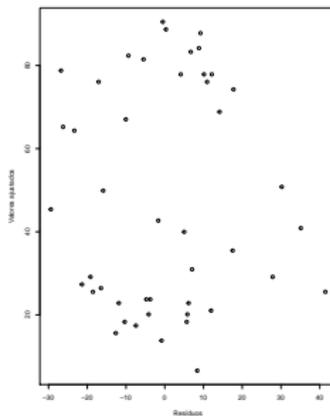
Propriedades

$$\sum_{i=1}^n \hat{\epsilon}_i = 0$$

$$\sum_{i=1}^n x_i \hat{\epsilon}_i = 0$$

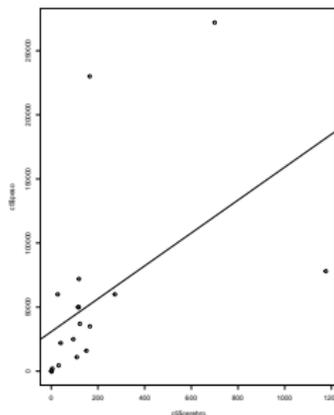
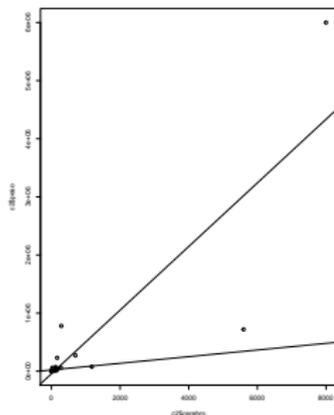
Valores ajustados e resíduos, \hat{y}

```
y_est <- fitted.values(lm.pres)
res <- residuals(lm.pres)
plot(res,est)
identify(res, est, labels=pres$profissao)
```



Pontos Extremos

Observações que assumem valores extremos comparado com as demais são denominados "outliers".



Critérios de Ajustamento

R^2 representa a proporção da variação explicada pelo modelo de regressão com relação à variação total, denominado Coeficiente de determinação.

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Obs: R^2 está definido no intervalo $[0,1]$

Erro Padrão da Regressão

$$s_p = \sqrt{\frac{1}{n-2} \sum_{i=1}^n \epsilon_i^2}$$

Regressão e Correlação

$$\hat{\beta} = r \frac{s_y}{s_x}$$

Exemplo:

Considere um conjunto de dados hipotéticos de 19 empresas que tem registrado os seus gastos com propaganda (em dólares) os valores das médias das vendas mensais (em dólares) de cada empresa.

- 1 Estimar β_0 e β_1 .
- 2 Montar a equação da regressão e interpretar os β s.
- 3 Fazer um scatterplot dos gastos versus vendas
- 4 Calcular \hat{y} considera um gasto com propaganda de 4 dolares, $x = 4$.
- 5 Calcule o coeficiente de determinação, o coeficiente de correlação e o erro da regressão.
- 6 Testar se os coeficientes da regressão são nulos.
- 7 Extrair os resíduos e os valores ajustados.
- 8 Fazer um gráfico dos gastos versus resíduos.
- 9 Fazer um histograma dos resíduos.
- 10 Fazer um scatterplot dos resíduos versus normal.

Residuals:

Min	1Q	Median	3Q	Max
-34.931	-16.983	-4.301	17.384	35.804

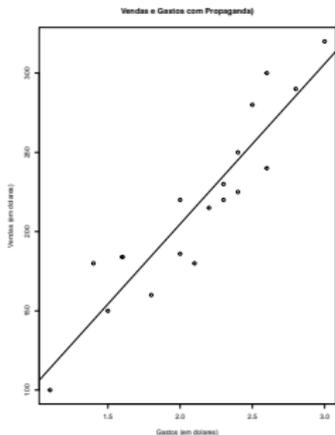
Coefficients:

	Estimate	Std. Error	t value	value
(Intercept)	2.727	21.954	0.056	0.956
gastos	101.049	10.102	10.003	1.54e-08

Residual standard error: 21.87 on 17 degrees of freedom Multiple R-squared: 0.8548, Adjusted R-squared: 0.8462
F-statistic: 100.1 on 1 and 17 DF, p-value: 1.541e-08

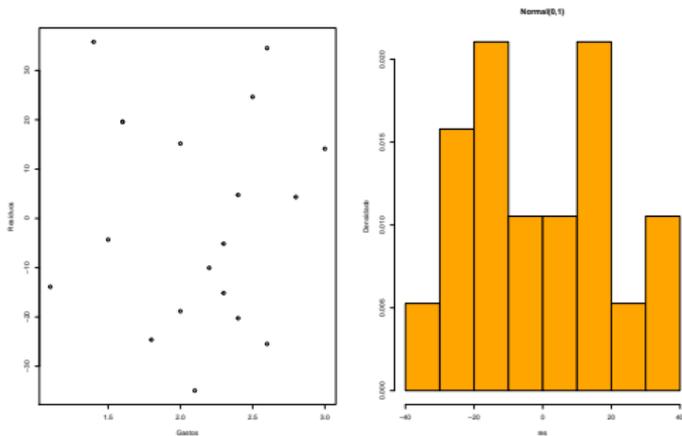
Analisando o output do R

- $\beta_0 = 2.727$ e $\beta_1 = 101.049$
- $\hat{y} = 2.727 + 101.049x$

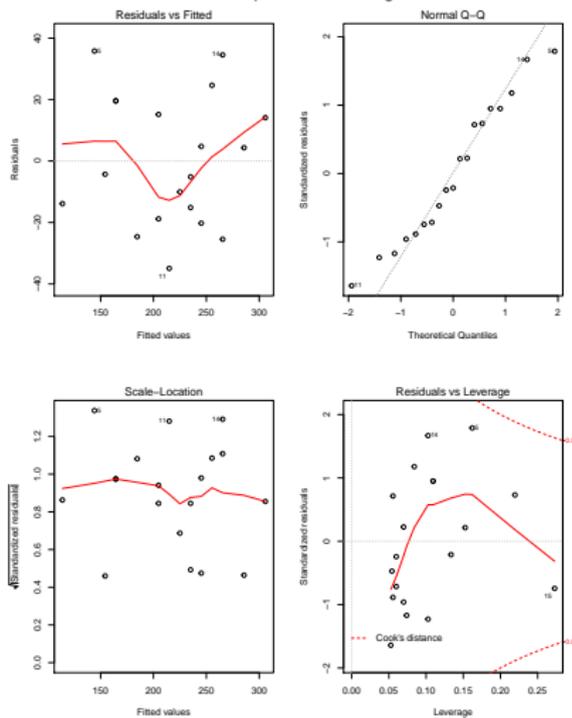


- $\hat{y} = 2.727 + 101.049 * 4 = 406.923$ Ou seja, se aumentar 4 dólares no gastos com propaganda terá um retorno de 406 dólares nas vendas.

- `aju <- fitted.values(fit)`
- `res <- residuals(fit)`



Análise de resíduos para modelo de regressão linear



Script para Análise de Regressão Linear

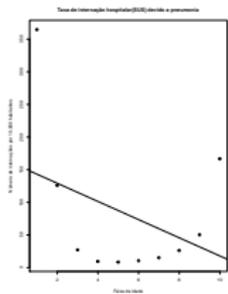
```
. setwd("/Curso de Estatística no R")
v <- read.csv("vendas.csv", sep=";", dec=".", header=TRUE)
head(vendas)
vendas <- v$vendas
gastos <- v$gastos
plot(gastos, vendas)
cor(gastos, vendas)
plot(gastos, vendas, xlab="Gastos (em milhares de dolares)", ylab="Vendas (em milhares de dolares)",
main="Vendas e Gastos com Propaganda")
fit <- lm(vendas ~ gastos)
summary(fit)
aju <- fitted.values(fit)
res <- residuals(fit)
m <- mean(vendas)
a <- (aju-m):2
b <- (vendas-m):2
R2 <- sum(a)/sum(b)
R2
plot(gastos, res, ylab="Resíduos", xlab="Gastos")
hist(res, probability=TRUE, col="orange", main="Normal(0,1)", ylab="Densidade")
r <- res:2
Sp <- sqrt(sum(r)/(19-2))
Sp
par(mfrow=c(2,2))
plot(fit)
mtext("Análise de resíduos para modelo de regressão linear",
outer=TRUE, line=-2, cex=1.4)
layout(1)
```

Exemplo: Taxa de internações devido a pneumonia por faixa de idade - DataSus

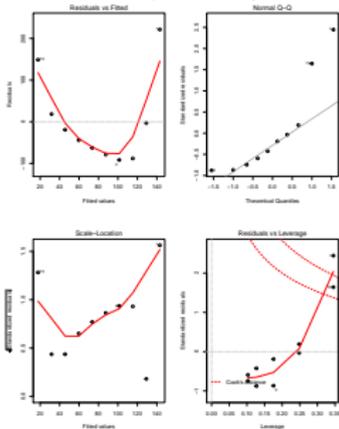
```
h <- read.csv("hospital.csv", sep=";", dec=".", header=TRUE)
h
```

idade	faixa	taxa
Menor 1 ano	1	364.58
1 a 4 anos	2	126.02
5 a 9 anos	3	27.19
10 a 19 anos	4	9.68
20 a 29 anos	5	8.52
30 a 39 anos	6	10.68
40 a 49 anos	7	15.28
50 a 59 anos	8	26.30
60 a 69 anos	9	50.34
70 anos e mais	10	166.65

```
plot(h$faixa,h$taxa, main="Taxa de internação hospitalar(SUS) devido a pneumonia", xlab="Faixa de idade",ylab="Número de internações por 10.000 habitantes")
fit <- lm(h$taxa ~ h$faixa)
abline(fit)
```



Análise de resíduos para modelo de internação linear



Exemplo: Taxa de mortalidade devido a neoplasias malignas

Modelo de regressão linear para gerar previsões em séries temporais.

```
n <- read.csv("neoplasia.csv", sep=";", dec=".",header=TRUE)
ano <- n$Ano
neo <- n$neoplasias_malignas
install.packages("forecast")
neo_ts <- ts(neo, start=c(1998,1), end=c(2012,1), frequency=1)
fit <- tslm(neo_ts ~ trend)
summary(fit)
```

Call:

tslm(formula = neo_ts ~ trend)

Residuals:

Min	1Q	Median	3Q	Max
-2.3849	-1.6397	-0.7218	1.6010	3.0999

Coefficients:

	Estimate	Std. Error	t value	value
(Intercept)	16.2163	1.0492	15.456	9.54e-10 ***
ano	0.5690	0.1154	4.931	0.000275

Residual standard error: 16.69 on 43 degrees of freedom, Multiple R-squared: 0.7258, Adjusted R-squared: 0.7194

Obs: A taxa de mortalidade de devido a neoplasias malignas aumenta em 0.569 a cada ano.

Equação do modelo

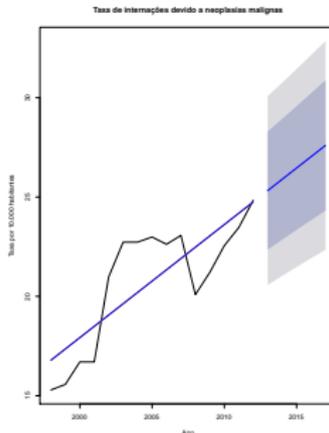
$$y_t = \beta_0 + \beta_1 t + \epsilon_t$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 t$$

$$\text{neoplasia} = 16.2163 + 0.5690 \text{Ano}$$

O Modelo de Previsão

```
f <- forecast(fit, h=5, level=c(80,95))  
plot(f, ylab="Taxa por 10.000 habitantes", main="Taxa de internações devido a neoplasias malignas", xlab="Ano")  
lines(fitted(f), col="blue")
```



Inferência da Regressão

$E(\hat{\beta}_0) = \beta_0$, onde $\hat{\beta}_0$ é um estimador não viesado de β_0

$$V(\hat{\beta}_0) = \frac{\sigma_\epsilon^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}$$

$E(\hat{\beta}_1) = \beta_1$, onde $\hat{\beta}_1$ é um estimador não viesado de β_1

$$V(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$\hat{\beta}_0$ tem distribuição $N(\beta_0, V(\hat{\beta}_0))$

$\hat{\beta}_1$ tem distribuição $N(\beta_1, V(\hat{\beta}_1))$

Intervalo de confiança

Um intervalo de confiança para β com $100(1 - \alpha)\%$ será:

$$(\hat{\beta} \pm t_{\frac{\alpha}{2}} SE(\hat{\beta}))$$

Onde, $SE(\hat{\beta}) = \sqrt{V(\hat{\beta})}$ confint(fit)

2.5% 97.5%

(Intercept) -9.9871443 10.555143

edu 0.7314745 1.072517

Análise de Variância

Fonte	Soma de Quadrados	GL	Quadrados Médios	Estatística
Regressão	$SSR = \sum_{i=1}^n (y_i - \hat{y})^2$	1	$SSRM = \frac{SSR}{1}$	$F = \frac{SSRM}{S_E}$
Resíduos	$SSE = \sum_{i=1}^n \epsilon^2$	n-2	$S_E = \frac{\sum_{i=1}^n \epsilon^2}{n-2}$	
Total	$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$	n-1		

Exemplo: Regressão do prestígio

```
summary(fit)
```

```
anova(fit)
```

Analysis of Variance Table

Response: pre					
Fonte	Soma de Quadrados	Gl	Quadrados médios	F	p-valor
educacao	31707	1	31707	113.8	1.17e-13***
residuals	11981	43	279		
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Exemplo:

Considere os dados de internamentos por neoplásias malignas do DataSus.

- 1 Estimar β_0 e β_1 .
- 2 Montar a equação da regressão e interpretar os β s.
- 3 Fazer um scatterplot dos gastos versus vendas
- 4 Calcular \hat{y} considera um gasto com propaganda de 4 dolares, $x = 4$.
- 5 Calcule o coeficiente de determinação, o coeficiente de correlação e o erro da regressão.
- 6 Testar os β s.
- 7 Extrair os resíduos e os valores ajustados.
- 8 Fazer um gráfico dos gastos versus resíduos.
- 9 Fazer um histograma dos resíduos.
- 10 Fazer um scatterplot dos resíduos versus normal.
- 11 Obter os intervalos de confiança de 80 e 90 por cento para os coeficientes do modelo.
- 12 Análisar a tabela da ANOVA.
- 13 Conclusões

Regressão Múltipla

A fórmula geral do modelo de regressão múltipla com k variáveis independentes é dada por:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_k x_{k,i} + \epsilon_i$$

Onde:

- 1 y_i a variável a ser predita
- 2 $x_{1,i}, \dots, x_{k,i}$ são as k variáveis preditoras
- 3 β_1, \dots, β_k são os coeficientes que medem o efeito marginal de cada preditor
- 4 ϵ_i os erros associados

Presupostos básicos sobre os erros ϵ_i

- 1 Os erros tem média zero
- 2 Os ϵ_i e ϵ_j são não correlacionados para todo $i \neq j$
- 3 Os ϵ_i são não correlacionados com os $x_{j,i}$
- 4 Os ϵ_i tem distribuição normal
- 5 A $\text{var}(\epsilon_i)$ é constante

Exemplo - Carteira de crédito bancária

O arquivo credito.csv contém uma amostra de 500 clientes de um banco australiano com as seguintes variáveis:

- Score de crédito do cliente
- Poupança
- Renda
- FTE - tempo de dedicação na empresa
- Solteiro
- Tempo que reside no domicílio
- Tempo de permanência no emprego

Obs: Onde a variável score é apresentada numa escala entre 0 e 100

O objetivo é prever o valor do score bancário utilizando diversas outras variáveis, caso tipo de cross-sectional data.

Lendo os dados

```
credito <- read.csv("credito.csv", sep=";", dec=".",header=TRUE)
head(credito)
```

escore	poupa	renda	fte	solteiro	reside	tempo
39.39981	0.012	111.168	TRUE	FALSE	27	8
51.79090	0.654	56.400	TRUE	FALSE	29	33

Define a função panel.hist

```
panel.hist <- function(x, ...)
usr <- par("usr"); on.exit(par(usr))
par(usr = c(usr[1:2], 0, 1.5) )
h <- hist(x, plot = FALSE)
breaks <- h$breaks; nB <- -length(breaks)
y <- h$counts; y <- -y/max(y)
rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
```

Comando para fazer o painel de histogramas das variáveis, onde teremos na diagonal do painel os respectivos histogramas e fora da diagonal os scatterplots das variáveis

```
pairs(credito, diag.panel = panel.hist)
```

Obs: Este painel contém variáveis categóricas as quais devem ser retiradas, pois não faz sentido construir histograma de categorias.

Modificando o dataframe

```
escore <- credito$escore
poupa <- credito$poupa
renda <- credito$renda
fte <- credito$fte
solteiro <- credito$solteiro
reside <- credito$reside
tempo <- credito$tempo
c <- cbind(escore,poupa,renda,reside,tempo)
head(c)
```

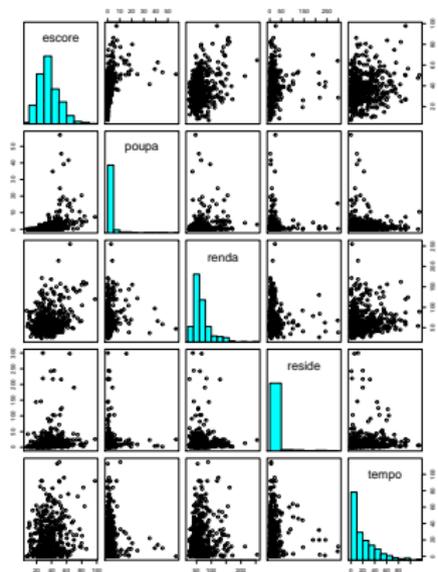
escore	poupa	renda	reside	tempo
39.39881	0.012	111.168	27	8
51.79090	0.654	56.400	29	33

Nomes das colunas

```
col <- c("escore","poupa","renda","reside","tempo")
dimnames(c) <- list(NULL,col)
c <- as.data.frame(c,rownames=NULL)
head(c)
```

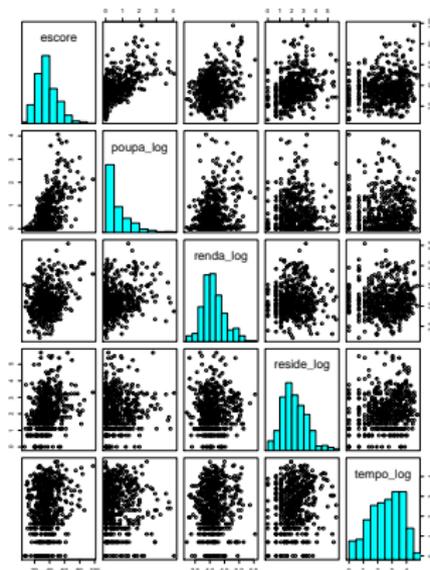
Novo painel

```
pairs(c, diag.panel = panel.hist)
```



Aplicando o log as variáveis independentes

```
poupa_log <- log(poupa + 1)
renda_log <- log(renda + 1)
reside_log <- log(reside + 1)
tempo_log <- log(tempo + 1)
clog <- cbind(escore,poupa_log,renda_log,reside_log,tempo_log)
clog <- as.data.frame(clog,rownames=NULL)
pairs(clog, diag.panel = panel.hist)
```



Estimação do modelo - Consiste em encontrarmos valores dos coeficientes os quais minimizam a soma de quadrados dos erros

Método dos mínimos quadrados - MQ

$$\sum_{i=1}^n \epsilon^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{1,i}, \dots, \beta_{k,i} x_{k,i})^2$$

Os preditores da variável resposta y podem ser obtido através da seguinte expressão:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_{k,i} x_{k,i}$$

Obs: a equação acima não contém o termo dos erros

Estimação dos erros

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1,i} - \dots - \hat{\beta}_{k,i} x_{k,i}$$

Coeficiente de determinação: $\frac{\text{Variação explicada pelo modelo estimado}}{\text{Variação total}}$

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Estimando o modelo: comando $\text{lm}(y \sim x)$

```
aju <- lm(formula = escore ~ poupa_log + renda_log + reside_log + tempo_log)
summary(aju)
```

Residuals:

Min	1Q	Median	3Q	Max
-26.133	-6.966	-1.125	5.379	37.446

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
Intercept	-0.2186	5.2309	-0.042	0.96668
poupa_log	10.3526	0.6124	16.904	2e-16***
renda_log	5.0521	1.2579	4.016	8.83e-05***
reside_log	2.6666	0.4345	6.137	1.72e-09***
tempo_log	1.3138	0.4094	3.209	0.00142**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

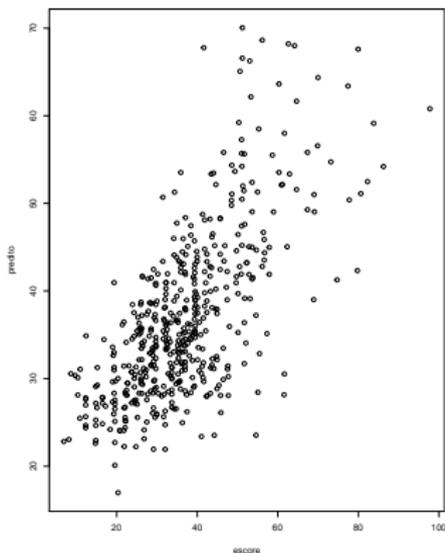
Residual standard error: 10.16 on 495 degrees of freedom

Multiple R-squared: 0.4701, Adjusted R-squared: 0.4658

F-statistic: 109.8 on 4 and 495 DF, p-value: < 2.2e-16

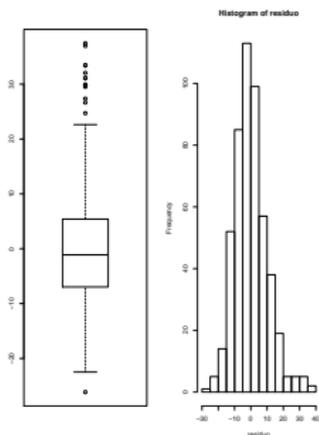
Valores Preditos(\hat{y}) versus Valores reais(y)

Se as variáveis independentes explicam bem a variação existente na variável dependente, espera-se que os valores ajustados de y estejam muito próximo dos valores observados dos escores dos 500 clientes da carteira de crédito.



Análises dos Resíduos

Após o ajustamento espera-se que a distribuição dos resíduos seja simétrica e aproximadamente normal com média zero e variância um. O seja, o seu modelo foi capaz de extrair toda informação contida na amostra e o que restou foi apenas um mero resíduo desprezível, caso contrário teremos que revrmos o conjunto de variáveis independentes, ou a formulação matemática do modelo.

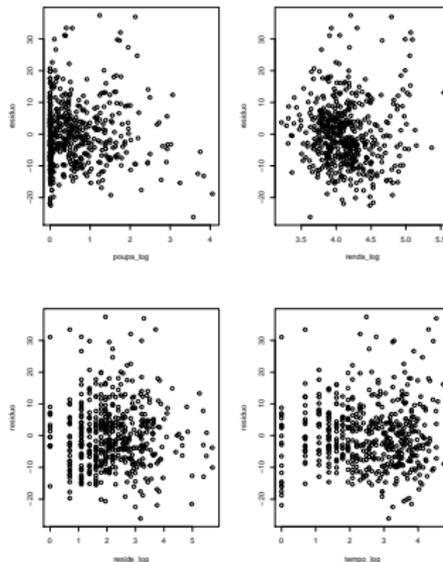


Gráficos dos Resíduos versus cada variável independente

```

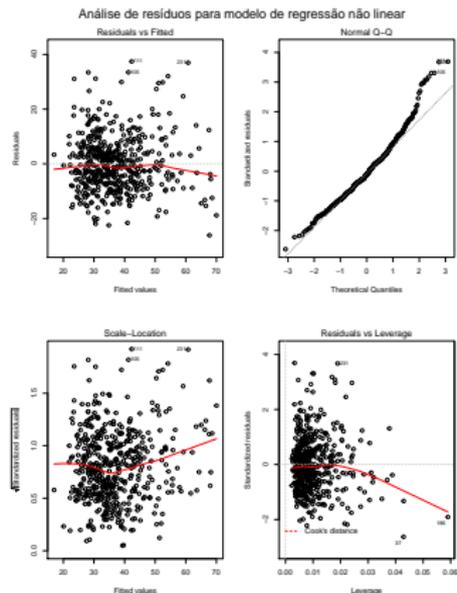
escore_fit <- fitted.values(aju) plot(escore_fit,escore) residuo <- residuals(aju) par(mfrow=c(2,2))
plot(poupa_log,residuo) plot(renda_log,residuo) plot(reside_log,residuo) plot(tempo_log,residuo) boxplot(residuo)
hist(residuo)

```



Comando `plot(aju)`: painel de gráficos dos elementos contidos no objeto `aju` após o ajustamento

```
par(mfrow=c(2,2))
plot(aju)
mtext("Análise de resíduos para modelo de regressão não linear",
      outer=TRUE, line=-2, cex=1.4)
layout(1)
```



R^2 Ajustado

O R^2 não representa uma boa medida quando estamos analisando o modelo com respeito a predição, pois tende a convergir para 1 quando aumentamos o número de variáveis independentes. O R^2 ajustado, corrigido, impõe uma penalidade a medida que incluímos novas variáveis no modelo, representado assim um bom critério para seleção de variáveis a serem incluídas no modelo

$$R^2_{aju} = 1 - (1 - R^2) \frac{n-1}{n-k-1}$$

AIC - Critério de Informação de Akaike

$$AIC = n \log\left(\frac{SSE}{n}\right) + 2(k + 2)$$

- n representa o número de observações
- k o número de variáveis independentes
- $SSE = \sum_{i=1}^n \epsilon^2$

BIC - Critério de Informação de Bayseano de Scharz

$$BIC = n \log\left(\frac{SSE}{n}\right) + (k + 2) \log(n)$$

- O BIC difere do AIC apenas pelo termo da penalidade do critério, o qual é multiplicado por 2 e o outro por $\log(n)$ respectivamente.

Exemplo: considere os dados da carteira de crédito de um banco australiano, obter os estimadores das seguinte regressões:

1 $y = \beta_0 + \beta_1 \text{poupa_log} + \beta_2 \text{renda_log} + \beta_3 \text{reside_log} + \beta_4 \text{tempo_log}$

2 $y = \beta_0 + \beta_1 \text{poupa_log} + \beta_2 \text{renda_log} + \beta_3 \text{reside_log}$

3 $y = \beta_0 + \beta_1 \text{poupa_log} + \beta_2 \text{reside_log} + \beta_3 \text{tempo_log}$

4 $y = \beta_0 + \beta_1 \text{poupa_log} + \beta_2 \text{reside_log}$

5 $y = \beta_0 + \beta_1 \text{poupa_log} + \beta_2 \text{renda_log} + \beta_3 \text{tempo_log}$

6 $y = \beta_0 + \beta_1 \text{poupa_log} + \beta_2 \text{tempo_log}$

7 $y = \beta_0 + \beta_1 \text{poupa_log} + \beta_2 \text{renda_log}$

8 $y = \beta_0 + \beta_1 \text{poupa_log}$

9 $y = \beta_0 + \beta_1 \text{renda_log} + \beta_2 \text{reside_log} + \beta_3 \text{tempo_log}$

10 $y = \beta_0 + \beta_1 \text{renda_log} + \beta_2 \text{reside_log}$

11 $y = \beta_0 + \beta_1 \text{renda_log} + \beta_2 \text{tempo_log}$

12 $y = \beta_0 + \beta_1 \text{reside_log} + \beta_2 \text{tempo_log}$

13 $y = \beta_0 + \beta_1 \text{reside_log}$

14 $y = \beta_0 + \beta_1 \text{renda_log}$

15 $y = \beta_0 + \beta_1 \text{tempo_log}$

16 $y = \beta_0$

Seleção do Melhor Modelo

Modelo	k	$2(k+2)$	$(k+2)\log(n)$	R2 ajustado	AIC	BIC
1	4	12	37,3	0,46	2325,8	2351,1
2	3	10	31,1	0,45	2334,1	2355,1
3	3	10	31,1	0,45	2339,8	2360,9
4	2	8	24,9	0,44	2349,2	2366,1
5	3	10	31,1	0,43	2360,4	2381,5
6	2	8	24,9	0,41	2373,4	2390,3
7	2	8	24,9	0,40	2377,7	2394,6
8	1	6	18,6	0,39	2392,1	2404,7
9	3	10	31,1	0,16	2551,6	2572,7
10	2	8	24,9	0,15	2553,8	2570,7
11	2	8	24,9	0,10	2586,7	2603,5
12	2	8	24,9	0,09	2591,4	2608,2
13	1	6	18,6	0,08	2594,6	2607,2
14	1	6	18,6	0,08	2595,3	2607,9
15	1	6	18,6	0,10	2584,7	2597,3
16	0	0	0,00	0,00	2641,3	2662,4

Exercício - I

Acessando elementos na matriz

Considere a matriz `região.dados`, use os comandos R abaixo e identifique os elementos

- `região.dados[1,]`
- `região.dados[1,]`
- `região.dados[1,]`
- `região.dados[1,]`
- `região.dados[,1]`
- `região.dados[,2]`
- `região.dados[,3]`
- `região.dados[,4]`
- `região.dados[,5]`

Exercício - II

- 1 Considere $x=12$ e $y=23$. Use o R e obtenha:
- a $x + y, \sqrt{x}$ e \sqrt{y}
 - b $3(x + y), \frac{x}{3}, \frac{y}{2}$ e xy
 - c $z = x^2 + y^2$ e \sqrt{z}
 - d $(x + y)^2$
- 2 Considere dois vetores $\text{vet1}=(8,10,7,3,2,15,20)$ e $\text{vet2}=(1,1,1,1,1,1,1)$. Calcule:
- a A soma dos dois vetores
 - b A soma acumulada de vet1 e vet2
 - c O vetor $\text{vet3} = \text{vet1} - \text{vet2}$
 - d A média e mediana de vet1 , vet2 e vet3
 - e Os o sumário de vet1 , vet2 e vet3
 - f Os percentis, 20, 25, 50, 70 e 75 de vet1 , vet2 e vet3
 - g As variâncias de vet1 , vet2 e vet3 para n e $n-1$.
- 3 Considere o vetor $\text{vet4} = \text{vet1} - \text{vet1}_{(1)} - \text{vet1}_{(6)} - \text{vet1}_{(7)}$. Calcule a média, mediana, quartil1, quartil3, o intervalo interquartilico e a variância de vet4 . Compare com os resultados de vet1 .

Exercício - III

Os dados abaixo representam as temperaturas (em graus F) dos anéis de combustível de espaçonaves que foram submetidos a testes de temperatura.

(84, 49, 61, 40, 83, 67, 45, 66, 70, 69, 80, 58, 68, 60, 67, 72, 73, 70, 57, 63, 70, 78, 52, 67, 53, 67, 75, 61, 70, 81, 76, 79, 75, 76, 58, 31)

- a Calcule a média e mediana das temperaturas
- b Encontre os quartis
- c Encontre os percentis 5 e 9
- d Retire da amostra o menor valor e recalcule os itens anteriores
- e Comente os resultados obtidos

Exercício - IV

- 2 Uma amostra de seis resistores fornece as seguintes resistências em ohms:
 $x_1 = 45, x_2 = 38, x_3 = 47, x_4 = 41, x_5 = 35, x_6 = 43.$
- a Calcule a média e mediana
 - b Encontre os quartis
 - c Encontre os percentis 5 e 9
 - d Retire da amostra o menor valor e recalcule os itens anteriores. Comente os resultados obtidos.
 - e Se as temperaturas fossem 450, 380, 470, 410, 350 e 430 como vc poderia comparar essas duas amostras?
 - f Construa um boxplot

Exercício - V

- 1 A amostra abaixo representa o grau de concentração de oxigênio na fabricação de circuitos integrados. Essa concentração avalia o nível de contaminação no silicone desses circuitos.
(3.15, 2.68, 4.31, 2.09, 3.82, 2.94, 3.47, 3.39, 2.81, 3.61).
- a Calcule a variância e o desvio padrão amostral usando os comandos R.
 - b Calcule a variância e o desvio padrão usando a definição.
 - c Subtraia 35 de cada elemento da amostra e calcule s^2 e s , compare com os resultados anteriores.
 - d Construa um boxplot

Enivaldo Rocha

Exercícios VI - Correlação



Apêndice I

Tabela: Funções matemáticas do R

Função	Descrição
<code>sqrt()</code>	raiz quadrada
<code>abs()</code>	valor absoluto
<code>sin()</code> <code>cos()</code> <code>tan()</code>	funções trigonométricas
<code>asin()</code> <code>acos()</code> <code>atan()</code>	funções trigonométricas inversas
<code>sinh()</code> <code>cosh()</code> <code>tanh()</code>	funções hiperbólicas
<code>asinh()</code> <code>acosh()</code> <code>atanh()</code>	funções hiperbólicas inversas
<code>exp()</code> <code>log()</code>	exponencial e logaritmo natural
<code>log10()</code> <code>log2()</code>	logaritmo na base 10 e na base 2
<code>gamma()</code>	funções Gamma de Euler
<code>factorial</code>	fatorial (n!)
<code>choose()</code>	número de combinações
<code>combn()</code>	todos conjuntos gerados pela combinação de certo número de elementos

Obrigado!!!

-  Brian S. Everitt and Torsten Hothorn, (2014). A Handbook of Statistical Analyses Using R, Third Edition. Chapman Hall Book.
-  Bussab, W. de O. e Morettin, P. A. (2003). Estatística Básica, 5ª ed. São Paulo: Editora Saraiva.
-  George E. P. Box, Stuart Hunter, William G. Hunter, (2005). Statistics for Experimenters Design, Innovation, and Discovery. Second Edition. John Wiley Sons Inc. Publication.
-  John Fox, (2016). Applied Regression Analysis and Generalized Linear Models, Third Edition. SAGE.
-  John Fox, (2017). Interface for R Using the R Commander. SAGE.
-  Rob J Hyndman, George Athanasopoulos, (2014). Forecasting Principles and Practice. Texts.
-  The R Project for Statistical Computing (2006). www.r-project.org.