

8. Medidas de associação

Parte 1

2012

Dados: (x_i, y_i) , $i = 1, \dots, n$. n pares de observações das variáveis x e y , que podem ser qualitativas ou quantitativas.

Os pares representam a ocorrência simultânea de x e y . Cada par (x_i, y_i) é indissociável.

Em várias situações há interesse em estudar a relação entre x e y , se existir.

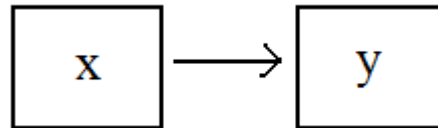
Uma possível relação entre x e y pode ser quantificada por uma medida **resumo**: medida de **associação**.

A associação entre variáveis pode ser negativa ou positiva (**sentido**). É fraca ou forte (**intensidade** ou **força**).

Gráficos são úteis no estudo de associação entre variáveis.

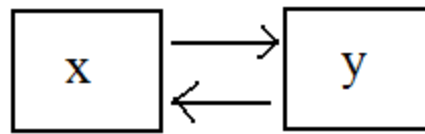
Tipos de relação

(a) **Causal unilateral**. y depende de x (ou x depende de y).



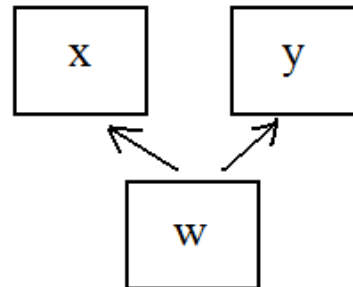
Exemplo. O preço de venda de um produto (y) depende da distância entre o local de produção e o local de venda (x). x depende de y ?

(b) **Causal bilateral** ou **interdependência**. y depende de x e x depende de y .



Exemplo. Relação entre peso (y) e altura (x) de uma pessoa.

(c) **Dependência indireta** (ou **associação espúria**). x e y são influenciadas por outra(s) variável(is).

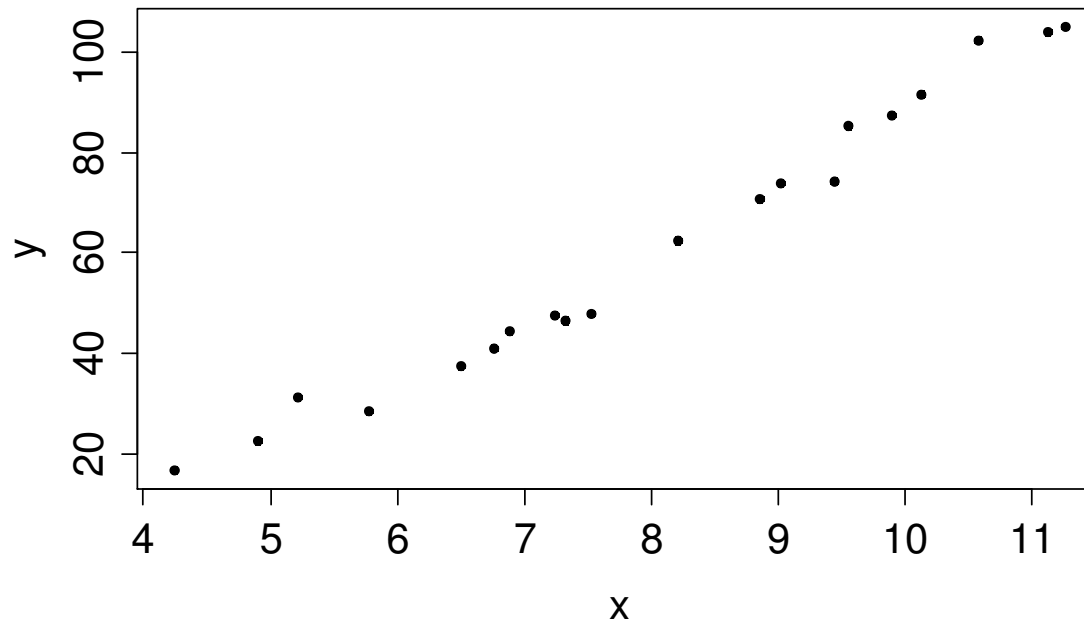


Exemplo. Relação entre o número anual de casos de insolação (x) e a produção anual de trigo (y). Causa comum: temperatura (w).

8.1. Variáveis quantitativas

$(x_1, y_1), \dots, (x_n, y_n)$: conjunto de dados **bivariado**.

Representação gráfica: **gráfico de dispersão** (*scatter plot*). Gráfico cartesiano dos pares (x_i, y_i) , $i = 1, \dots, n$.



Covariância entre x e y : medida da variação **conjunta** (ou concomitante ou simultânea) de x e y em relação às suas médias.

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad -\infty < \text{cov}(x, y) < \infty.$$

8.1. Variáveis quantitativas

Obs. (a) $\text{cov}(x, y) = \text{cov}(y, x)$ e (b) $\text{cov}(x, x) = s_x^2$.

Coeficiente de correlação linear de Pearson (r):

$$\text{cor}(x, y) = r = \frac{\text{cov}(x, y)}{s_x s_y} = \frac{1}{n-1} \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y},$$

sendo que s_x e s_y denotam os desvios padrão de x e y . Se $s_x = 0$ e/ou $s_y = 0$, r não está definido.

Propriedades: **P1.** $\text{cor}(x, x) = 1$. **P2.** $-1 \leq r \leq 1$.

P3. $r = 1$ se, e somente se, a relação entre x e y for **linear** ($y = a + bx$) e $b > 0$.

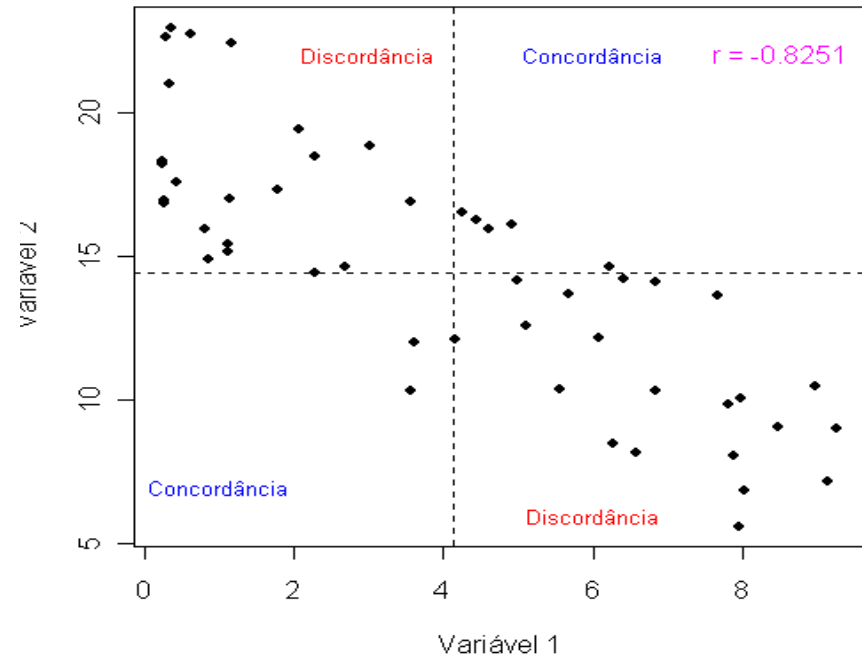
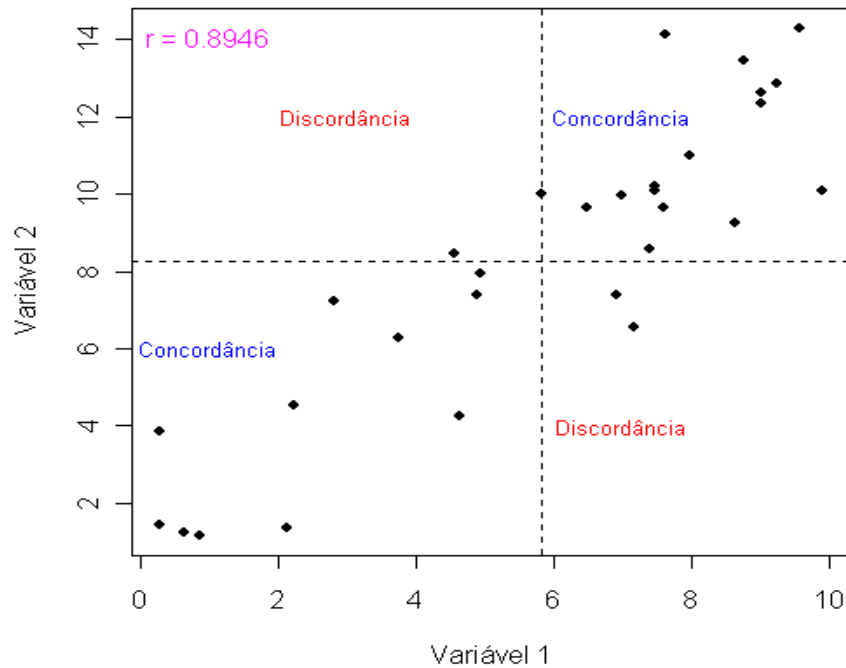
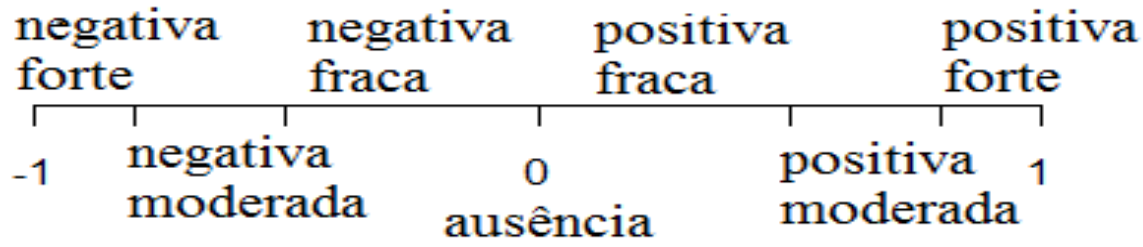
P4. $r = -1$ se, e somente se, a relação entre x e y for **linear** ($y = a + bx$) e $b < 0$.

P5. Invariância. Se $b_1 > 0$ e $b_2 > 0$, então $\text{cor}(x, y) = \text{cor}(a_1 + b_1 x, a_2 + b_2 y)$, em que a_1 e a_2 são reais quaisquer.

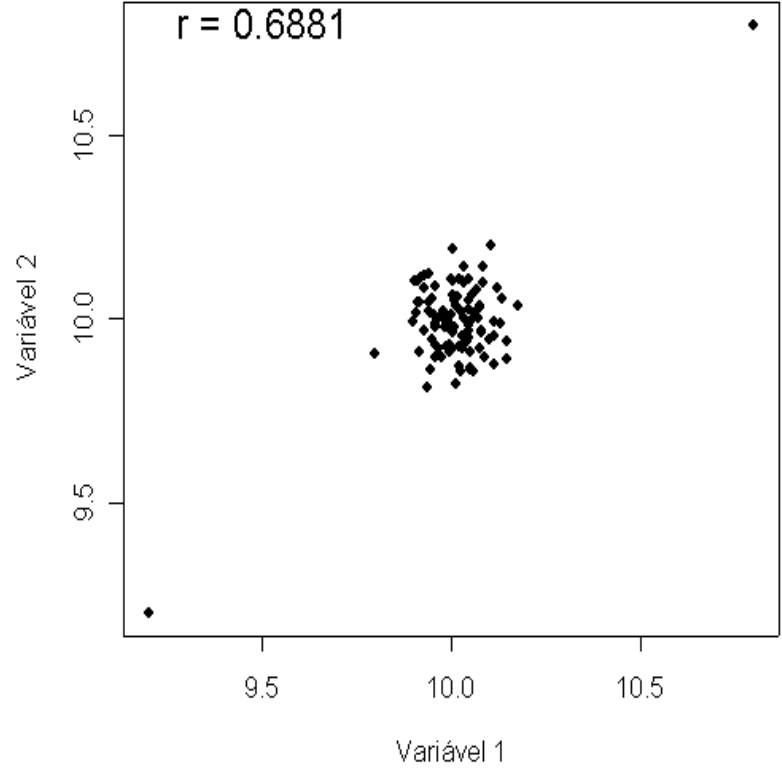
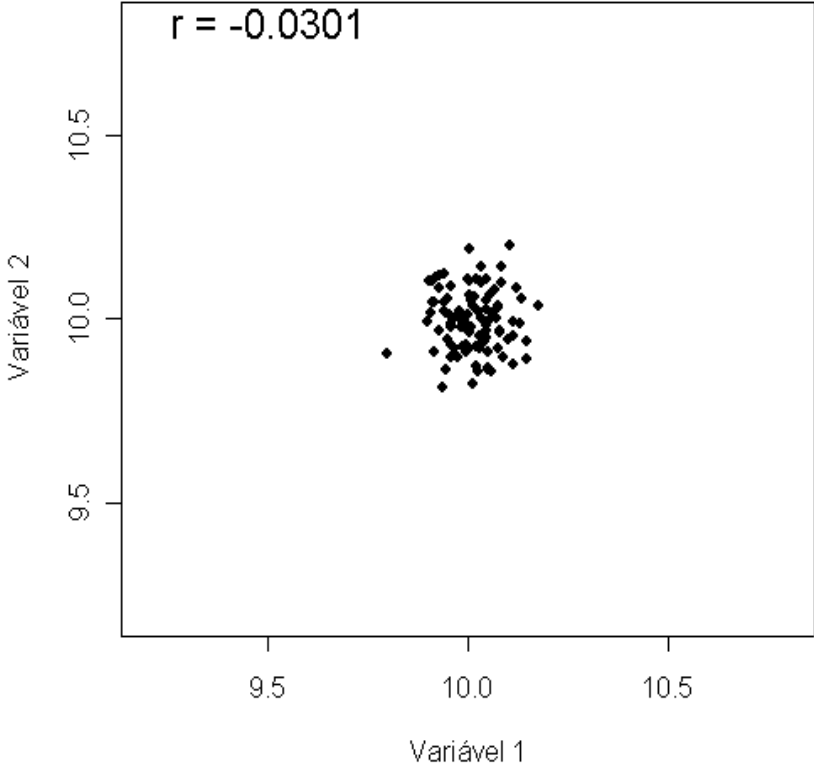
Exercício. Se $b_1 < 0$ e $b_2 > 0$ ou $b_1 > 0$ e $b_2 < 0$ ou $b_1 < 0$ e $b_2 < 0$, o que se pode afirmar sobre $\text{cor}(a_1 + b_1 x, a_2 + b_2 y)$?

8.1. Variáveis quantitativas

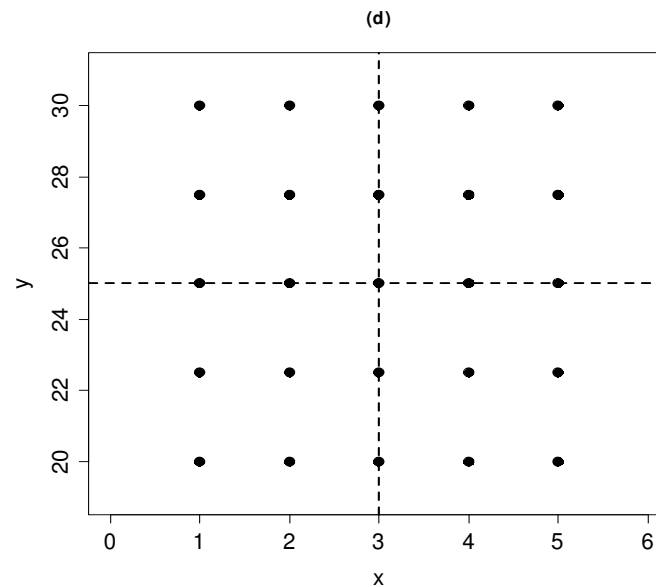
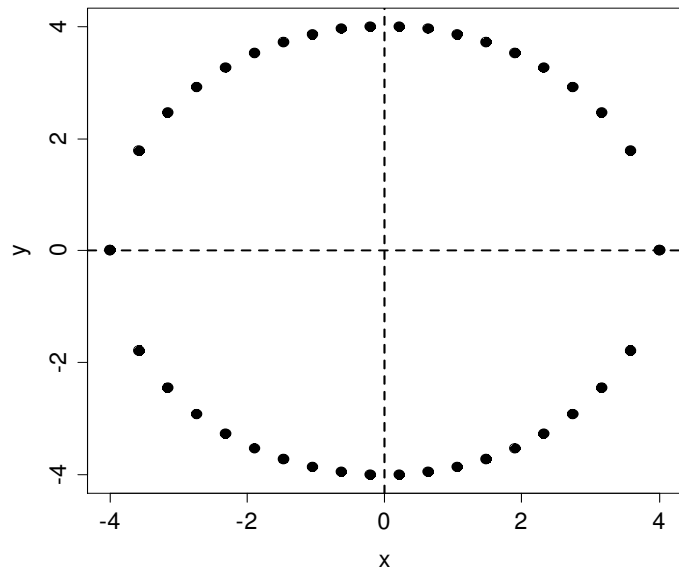
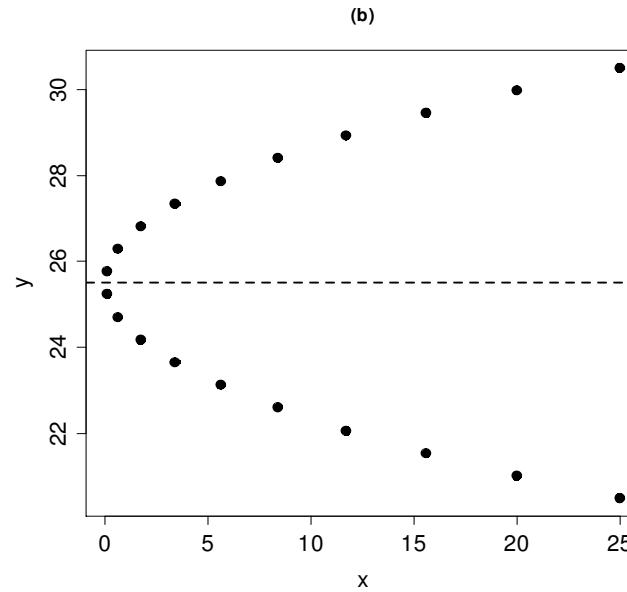
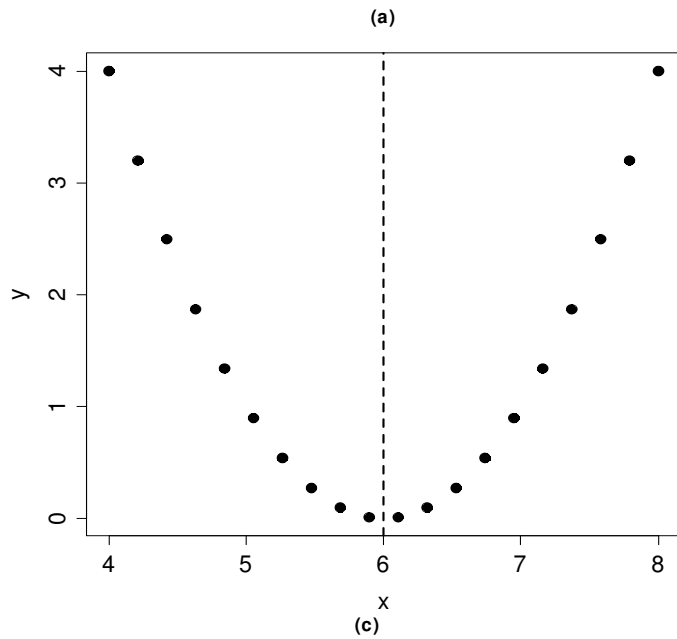
Sentido e força de r (correlação)



8.1. Variáveis quantitativas



8.1. Variáveis quantitativas

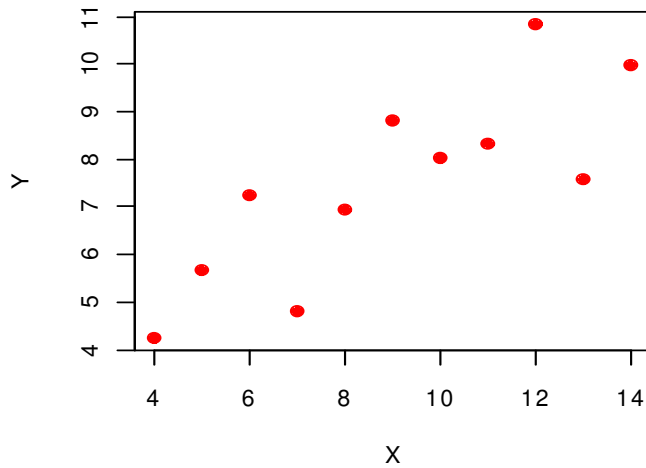


Exercício. Prove que se houver **simetria** em x e/ou y , então $r = 0$.

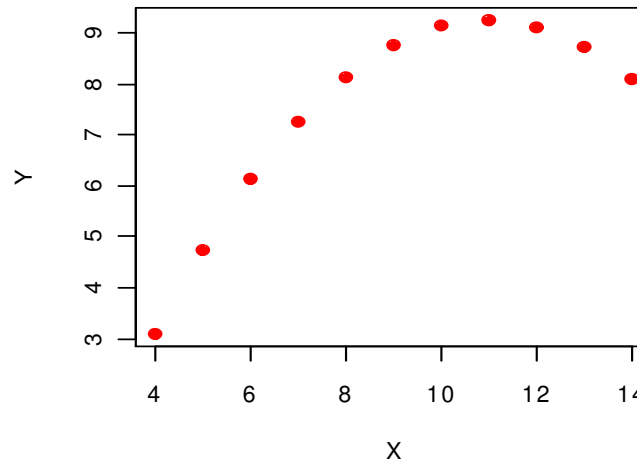
Obs. $r = 0$ não significa ausência de associação.

8.1. Variáveis quantitativas

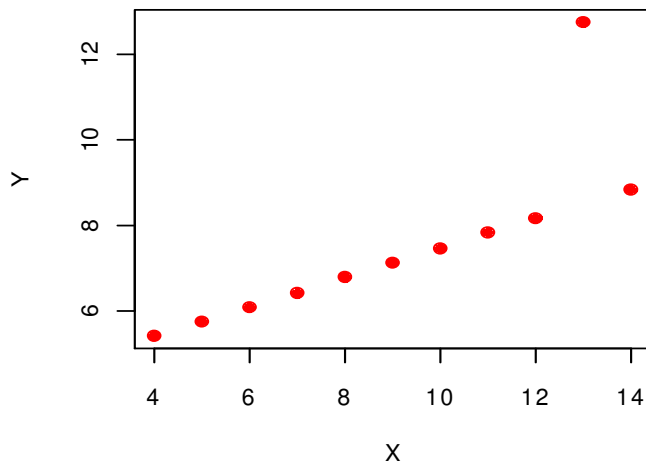
Exemplo 1



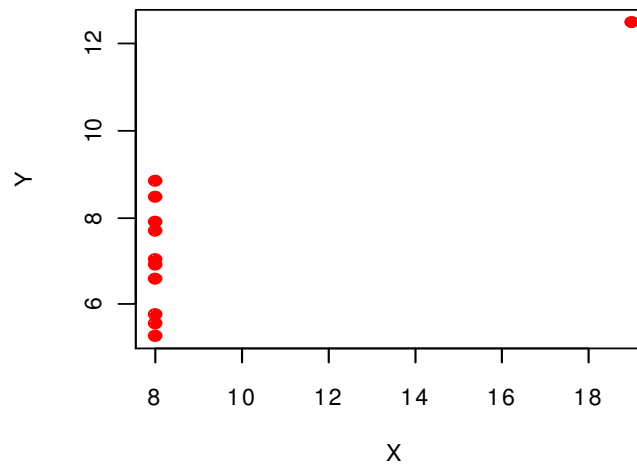
Exemplo 2



Exemplo 3



Exemplo 4



Dados anscombe em R

```
> ?anscombe
```

Valores de r:

Exemplo 1:
0,8164

Exemplo 2:
0,8162

Exemplo 3:
0,8163

Exemplo 4:
0,8165

Veja também <http://www.jerrydallal.com/LHSP/corr.htm>

Correlação em R

Funções `cor`, `cov` e `cov2cor`.

```
> x = c(5.5,6.7,9.5,4.2,9.0,11.6,4.5,9.6,6.2,11.6,8.8,8.6,7.8,4.8,  
10.1)
```

```
> y = c(11.6,11.3,17.5,9.1,15.7,16.9,8.1,21.2,11.7,18.7,13.9,15.0,  
11.6,7.0, 15.6)
```

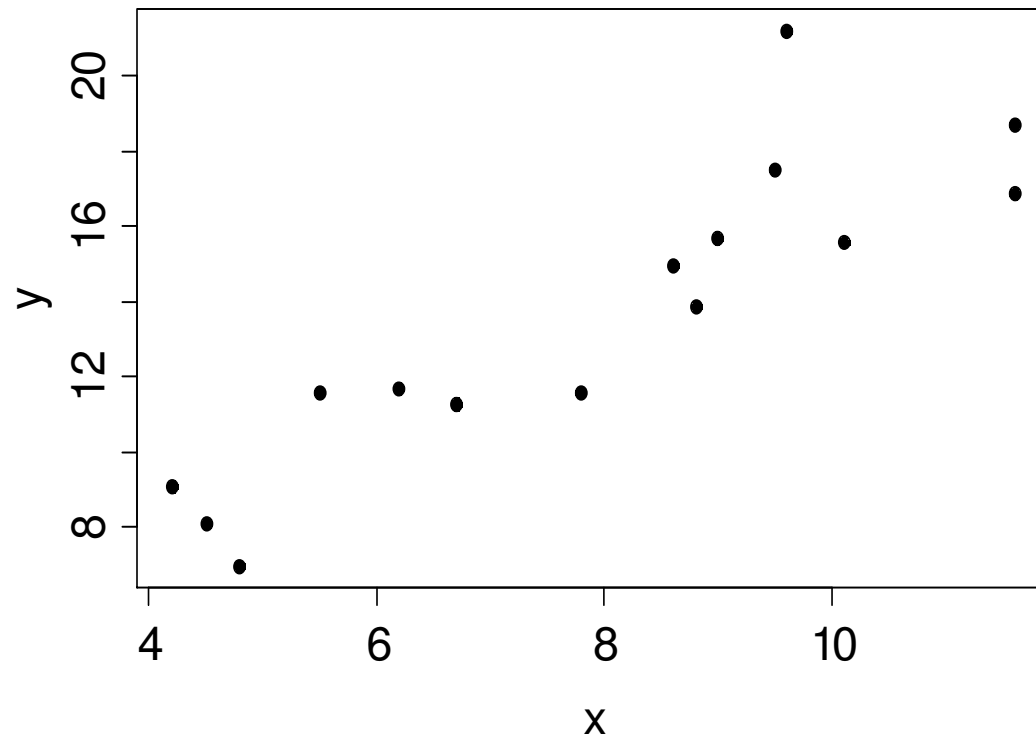
```
> length(x)
```

```
[1] 15
```

```
> cor(x, y)
```

```
[1] 0.8908723
```

```
> plot(x, y, pch = 20)
```



Correlação em R

```
> ? USArrests
```

Description

This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

Número de prisões por assalto, homicídio e estupro por 100 000 hab. e proporção da população urbana.

```
> names(USArrests)
```

```
[1] "Murder" "Assault" "UrbanPop" "Rape"
```

```
> rownames(USArrests)
```

```
[1] "Alabama" "Alaska" "Arizona" "Arkansas" "California" etc  
[50] "Wyoming"
```

```
> class(USArrests)
```

```
[1] "data.frame"
```

Classe “folha de dados”.

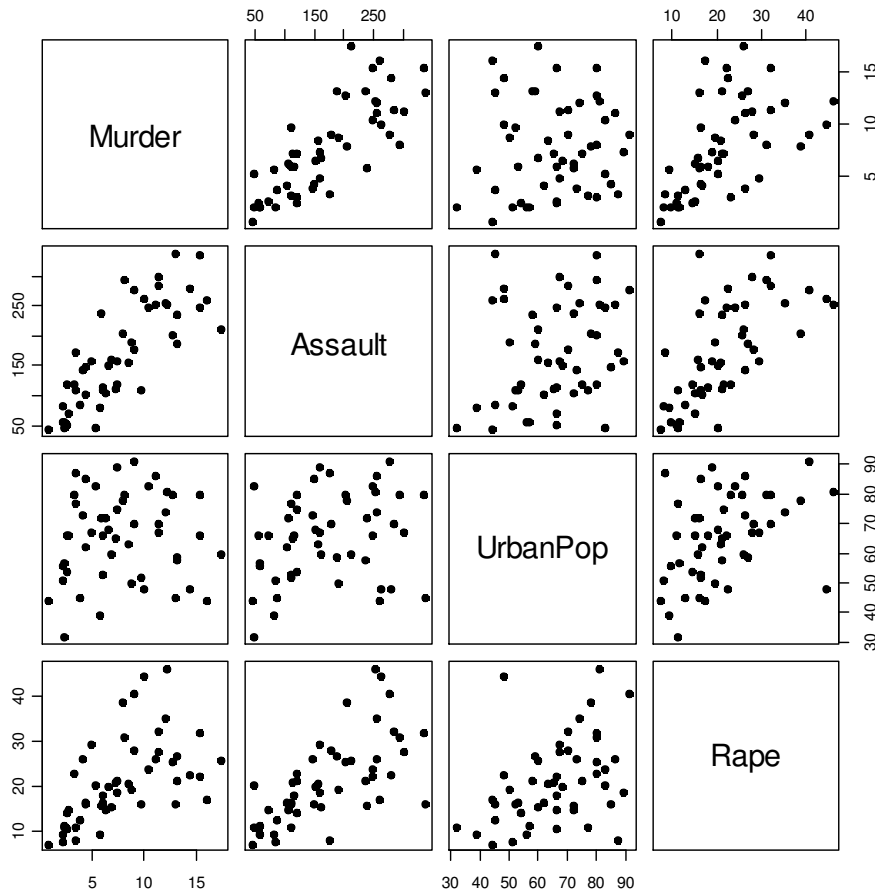
```
> summary(USArrests)
```

Murder	Assault	UrbanPop	Rape
Min. : 0.800	Min. : 45.0	Min. :32.00	Min. : 7.30
1st Qu.: 4.075	1st Qu.:109.0	1st Qu.:54.50	1st Qu.:15.07
Median : 7.250	Median :159.0	Median :66.00	Median :20.10
Mean : 7.788	Mean :170.8	Mean :65.54	Mean :21.23
3rd Qu.:11.250	3rd Qu.:249.0	3rd Qu.:77.75	3rd Qu.:26.18
Max. :17.400	Max. :337.0	Max. :91.00	Max. :46.00

Correlação em R

Gráficos de dispersão: função `pairs`.

```
> pairs(USArrests, pch = 20)
```

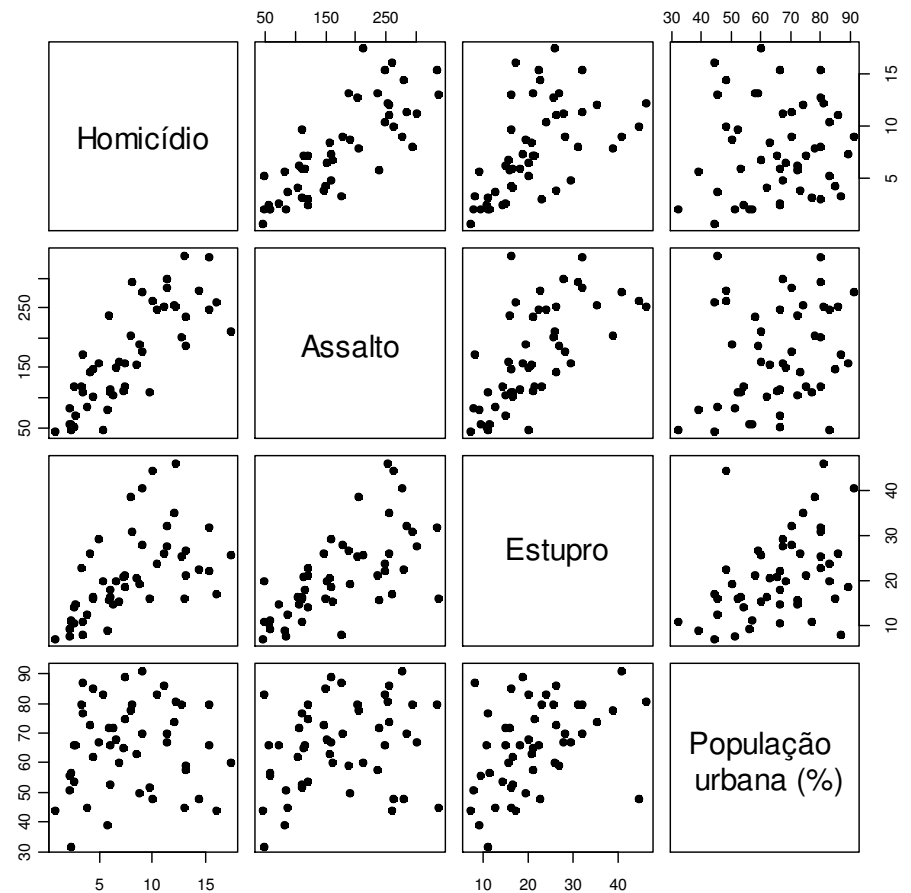


Matriz de gráficos de dispersão (*scatter plot matrix*).

```
> ordem = c("Murder", "Assault",  
"Rape", "UrbanPop")
```

```
> nomes = c("Homicídio", "Assalto",  
"Estupro", "População \n urbana (%)")
```

```
> pairs(USArrests[, ordem], pch = 20,  
labels = nomes)
```



Correlação em R

Matriz de covariâncias:

```
> cov(USArrests[, ordem])
```

	Murder	Assault	Rape	UrbanPop
Murder	18.970465	291.0624	22.99141	4.386204
Assault	291.062367	6945.1657	519.26906	312.275102
Rape	22.991412	519.2691	87.72916	55.768082
UrbanPop	4.386204	312.2751	55.76808	209.518776

Obs. É uma matriz simétrica com as variâncias na diagonal principal.

Matriz de correlações:

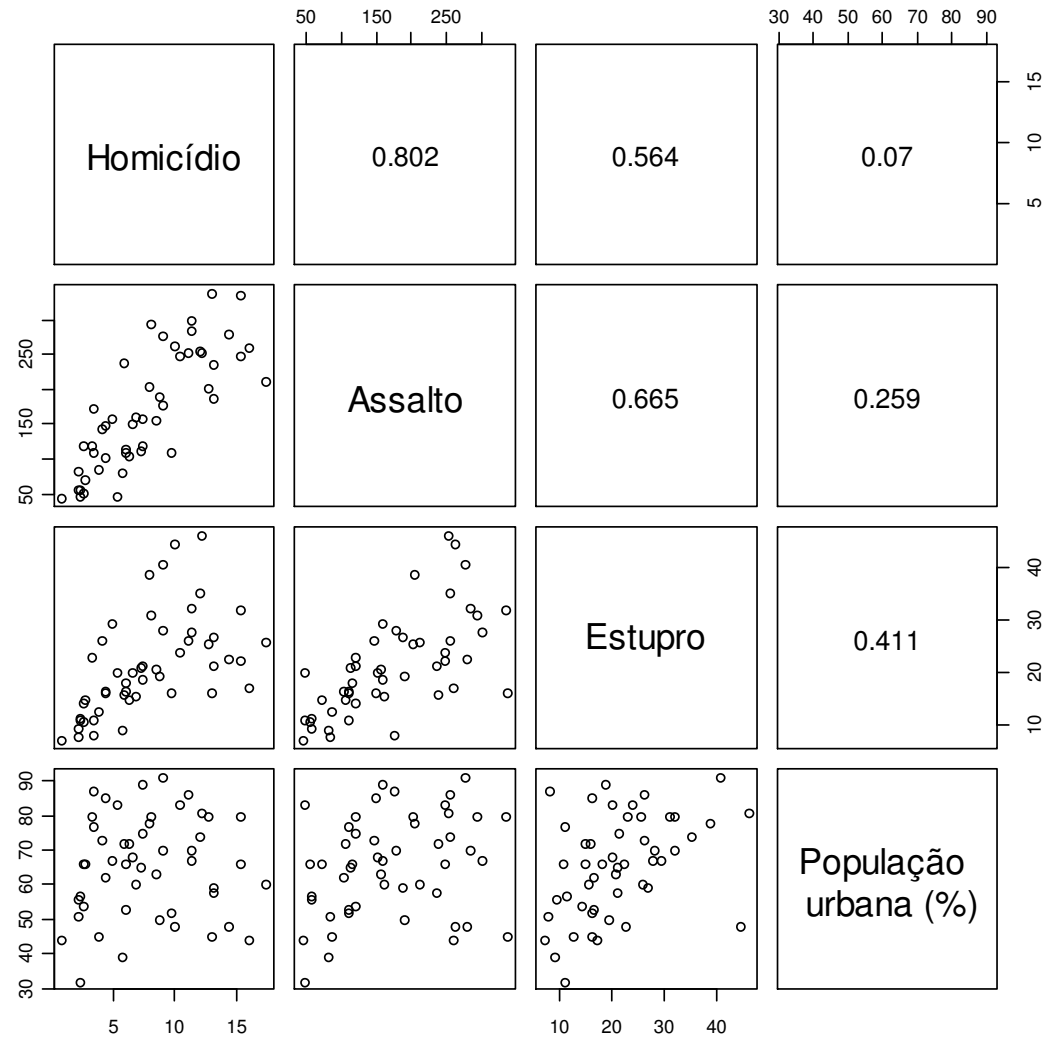
```
> cor(USArrests[, ordem])
```

	Murder	Assault	Rape	UrbanPop
Murder	1.00000000	0.8018733	0.5635788	0.06957262
Assault	0.80187331	1.00000000	0.6652412	0.25887170
Rape	0.56357883	0.6652412	1.00000000	0.41134124
UrbanPop	0.06957262	0.2588717	0.4113412	1.00000000

Obs. A função `cov2cor` transforma uma matriz de covariâncias em uma matriz de correlações.

Correlação em R

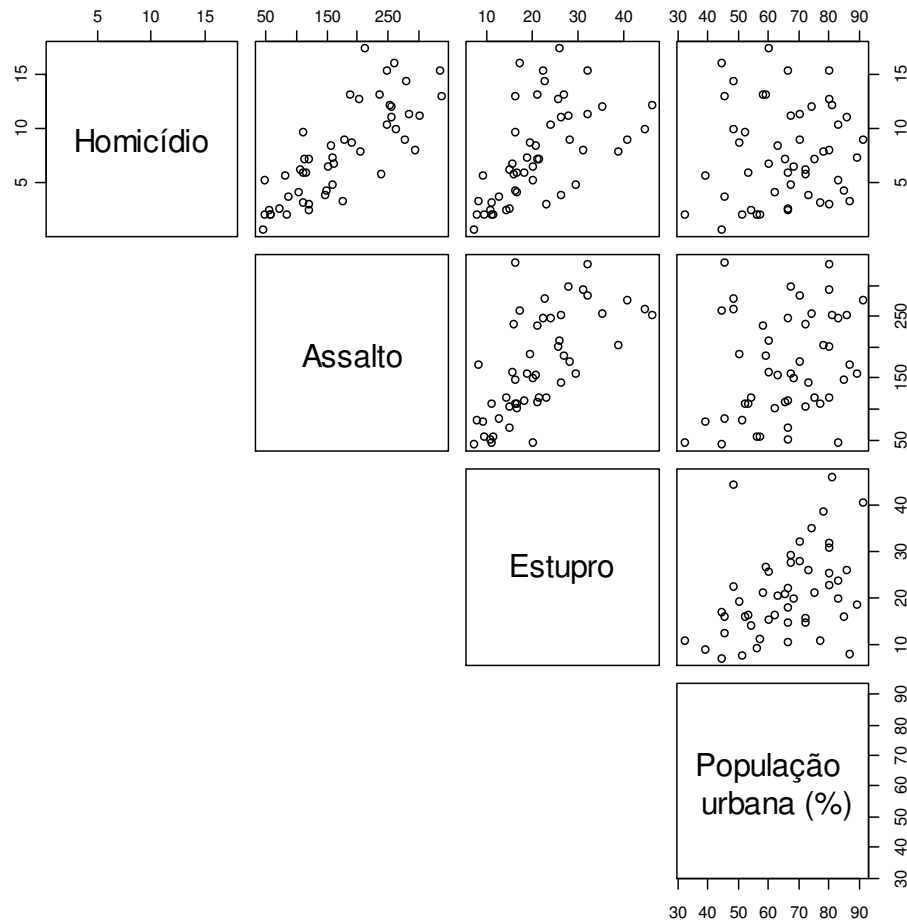
```
> panel.cor = function(x, y,
  digits = 3)
{
  usr = par("usr")
  on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r = cor(x, y)
  text(0.5, 0.5, round(r,
    digits), cex = 1.5)
}
> pairs(USArrests[, ordem],
  labels = nomes, upper.panel =
  panel.cor)
```



Correlação em R

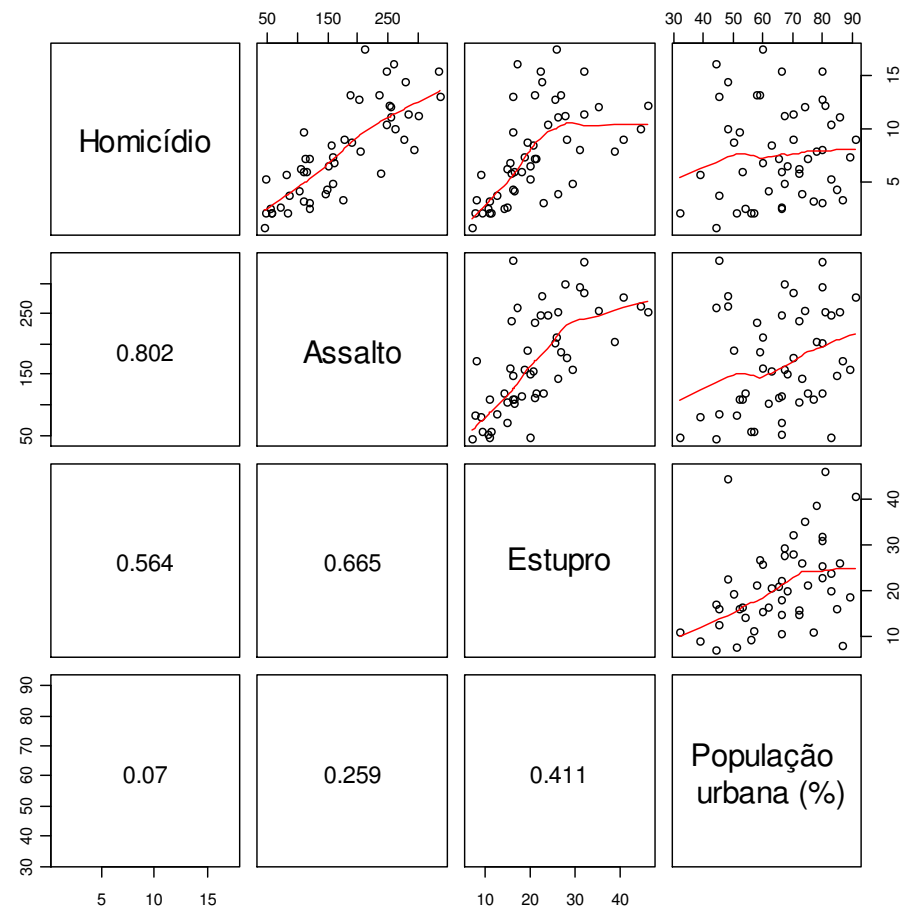
Omitindo a parte inferior da matriz:

```
> pairs(USArrests[, ordem],  
labels = nomes, lower.panel =  
NULL)
```



Correlações e linhas de tendência:

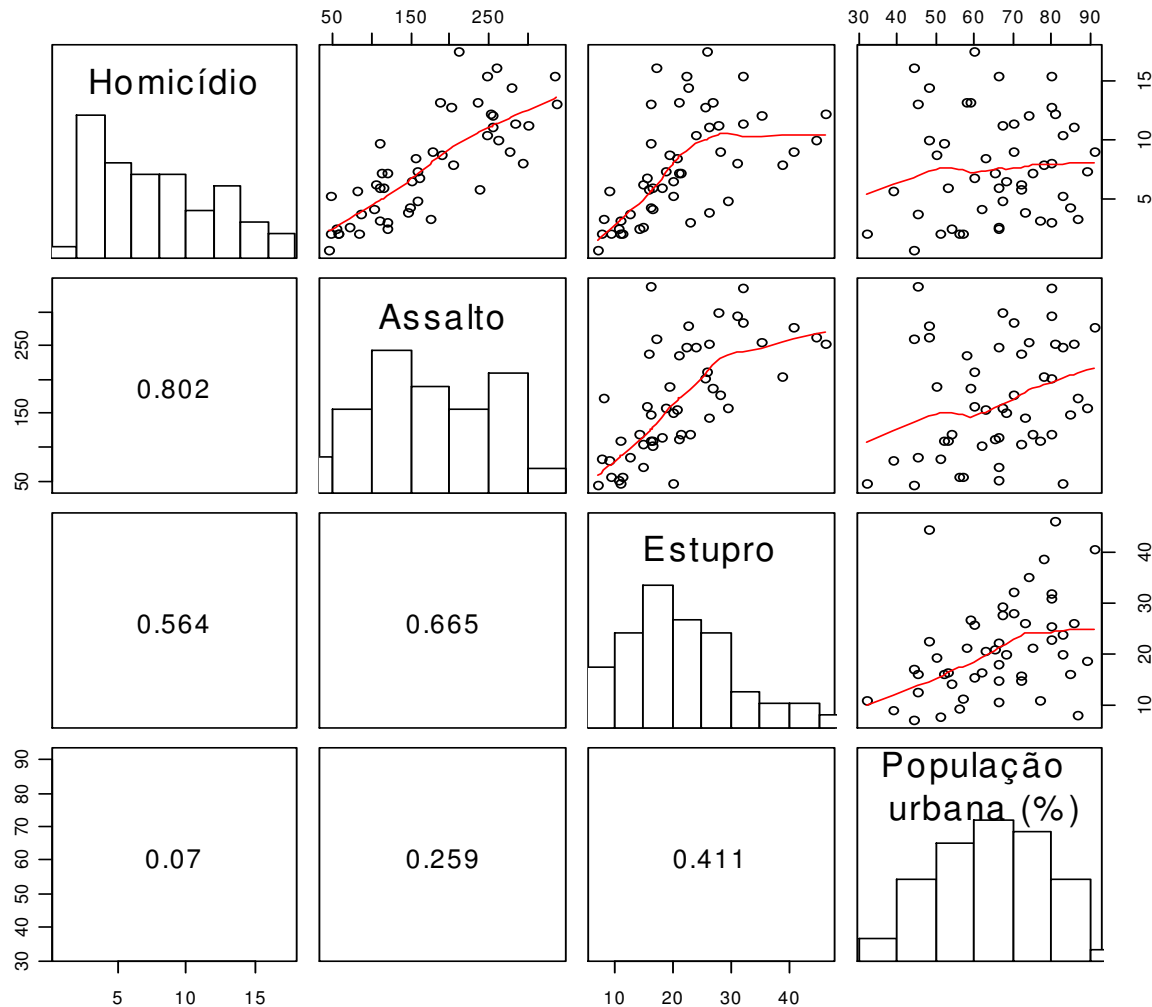
```
> pairs(USArrests[, ordem],  
labels = nomes, upper.panel =  
panel.smooth, lower.panel =  
panel.cor)
```



Correlação em R

Correlações, linhas de tendência e histogramas (utilize `?pairs`):

```
> pairs(USArrests[, ordem], labels = nomes, upper.panel =  
panel.smooth, lower.panel = panel.cor, diag.panel = panel.hist)
```



Quais pares apresentam as correlações mais fracas e mais fortes?

O efeito de urbanização está mais associado a qual tipo de crime?

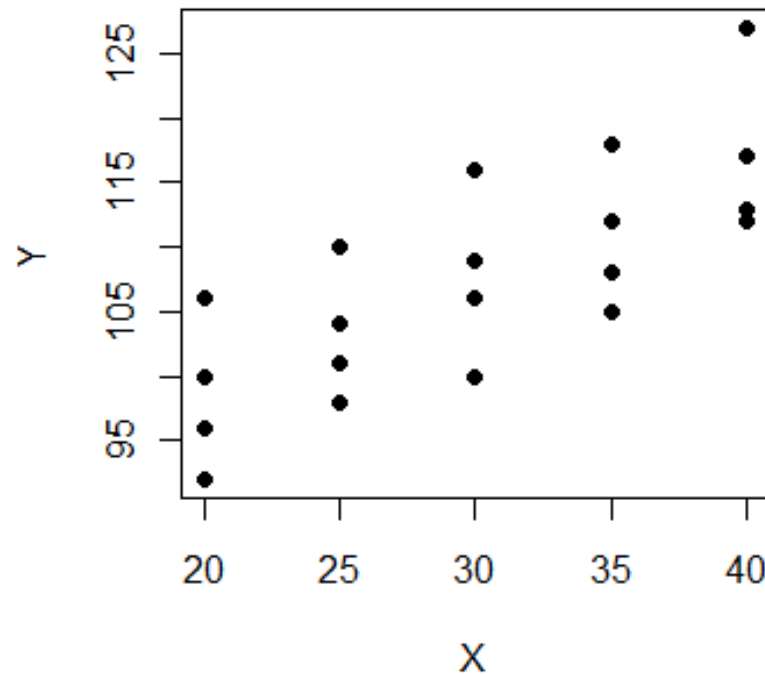
Uma grande quantidade de assaltos resultou em homicídios?

Que outras variáveis poderiam estar relacionadas à ocorrência dos crimes?

Modelo de regressão linear simples

Dados: (x_i, y_i) , $i = 1, \dots, n$. n pares de observações das variáveis x e y (quantitativas). Queremos obter a melhor reta para explicar a (possível) relação linear entre x e y .

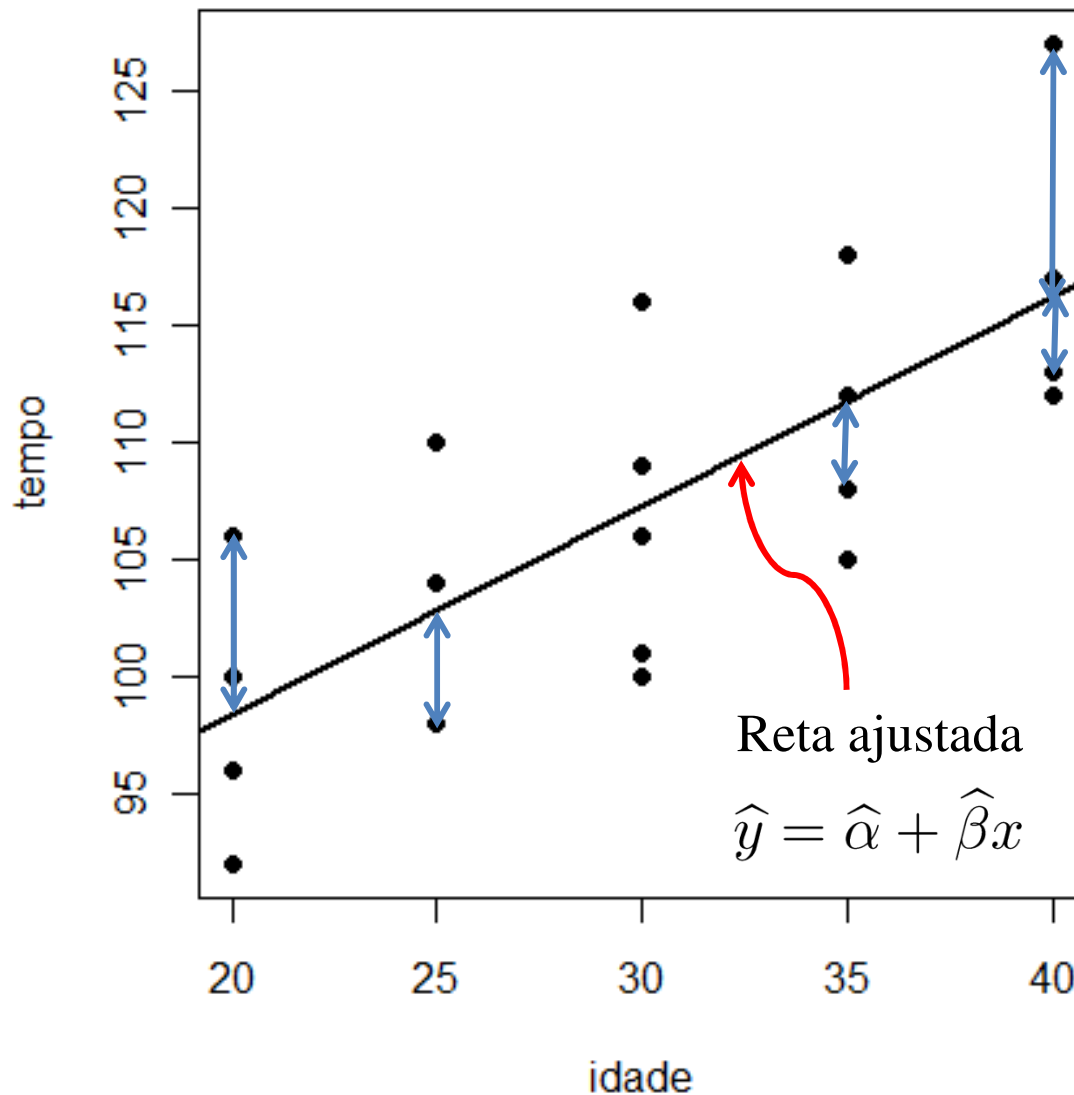
Exemplo:
Observações do tempo de reação (Y) a um certo estímulo e idade (X) de 20 indivíduos



Modelo de regressão linear simples

Ajustar o melhor modelo do tipo $y_i = \alpha + \beta x_i + \epsilon_i$, $i = 1, \dots, n$

α : coeficiente linear (intercepto) e β : coeficiente angular da reta



Para isso, pode-se minimizar soma quadrática dos erros,

$$Q = \sum_{i=1}^n \epsilon_i^2$$

Estimadores de mínimos quadrados de α e β :

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\hat{\beta} = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$$

Modelo de regressão linear simples

Comandos em R:

```
> Acuidade <-read.table(  
"http://wiki.icmc.usp.br/images/0/0f/Acuidade.txt", header=TRUE)  
  
> X <- Acuidade$idade  
> Y <- Acuidade$tempo  
  
> plot(X, Y, pch=16)  
> lm(Y~X)  
> abline(lm(Y~X), col=2)  
> summary(lm(Y~X))
```

Modelo de regressão linear simples

Resultado do ajuste e coeficiente de determinação R^2

```
> summary(lm(Y~X))
```

```
Call:
```

```
lm(formula = Y ~ X)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-7.500 -4.125 -0.750  2.625 10.500
```

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	80.5000	5.4510	14.768	1.67e-11	***
X	0.9000	0.1769	5.089	7.66e-05	***

```
---
```

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

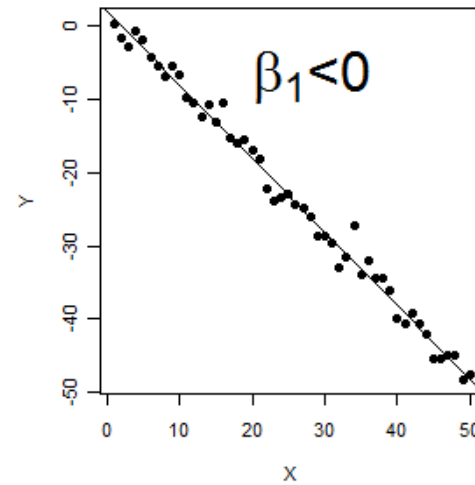
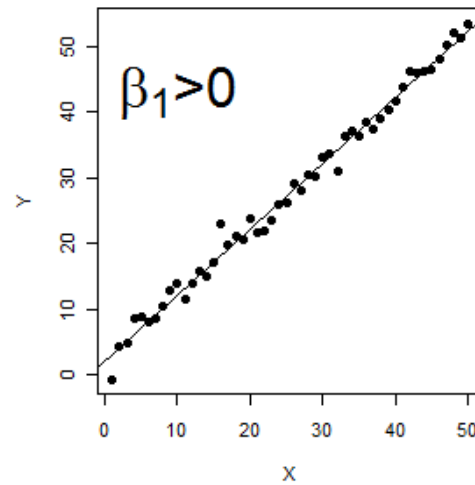
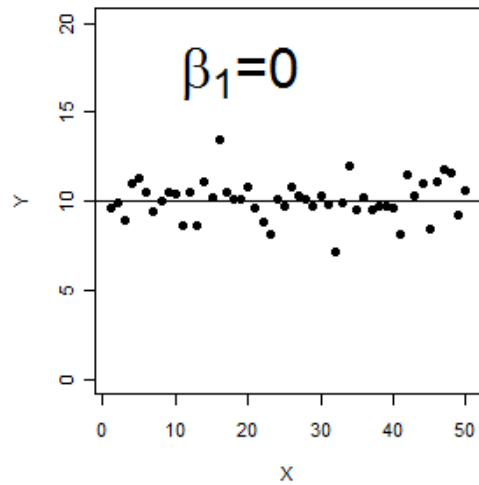
```
Residual standard error: 5.593 on 18 degrees of freedom
```

```
Multiple R-squared: 0.5899, Adjusted R-squared: 0.5672
```

```
F-statistic: 25.9 on 1 and 18 DF, p-value: 7.662e-05
```

Modelo de regressão linear simples

Coeficiente angular no modelo de Regressão Linear Simples



Exemplo em Wainer (2010)

Adaptado de Wainer, W. (2009), *Picturing the Uncertain World*, Princeton: Princeton, NJ

Número médio de **peças por cômodo** em **60** países ou regiões.

Dados: <http://unstats.un.org/unsd/demographic/products/socind/housing.htm>

```
> dados = read.csv("http://www.icmc.usp.br/~cibele/Dados/Housing_Dec2010.csv",  
header = TRUE, sep = ";")
```

```
> names(dados)
```

```
[1] "countryarea" "year" "total" "urban" "rural"
```

```
> summary(dados)
```

year	total	urban	rural
Min. :1976	Min. :0.500	Min. :0.500	Min. : 0.400
1st Qu.:1990	1st Qu.:0.700	1st Qu.:0.700	1st Qu.: 0.700
Median :1991	Median :1.000	Median :1.000	Median : 1.000
Mean :1991	Mean :1.141	Mean :1.153	Mean : 1.230
3rd Qu.:1995	3rd Qu.:1.300	3rd Qu.:1.300	3rd Qu.: 1.400
Max. :1998	Max. :3.000	Max. :3.100	Max. : 3.300
	<u>NA's :2.000</u>	<u>NA's :8.000</u>	<u>NA's :10.000</u>

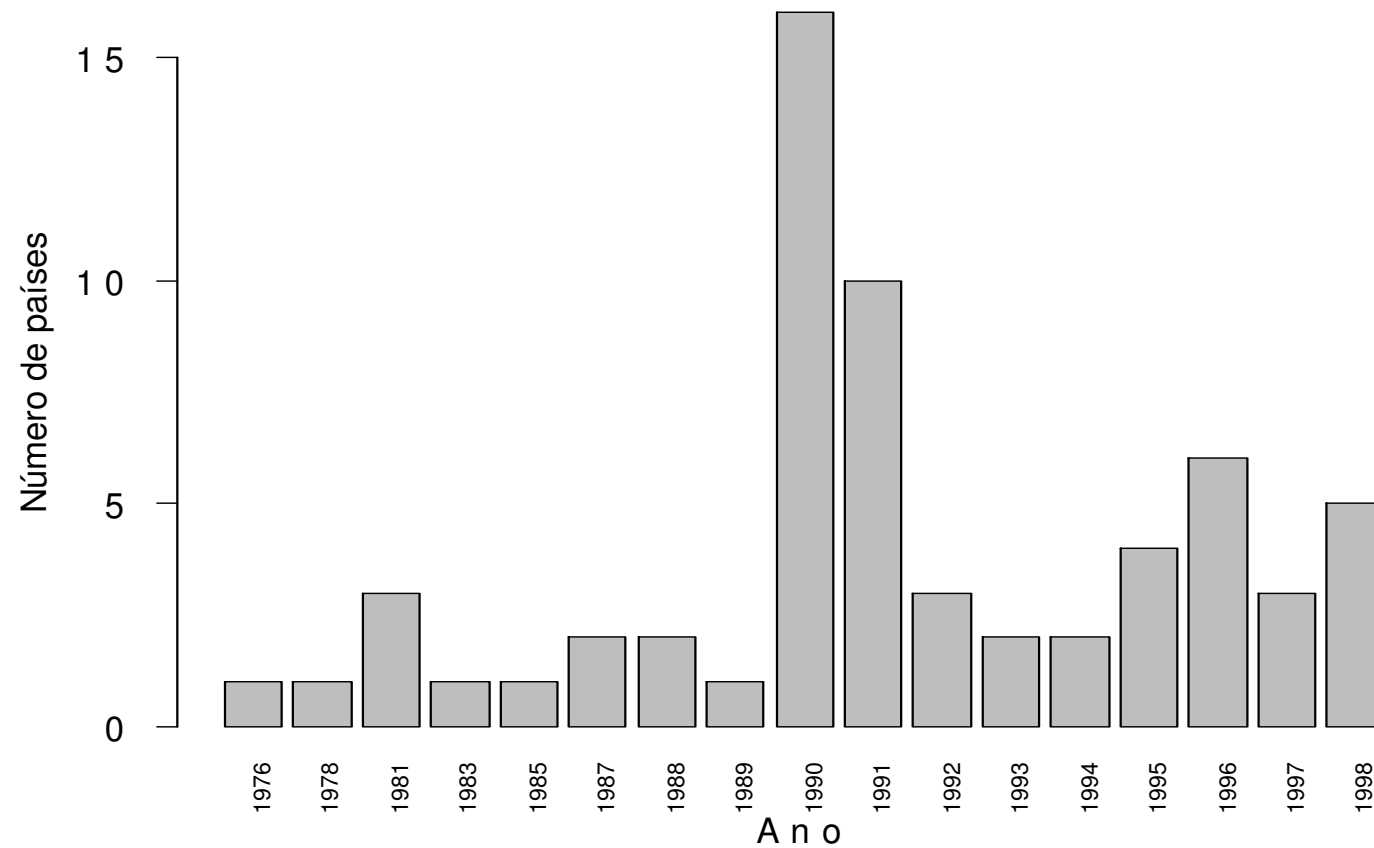
É possível comparar dados coletados de **1976** com os de **1998**?

Exemplo em Wainer (2009)

```
> attach(dados)
```

```
> table(year)
```

```
> barplot(table(year), xlab = "Ano", ylab = "Número de países",  
las = 2)
```

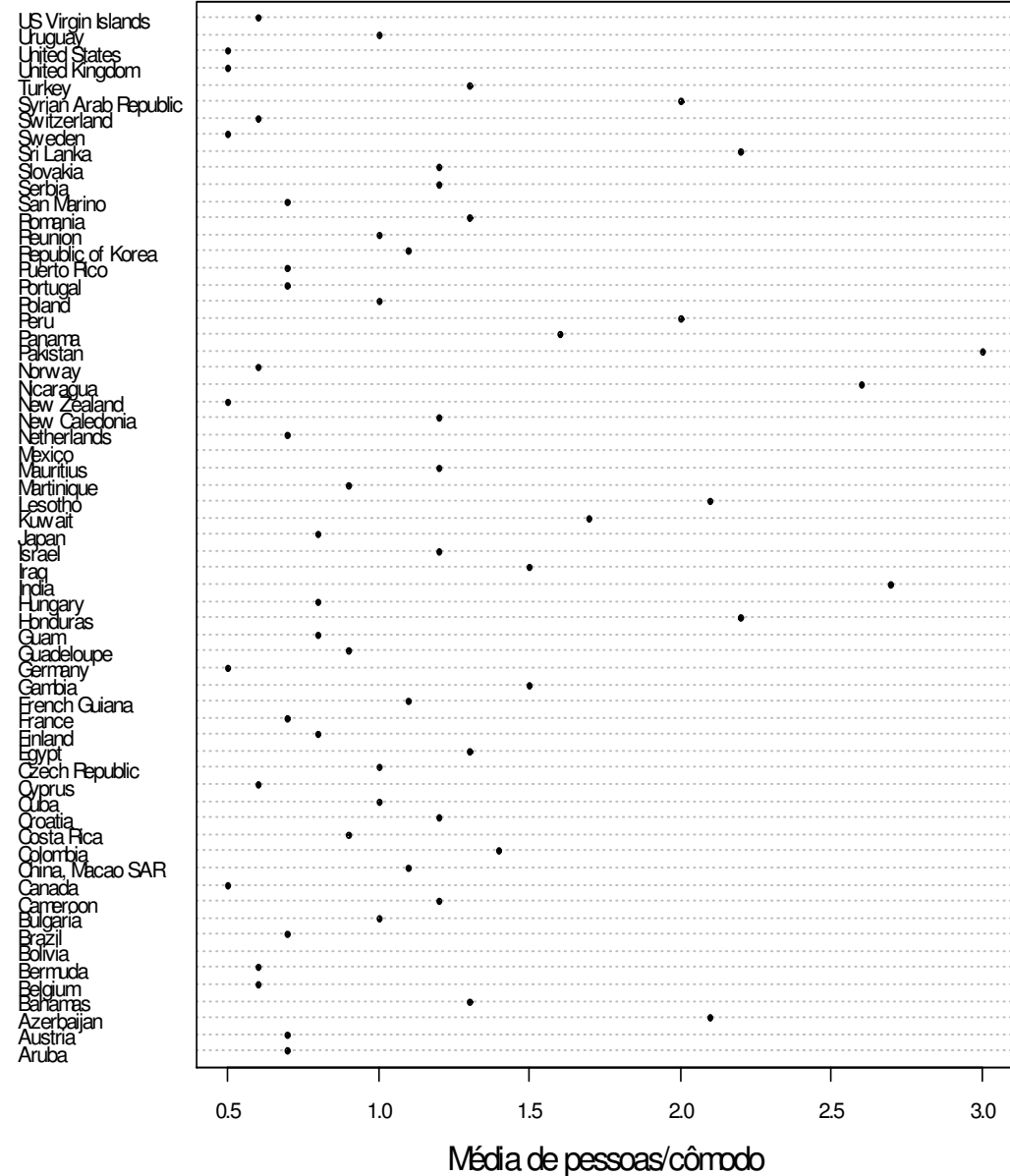


Exemplo em Wainer (2009)

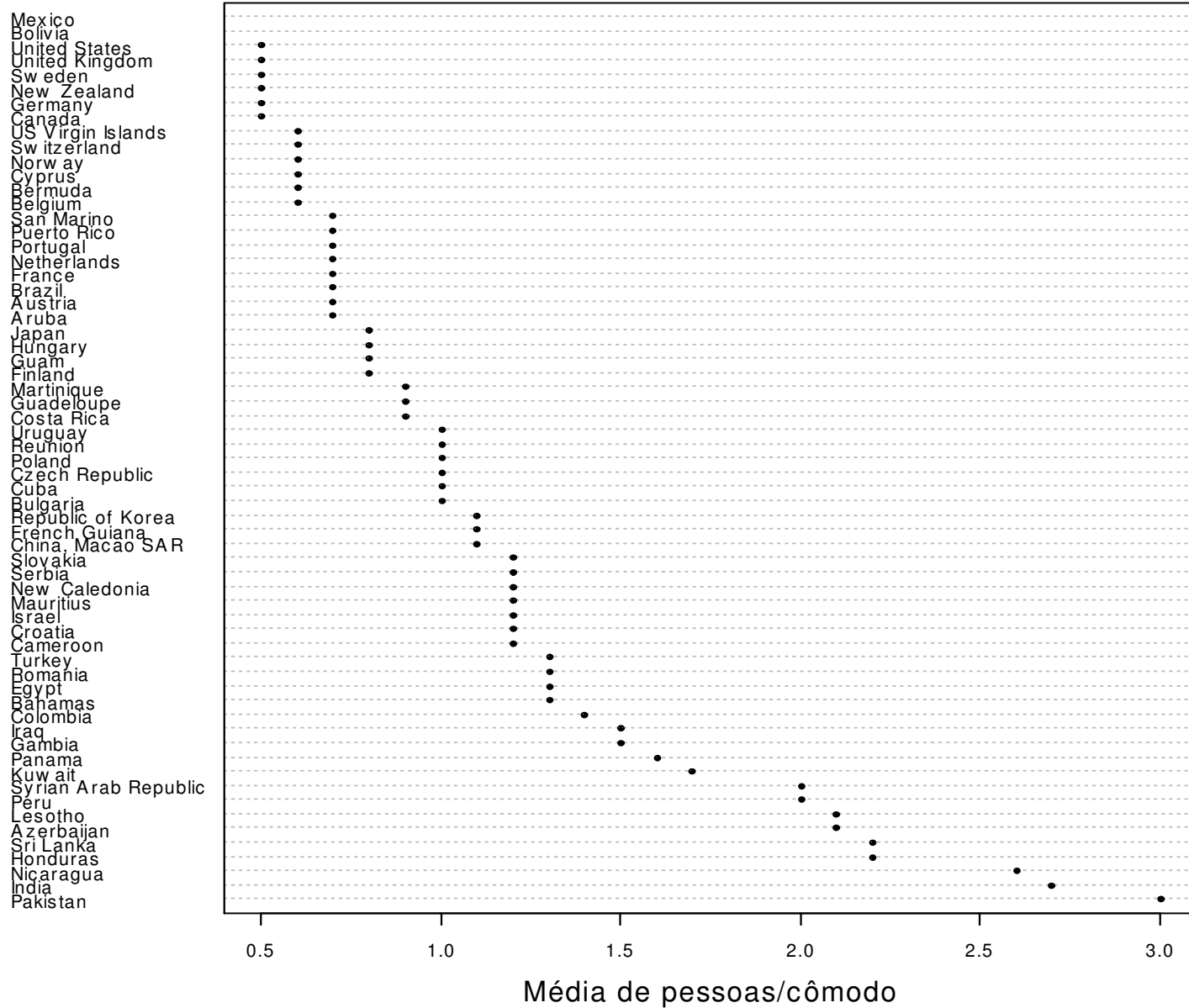
```
> countryarea[year == 1976]
[1] Cameroon
> countryarea[year == 1998]
[1] Azerbaijan  Brazil
Finland      Netherlands
Pakistan
> dotchart(total, labels =
countryarea, xlab = "Média
de pessoas/cômodo", pch =
20, cex = 0.7, cex.lab =
1.5)
```

Por que utilizar a ordem
alfabética?

```
> ordem = order(total,
decreasing = TRUE)
> dotchart(total[ordem],
labels =
countryarea[ordem], xlab =
"Média de pessoas/cômodo",
pch = 20, cex = 0.7,
cex.lab = 1.5)
```



Exemplo em Wainer (2009)



Exemplo em Wainer (2009)

```
> plot(year, total, xlab = "Ano", ylab = "Média de pessoas/cômodo",  
pch = 20)
```

```
> abline(lm(total ~  
year), lty = 2)
```

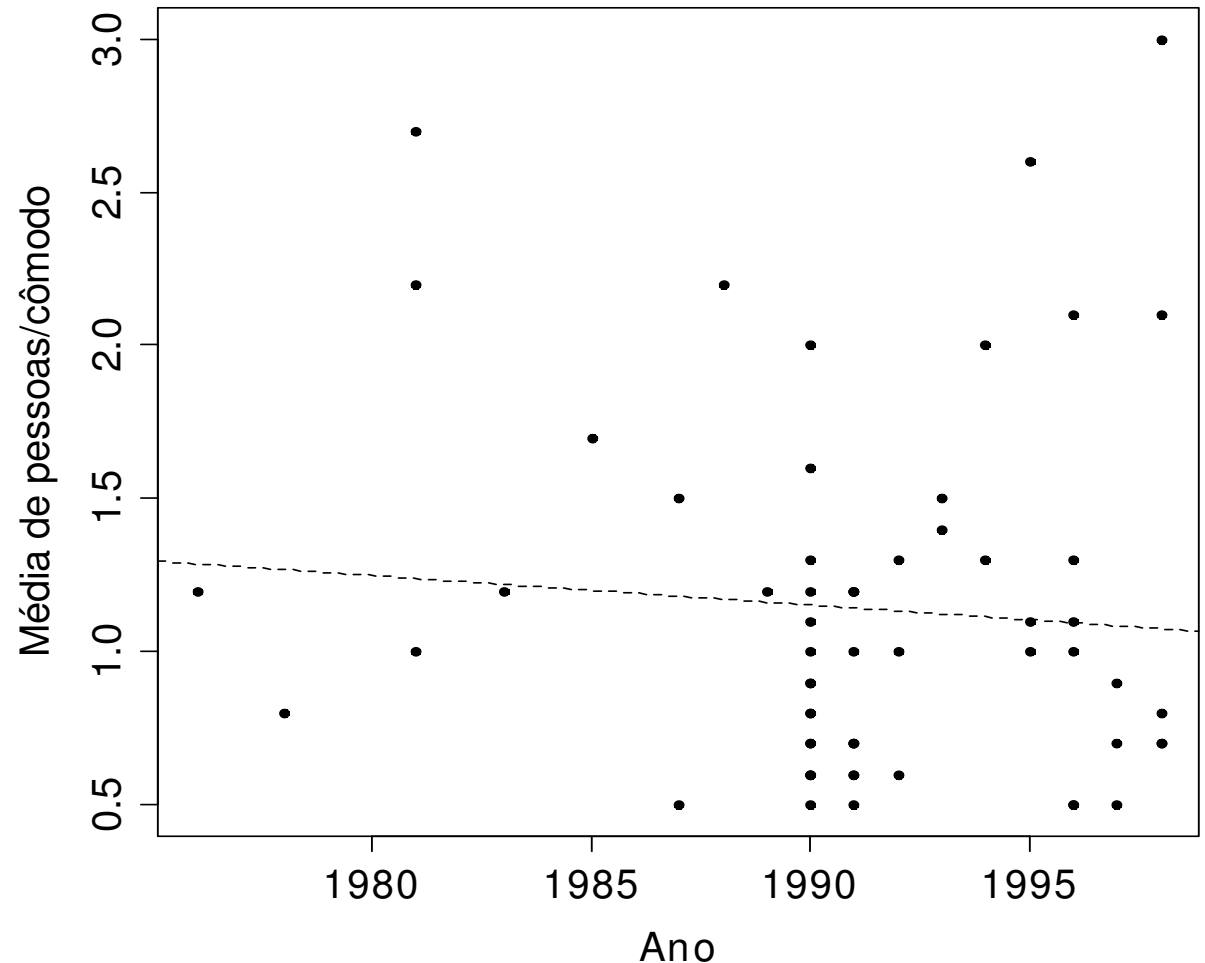
```
> cor(year, total)
```

```
[1] NA
```

```
> cor(year, total,  
use = "complete")
```

```
[1] -0.07985232
```

Não há indício de relação entre a densidade de ocupação e o ano em que o dado foi coletado.



Há diferença entre a ocupação nos meios rural e urbano?

Se a resposta for não, podemos trabalhar com a média geral (`total`).

Exemplo em Wainer (2009)

```
> plot(rural, urban, xlab = "Média de pessoas/cômodo - rural",  
ylab = "Média de pessoas/cômodo - urbano", pch = 20)
```

```
> abline(0, 1, lty = 2)
```

```
> cor(rural, urban,  
use = "complete")
```

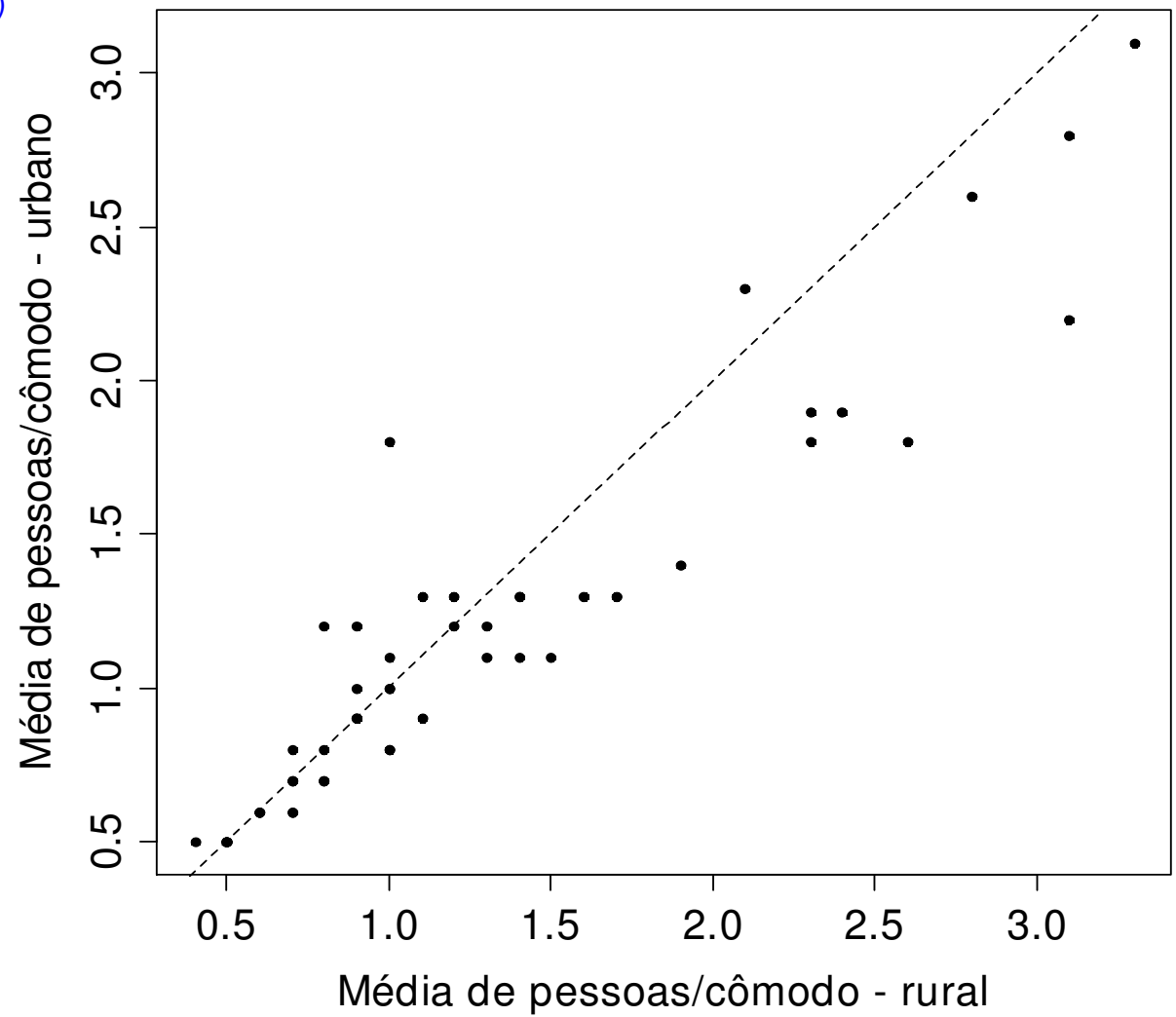
```
[1] 0.9385013
```

Correlação positiva forte.

Tendência de maiores médias no meio rural.

Situação econômica pode estar associada à densidade de ocupação?

Variável: PIB *per capita*.



Exemplo em Wainer (2009)

```
> pib = read.csv("Income_Dec2009.csv", header = TRUE, sep = ";")
```

```
> names(pib)
```

```
[1] "countryarea" "year" "GDPcapita"
```

```
> summary(pib)
```

countryarea	year	GDPcapita
Afghanistan: 1	Min. 2008	Min. : 138
Albania : 1	1st Qu. 2008	1st Qu.: 1218
Algeria : 1	Median 2008	Median : 4874
Andorra : 1	Mean 2008	Mean : 15772
Angola : 1	3rd Qu. 2008	3rd Qu.: 19291
Anguilla : 1	Max. 2008	Max. : 211501
(Other) :209	<u>NA's : 6</u>	<u>NA's : 6</u>

```
> pib$country[which.min(pib$GDPcapita)]
```

```
[1] Burundi
```

```
> pib$country[which.max(pib$GDPcapita)]
```

```
[1] Monaco
```

```
> dim(pib)
```

```
[1] 215 3
```

Dados de 2008
serão utilizados
apenas como
ilustração.

GDP: *per capita*
gross domestic
product (em US\$).

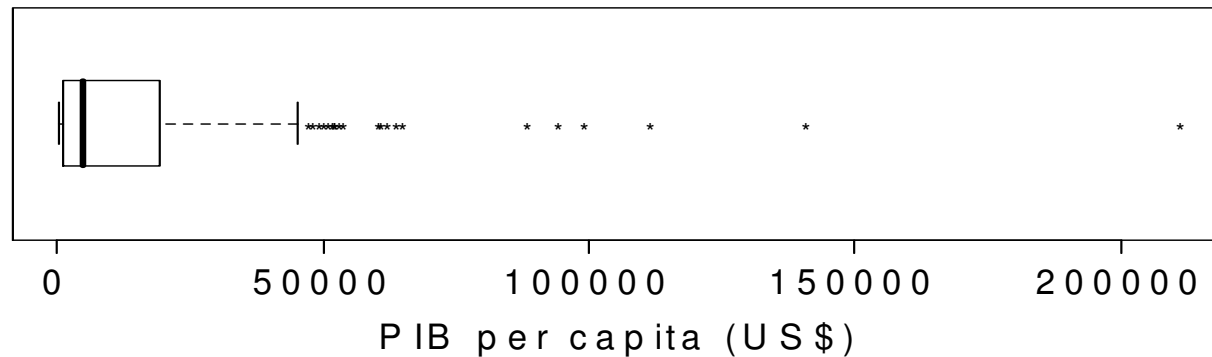
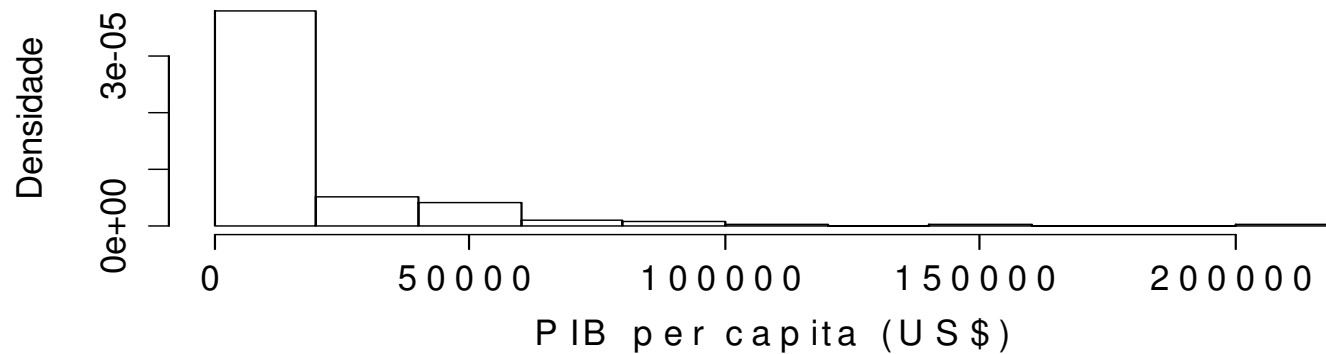
[http://unstats.un.org/
unsd/snaama/dnllist.
asp](http://unstats.un.org/unsd/snaama/dnllist.asp)

```
> pib$GDP[pib$country ==  
"Brazil"]
```

```
[1] 8311
```

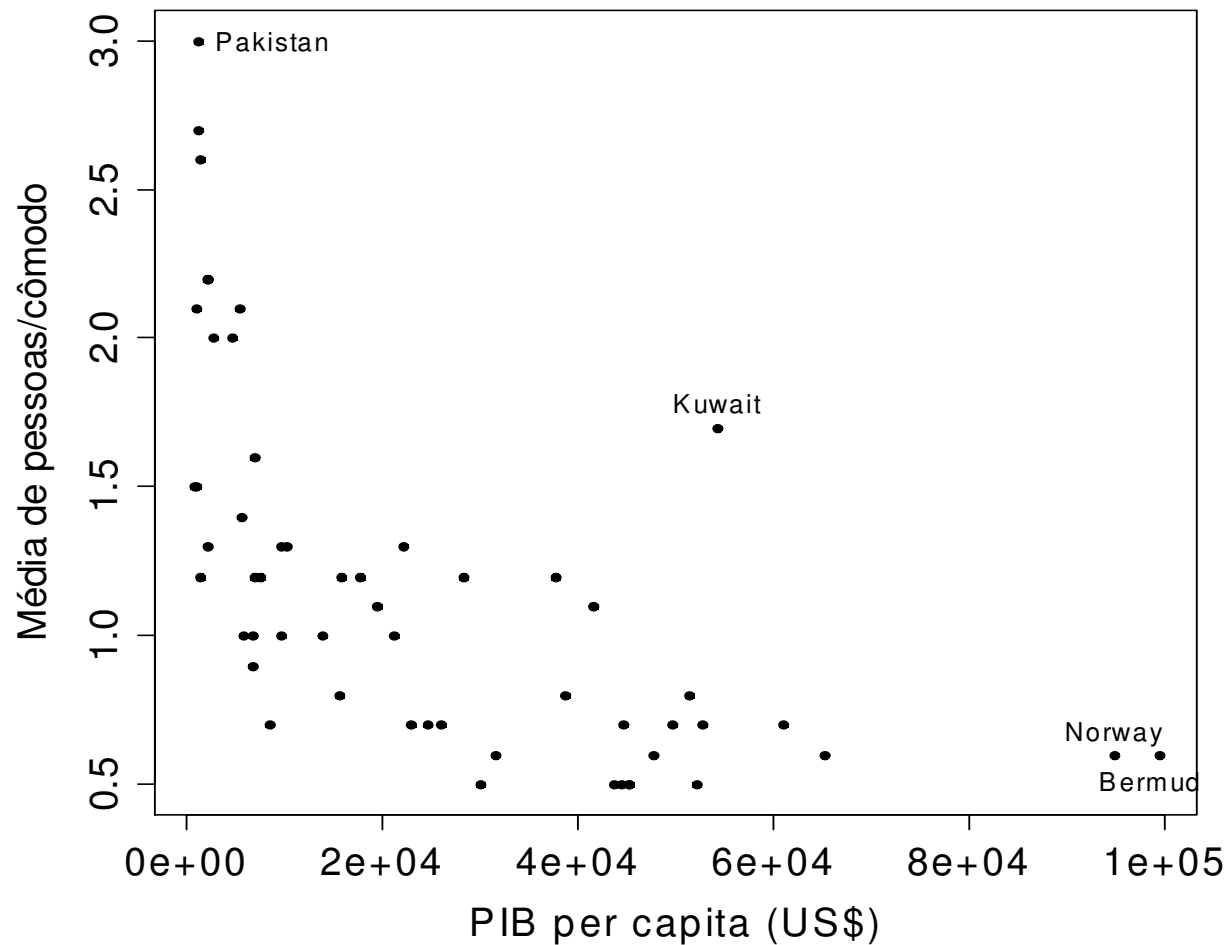
Exemplo em Wainer (2009)

```
> par(mfrow = c(2, 1))  
> hist(pib$GDP, freq = FALSE, xlab = "PIB per capita (US$)", ylab =  
"Densidade", main = "")  
> boxplot(pib$GDP, xlab = "PIB per capita (US$)", pch = "*",  
horizontal = TRUE)
```



Exemplo em Wainer (2009)

```
> pib60 = pib$GDP[match(countryarea, pib$country)]  
> plot(pib60, total, pch = 20, ylab = "Média de pessoas/cômodo",  
       xlab = "PIB per capita (US$)")  
> identify(pib60, total, countryarea)
```



Associação **negativa**.

Assimetria em PIB
per capita.

Transformações de variáveis

Alguns objetivos: (a) **simetrizar** os dados e (b) **linearizar** a relação entre as variáveis.

$$\text{Família de transformações: } t = t(x) = \begin{cases} x^\lambda, & \text{se } \lambda \neq 0, \\ \log(x), & \text{se } \lambda = 0, \end{cases} \text{ se } x > 0.$$

λ deve ser **escolhido** de modo a atingir o(s) objetivo(s), pelo menos aproximadamente.

$t(x)$ é monótona em x :

$$(1) \lambda \geq 0. \quad x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)} \Leftrightarrow t(x_{(1)}) \leq t(x_{(2)}) \leq \dots \leq t(x_{(n)}).$$

$$(2) \lambda < 0. \quad x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)} \Leftrightarrow t(x_{(n)}) \leq t(x_{(n-1)}) \leq \dots \leq t(x_{(1)}).$$

Posições são **preservadas** em (1) e são **invertidas** em (2).

Obs. Se M é a mediana de x , então $t(M)$ é a mediana de t .

Transformações **comuns**: $\log(x)$, $x^{1/2}$, $1/x$ e $1/x^2$.

Exemplo em Wainer (2009)

Transformação **logarítmica** da variável PIB *per capita*.

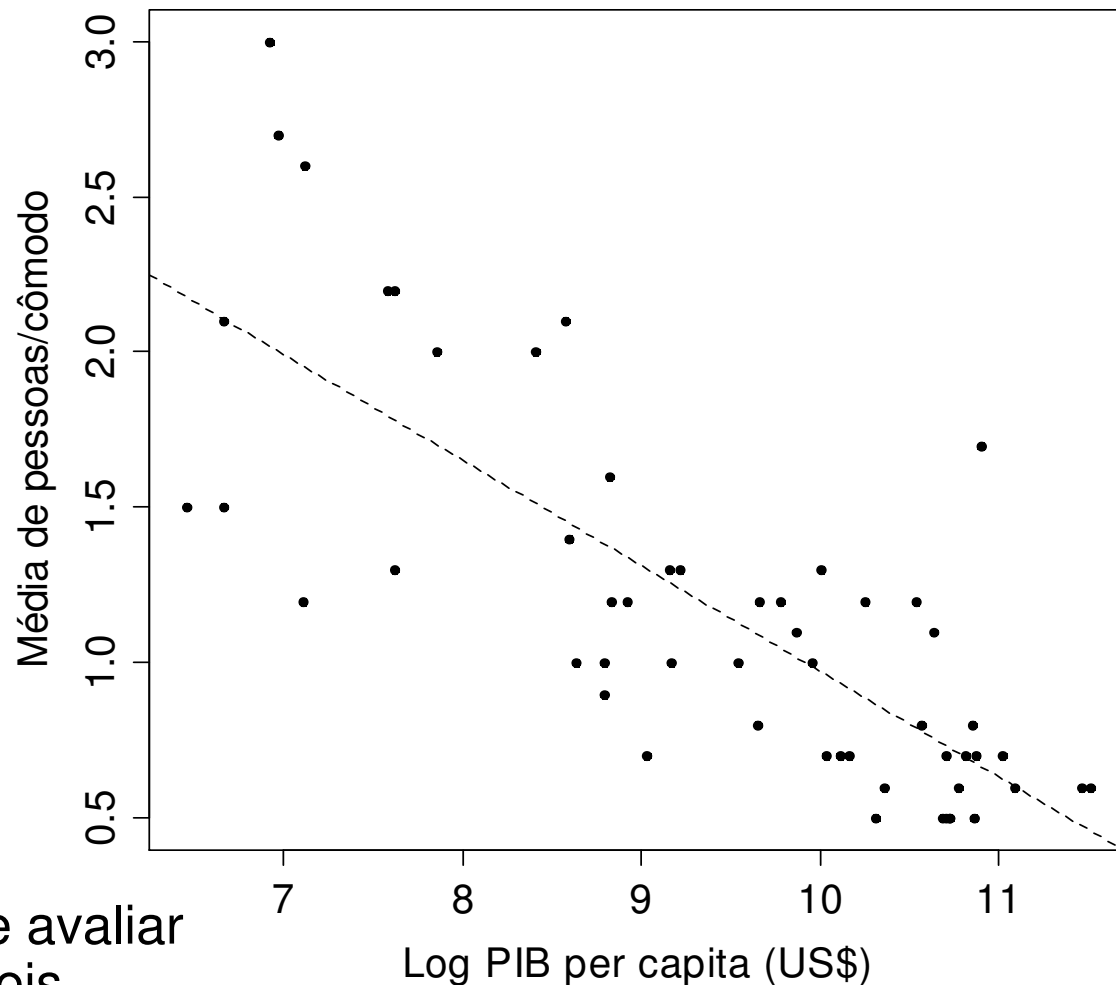
```
> plot(log(pib60),  
total, pch = 20, ylab =  
"Média de  
pessoas/cômodo", xlab =  
"Log PIB per capita  
(US$)")
```

```
> abline(lm(total ~  
log(pib60)), lty = 2)
```

```
> cor(log(pib60),  
total, use =  
"complete")
```

```
[1] -0.7787283
```

Outras variáveis:
fertilidade e
desemprego
feminino.



Exercício: Baixar dados e avaliar associações entre variáveis

<http://unstats.un.org/unsd/demographic/products/socind/>