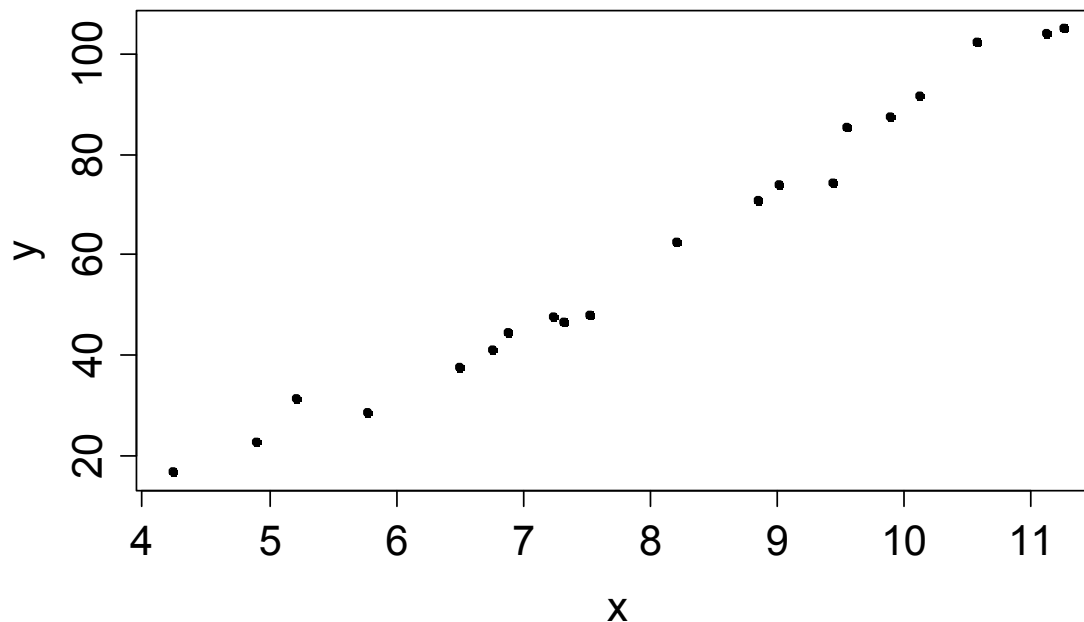


1. Coeficiente de correlação

$(x_1, y_1), \dots, (x_n, y_n)$: 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.

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

1. Coeficiente de correlação

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):

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

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 $b < 0$.

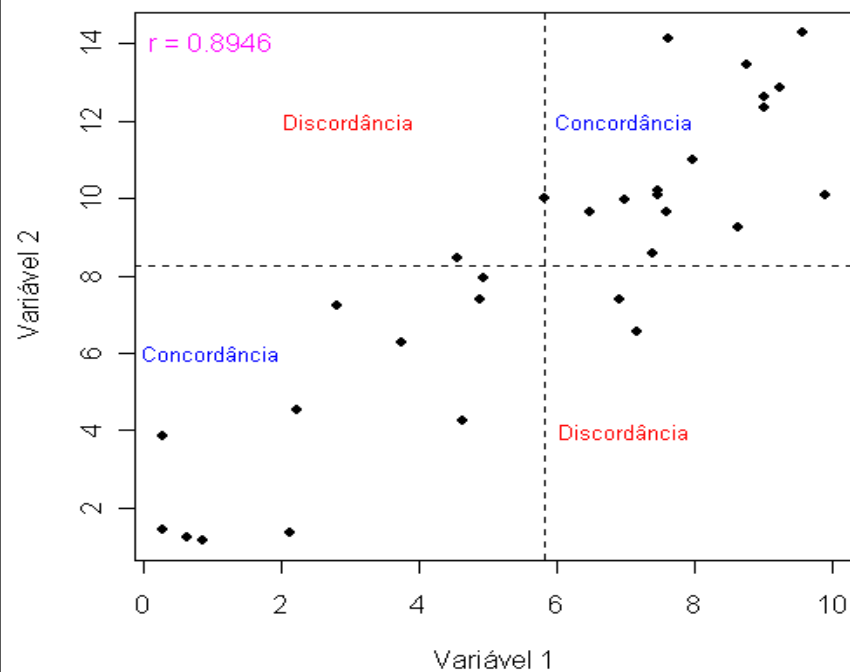
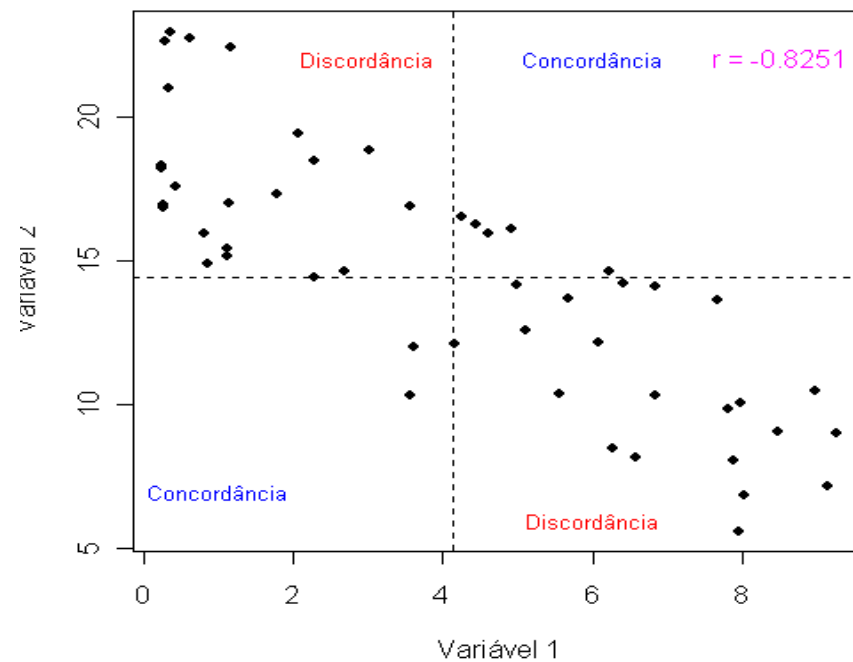
P5. Invariância. Se $b_1 > 0$ e $b_2 > 0$, então $\text{cor}(x, y) = \text{cor}(a_1 + b_1x, a_2 + b_2y)$, em que a_1 e a_2 são reais quaisquer.

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 $\text{cor}(a_1 + b_1x, a_2 + b_2y)$?

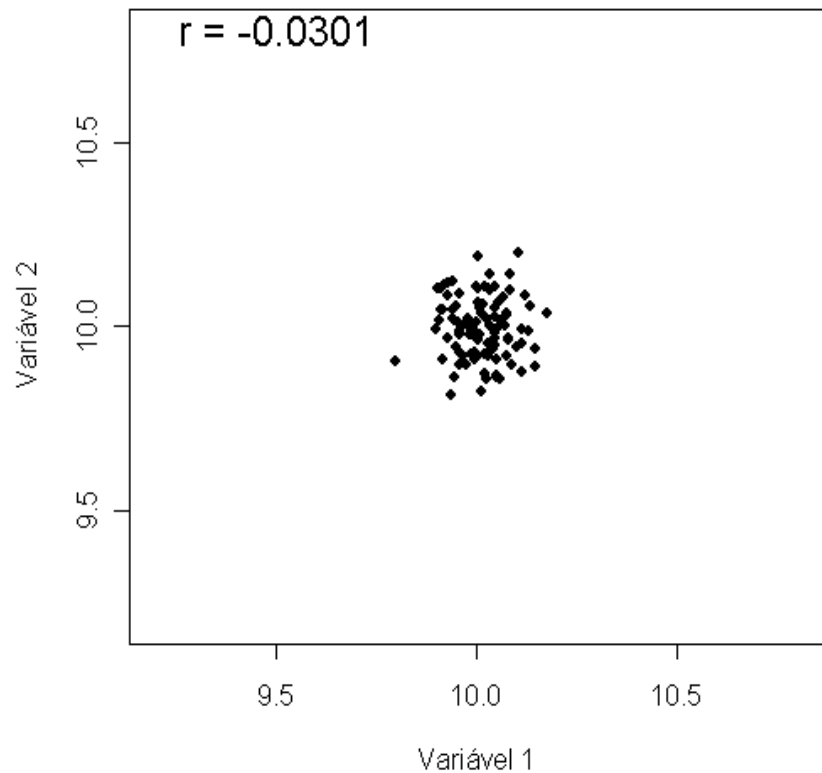
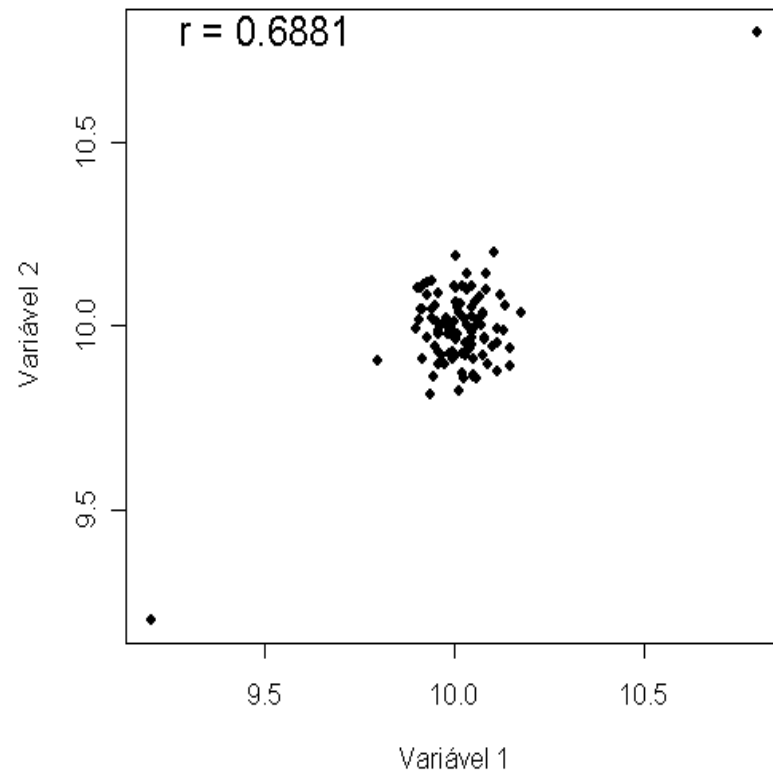
1. Coeficiente de correlação

Sentido e força de r

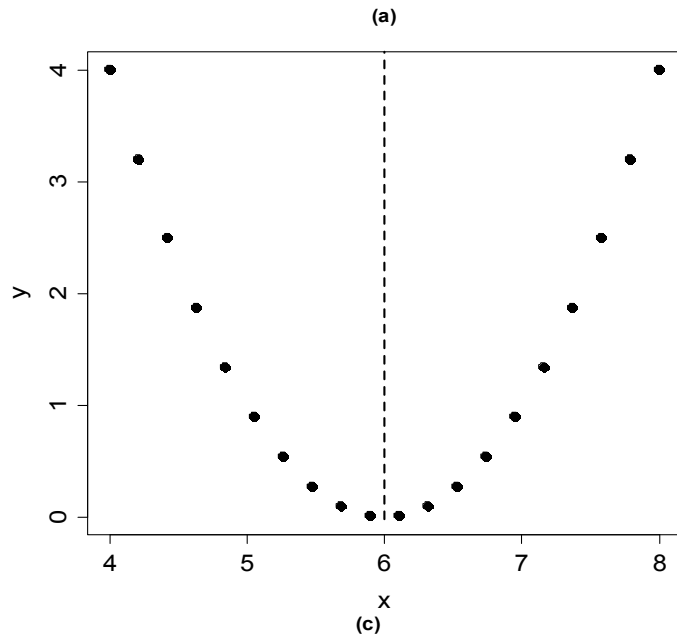
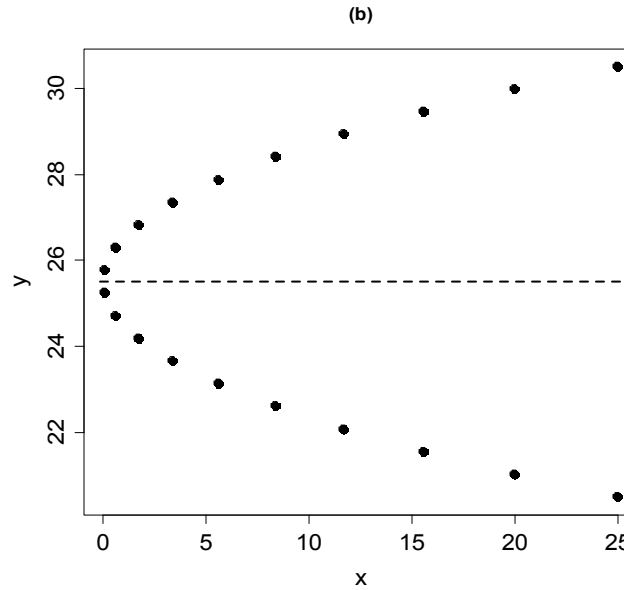
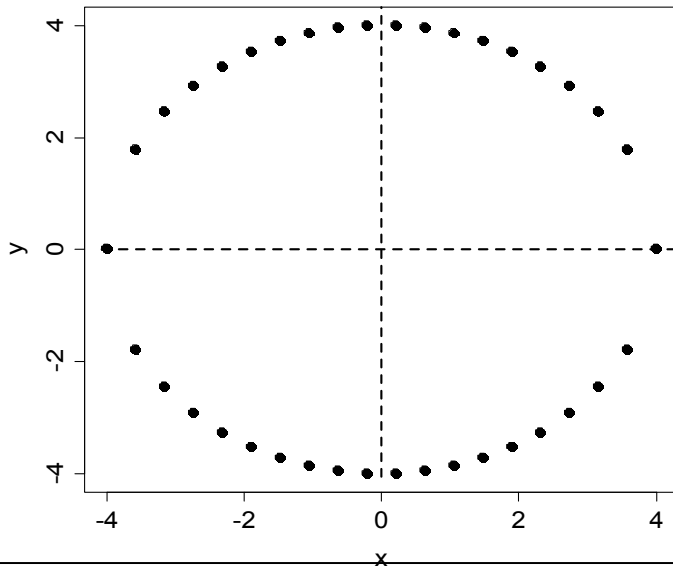
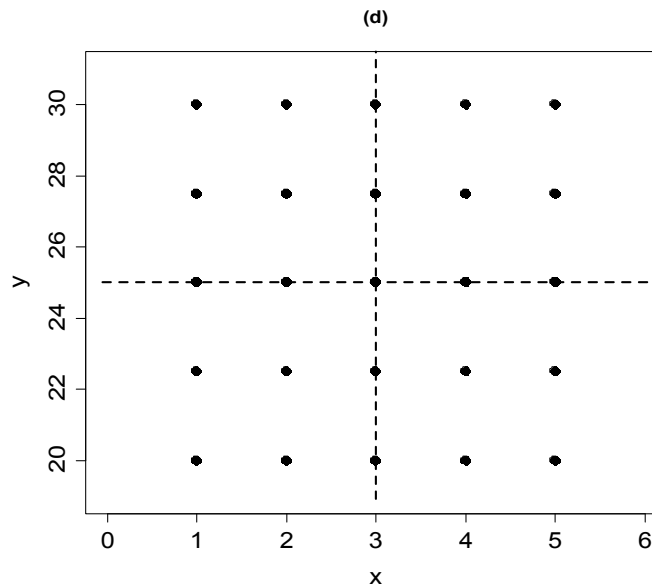
-1	negativa moderada	negativa fraca	0 ausência	positiva fraca	positiva moderada	1
negativa forte						positiva forte



1. Coeficiente de correlação



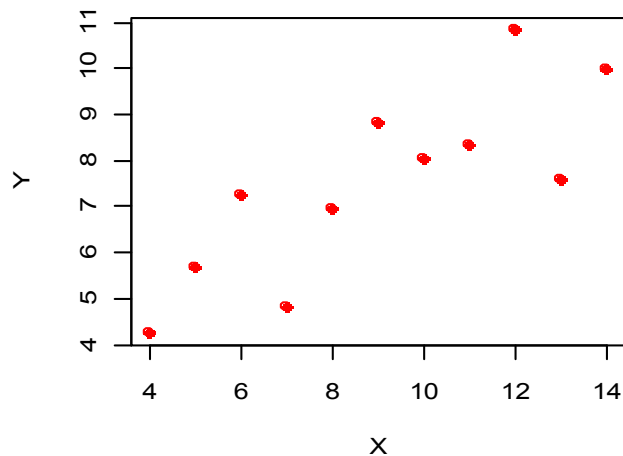
1. Coeficiente de correlação



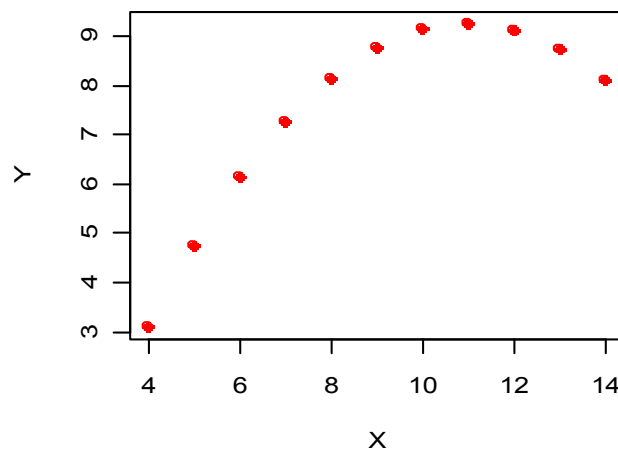
Obs. $r = 0$ não significa ausência de associação.

1. Coeficiente de correlação

Exemplo 1



Exemplo 2



Dados anscombe em R

`> ?anscombe`

Valores de r:

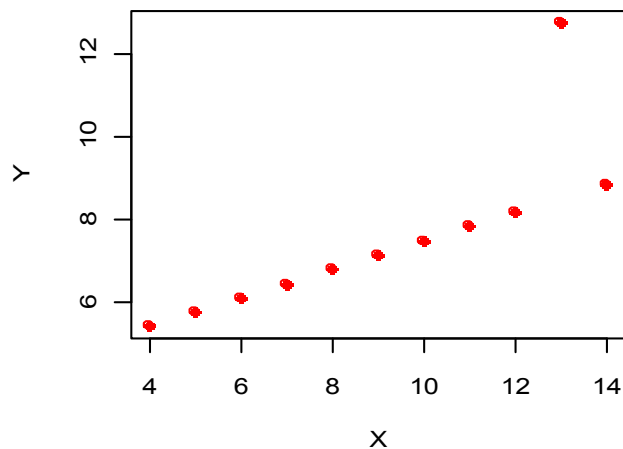
Exemplo 1:
0,8164

Exemplo 2:
0,8162

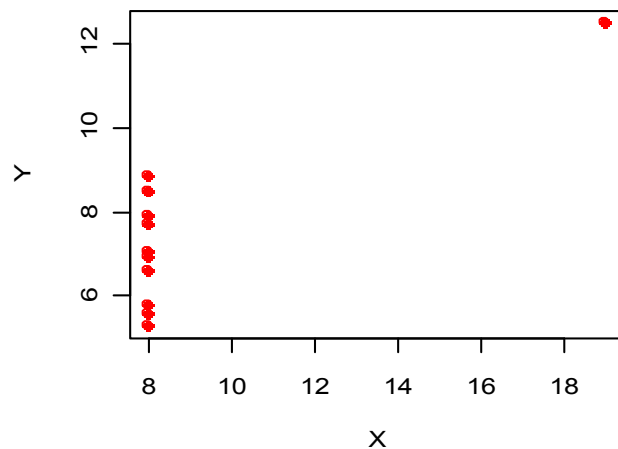
Exemplo 3:
0,8163

Exemplo 4:
0,8165

Exemplo 3



Exemplo 4



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

1. Coeficiente de correlação

Funções em R: `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)
```

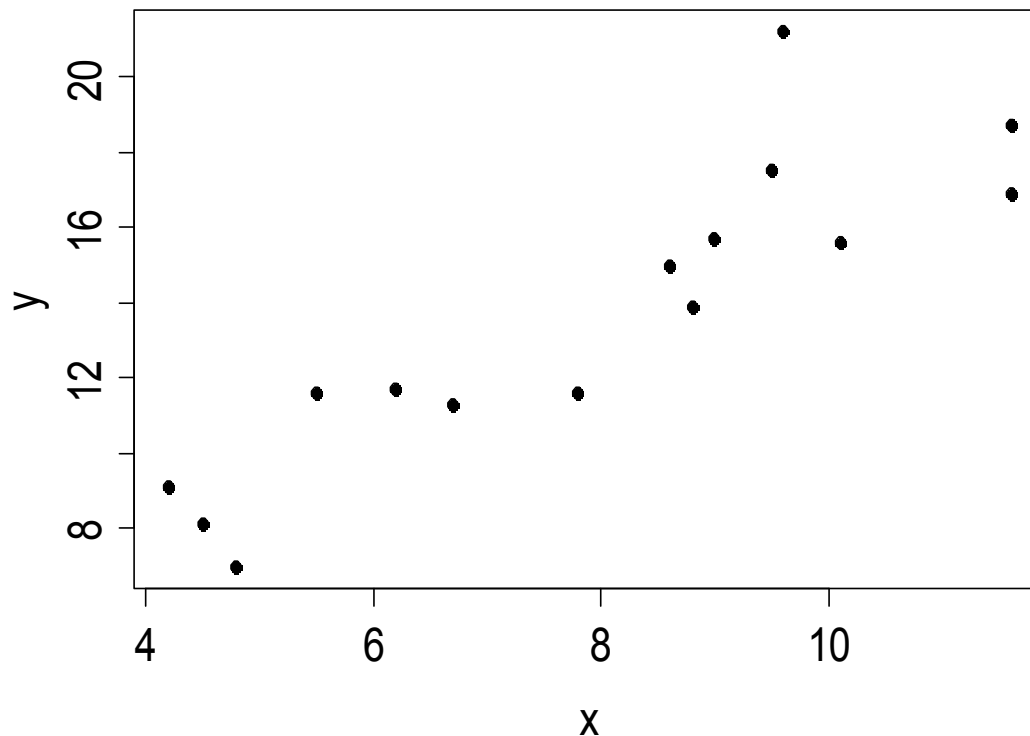
```
> length(x)
```

```
[1] 15
```

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

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

```
[1] 0.8908723
```



1. Coeficiente de correlação

```
> ? 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 100000 hab. e proporção da população urbana.

```
> names(USArrests)
```

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

```
> rownames(USArrests)
```

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

```
> 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

```
> class(USArrests)
```

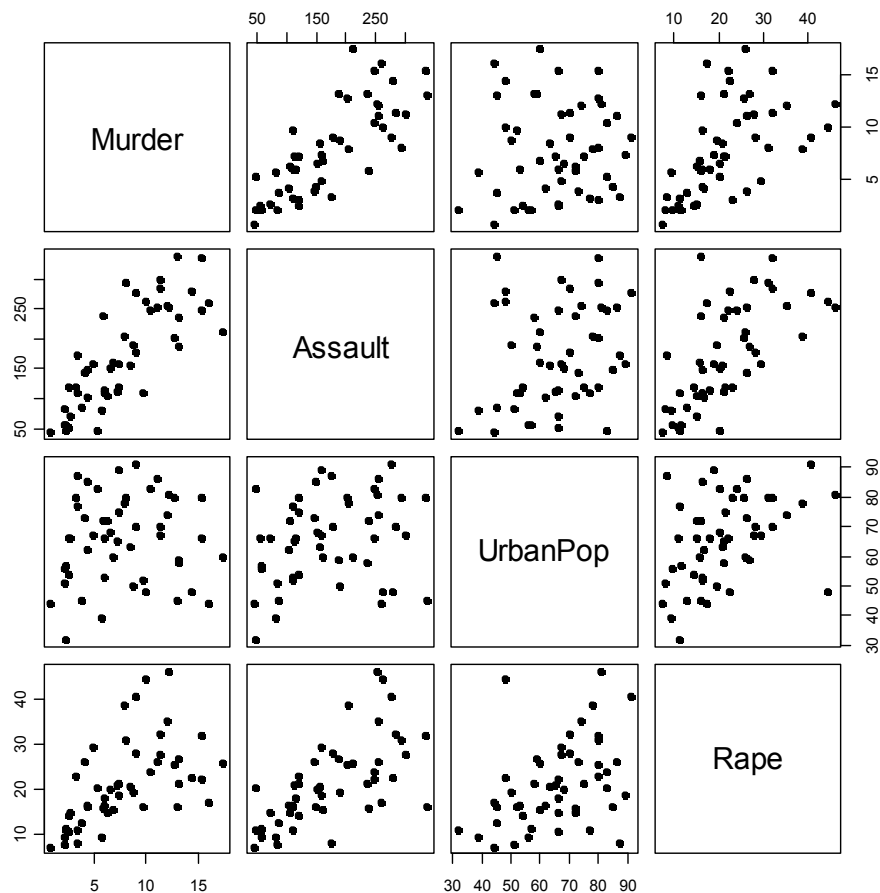
```
[1] "data.frame"
```

Classe “folha de dados”.

1. Coeficiente de correlação

Gráficos de dispersão: função `pairs`.

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

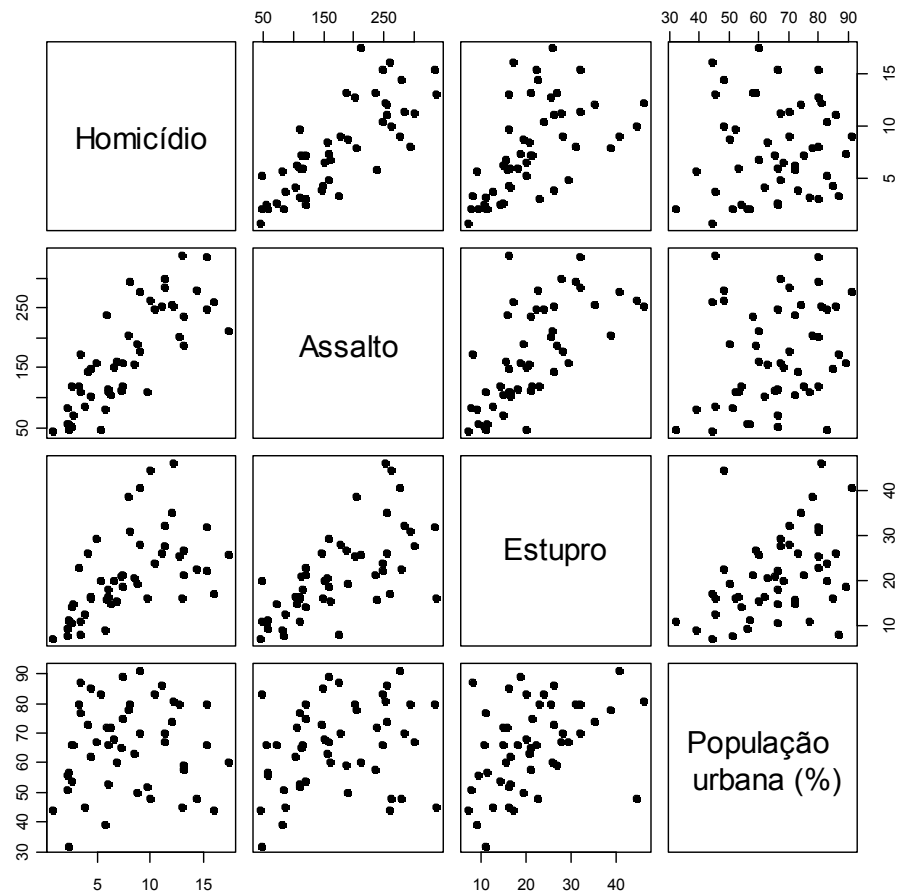


Matriz de gráficos de dispersão
(*scatter plot matrix*).

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

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

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



1. Coeficiente de correlação

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])
```

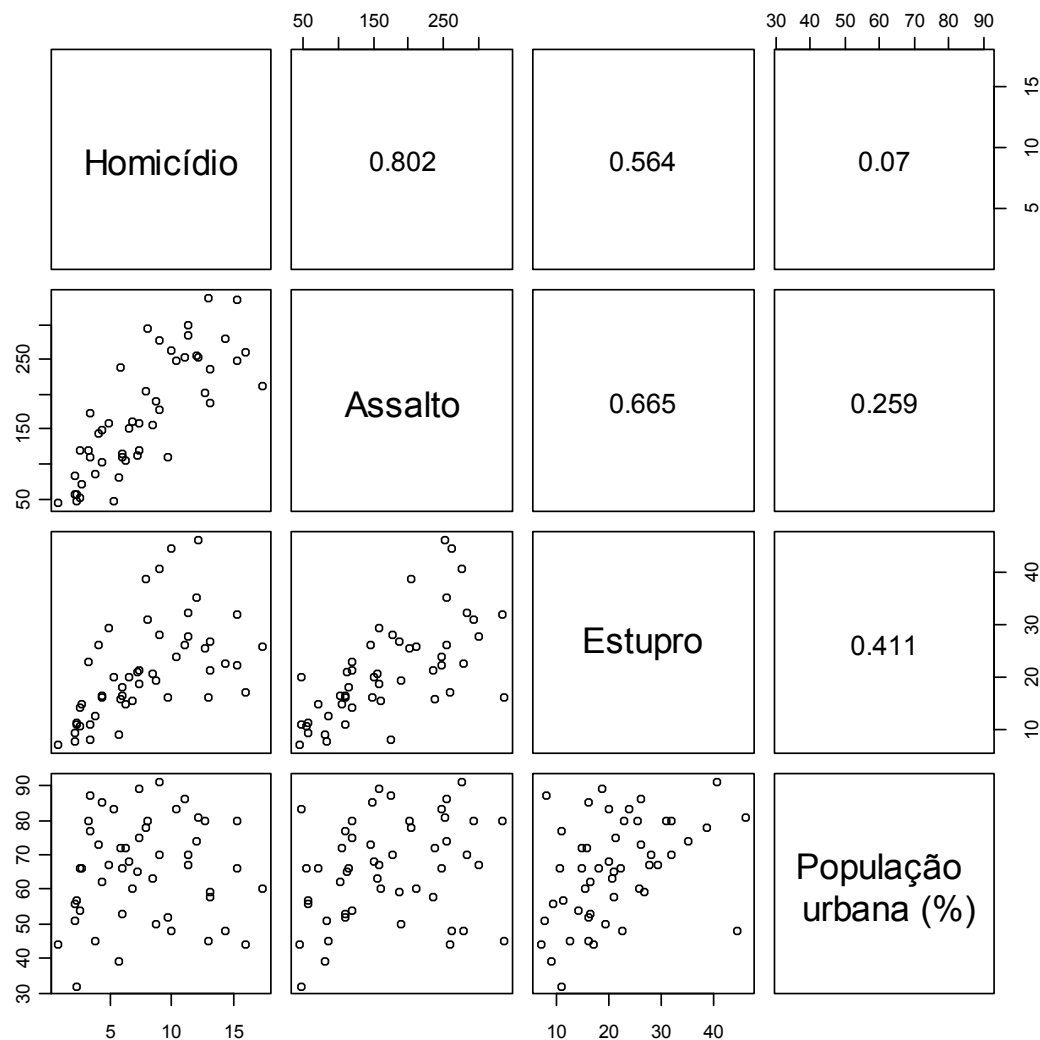
	Murder	Assault	Rape	UrbanPop
Murder	1.00000000	0.8018733	0.5635788	0.06957262
Assault	0.80187331	1.0000000	0.6652412	0.25887170
Rape	0.56357883	0.6652412	1.0000000	0.41134124
UrbanPop	0.06957262	0.2588717	0.4113412	1.00000000

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

1. Coeficiente de correlação

```
> 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)
}

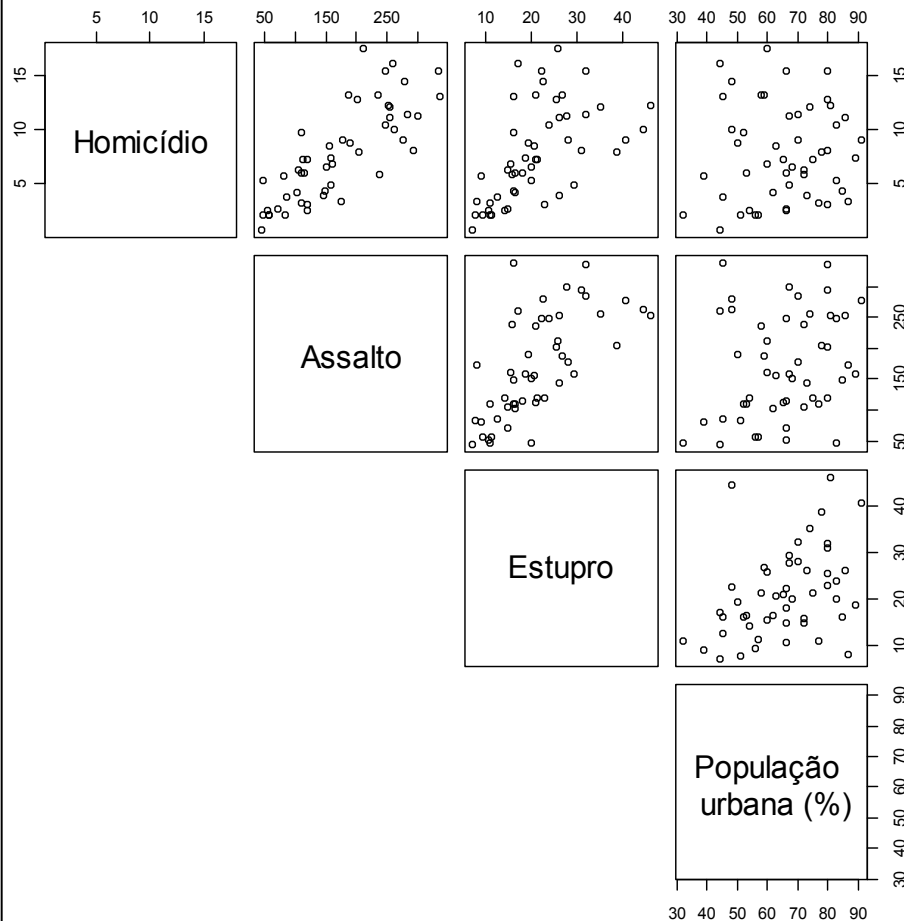
> pairs(USArrests[, ordem],
  labels = nomes,
  upper.panel = panel.cor)
```



1. Coeficiente de correlação

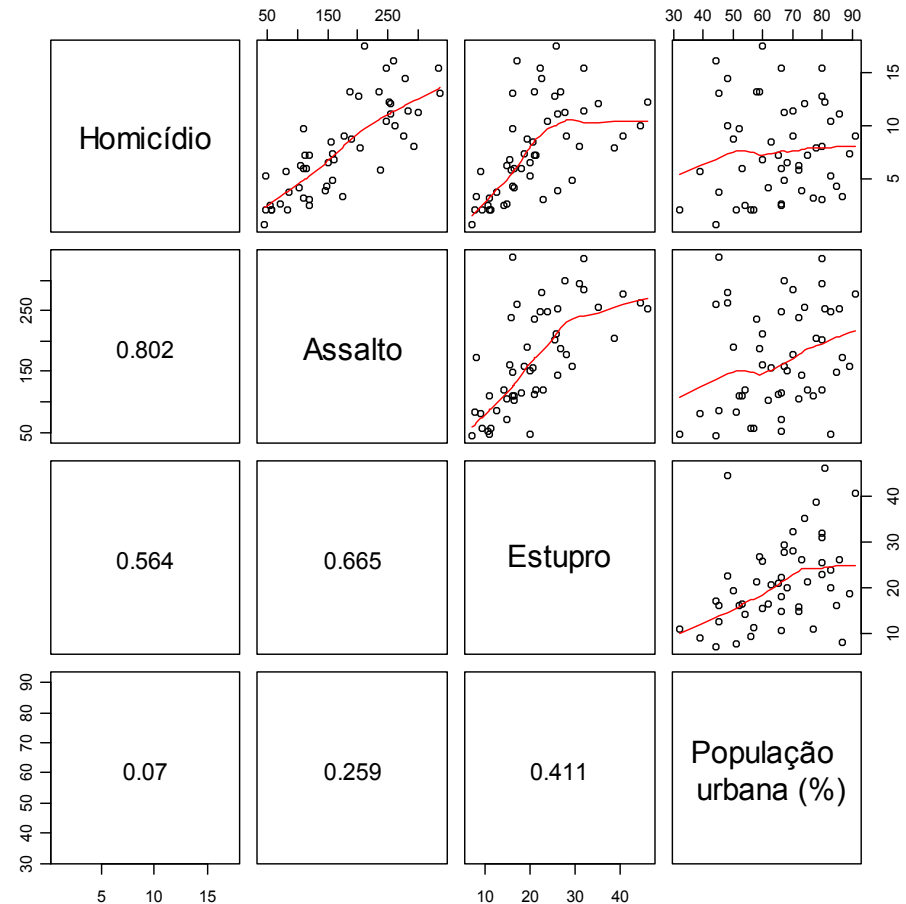
Omitindo a parte inferior da matriz:

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



Correlações e linhas de tendência:

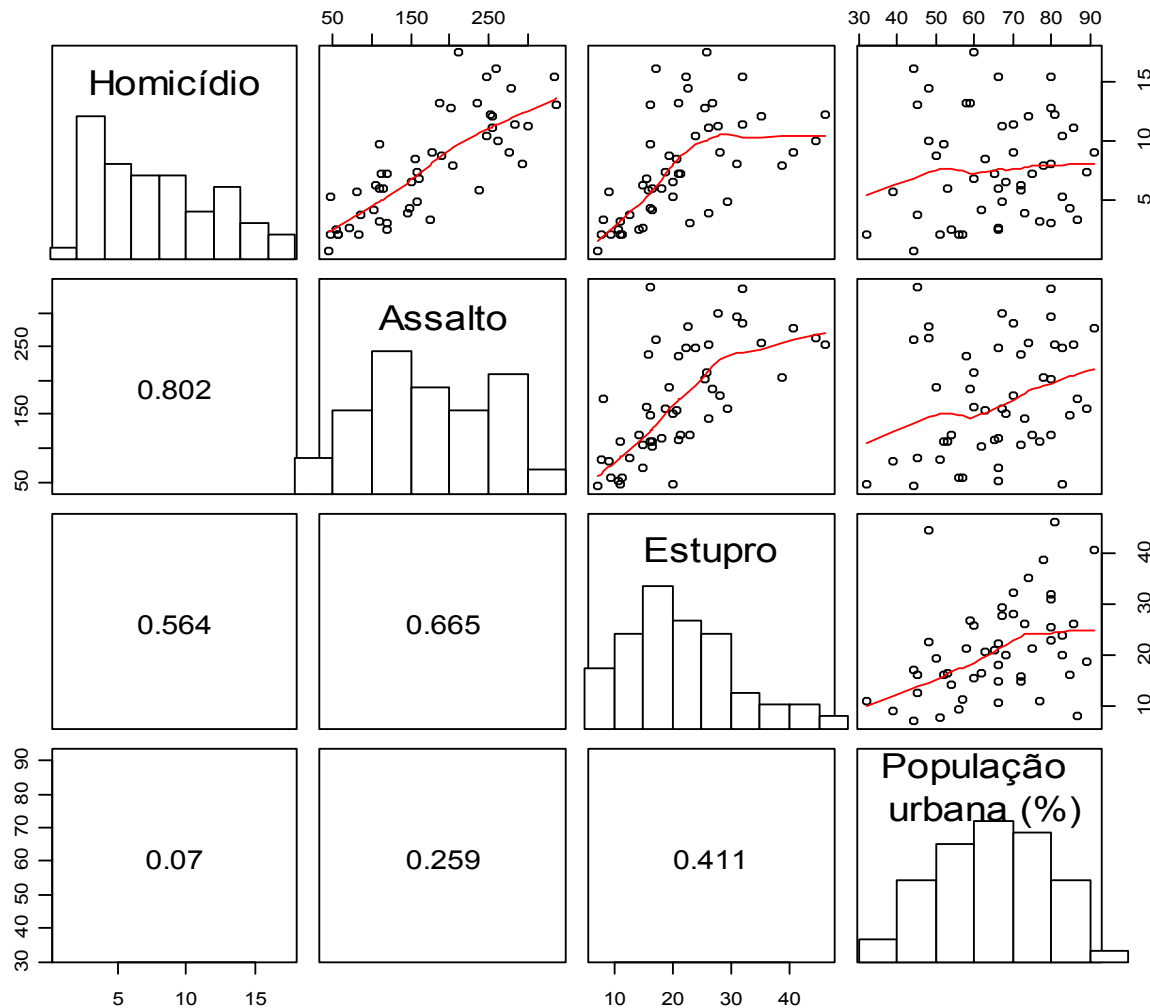
```
> pairs(USArrests[, ordem],  
labels = nomes, upper.panel =  
panel.smooth, lower.panel =  
panel.cor)
```



1. Coeficiente de correlação

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?

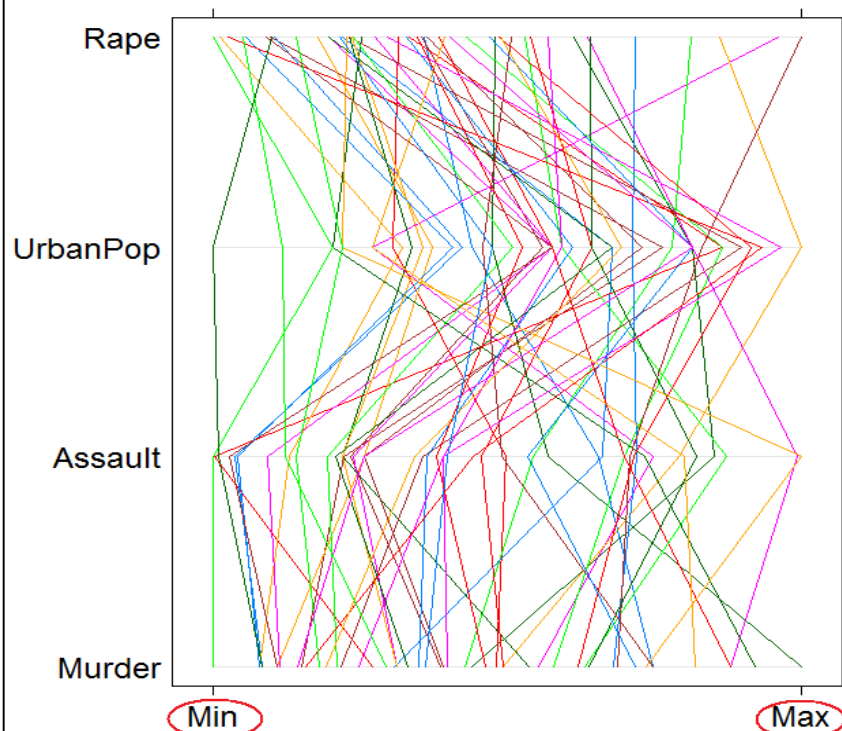
Que outras variáveis poderiam estar relacionadas à ocorrência dos crimes?

2. Gráficos

Função `parallel` (`lattice`): gráfico de coordenadas paralelas.

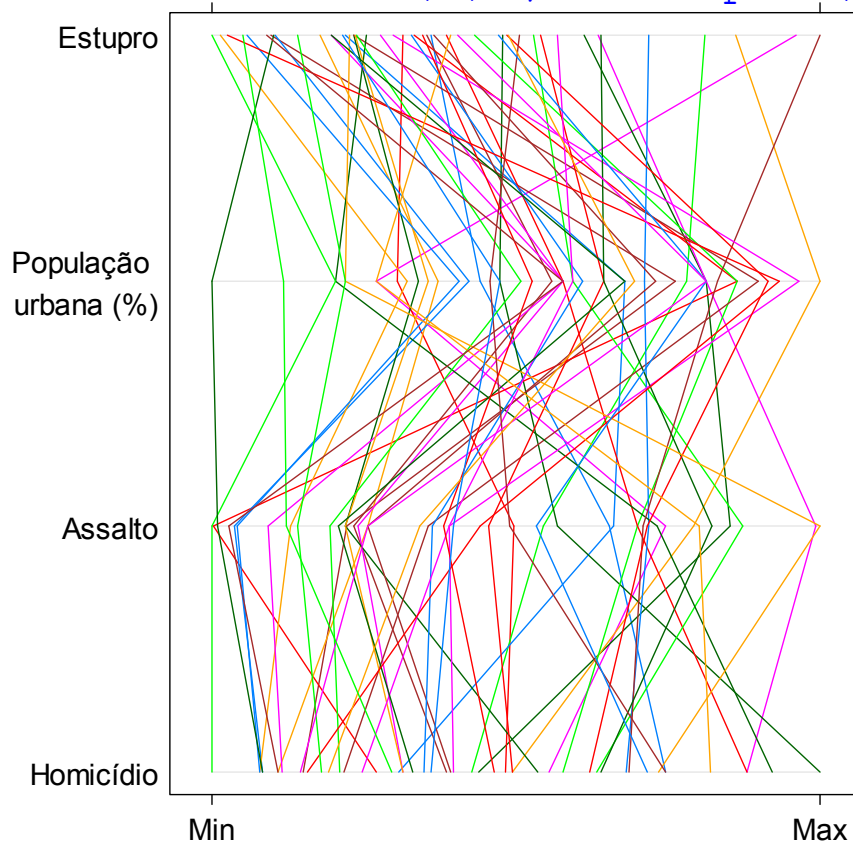
$p - 1$ segmentos de **retas** para cada observação unindo os valores escalonados em **[Min, Max]** para cada variável.

```
> parallel(USArrests)
```



Podem ser úteis para identificar grupos de observações (*cluster analysis*).

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



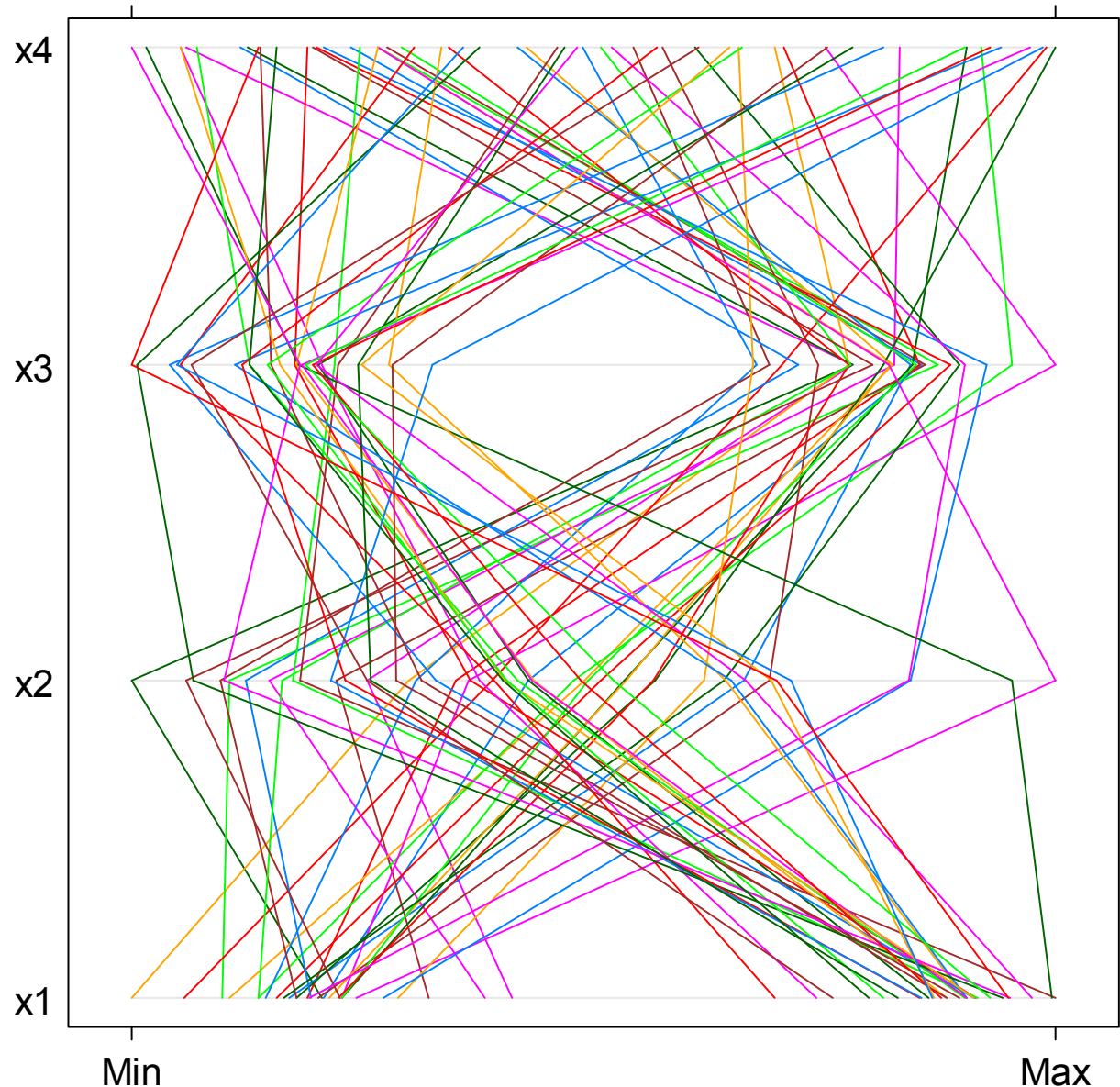
2. Gráficos

As variáveis x_1 e x_3 separam as observações em dois grupos.

Em um dos grupos os valores de x_1 são os menores e os valores de x_3 são os maiores.

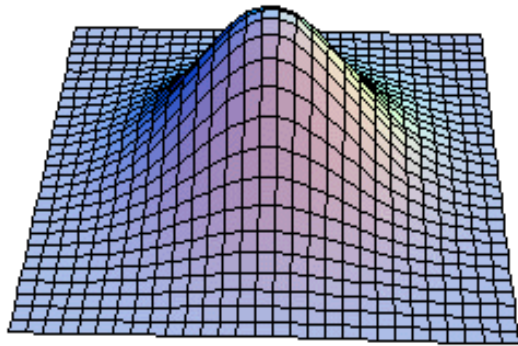
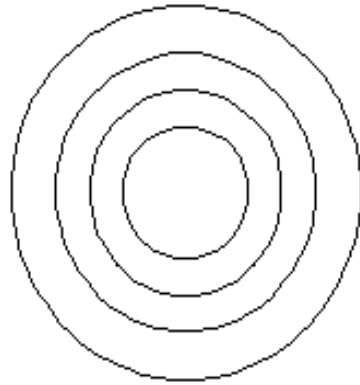
No outro grupo há uma inversão.

As variáveis x_2 e x_4 não permitem uma separação tão nítida quanto x_1 e x_3 .

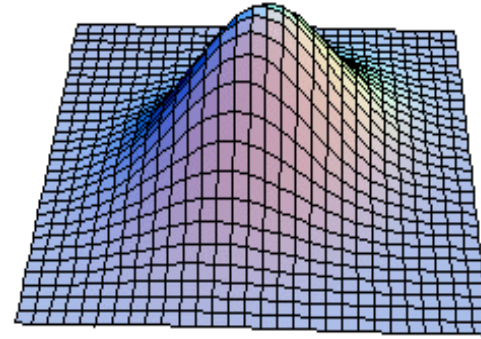
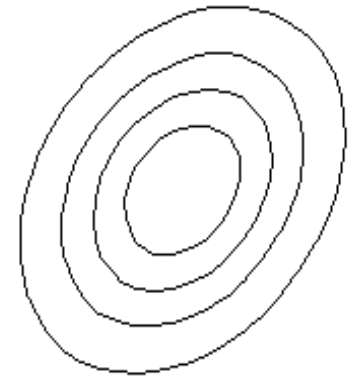


3.1. Distribuição normal bivariada

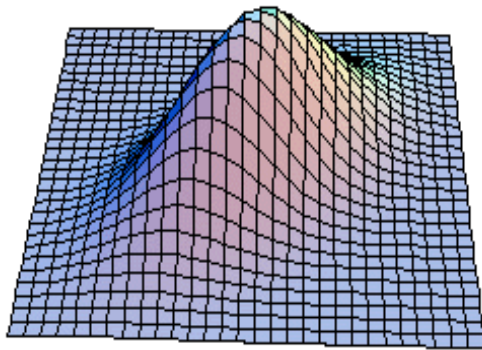
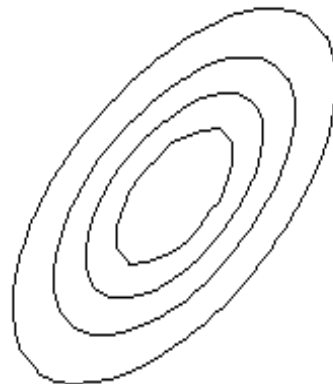
Correlação = 0 (independência)



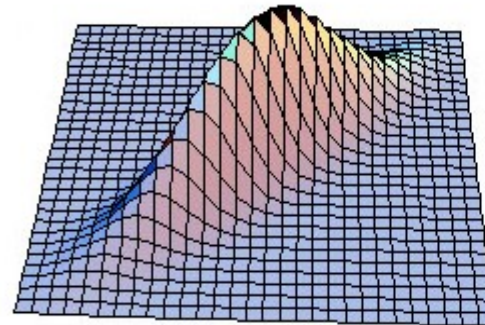
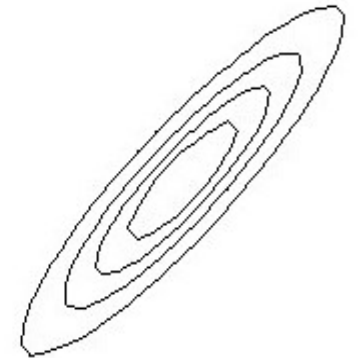
Correlação = 0,3



Correlação = 0,6



Correlação = 0,9



4. Pontos aberrantes

