
SME0803 Visualização e Exploração de Dados

# Representação de dados multidimensionais

Prof. Cibele Russo

cibele@icmc.usp.br

Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

Dados: $x_i$, i = 1,...,n, vetores $p \times 1$ ($p \geq 2$) cujos componentes podem ser p variáveis qualitativas, p variáveis quantitativas ou de ambos os tipos.

Problema central. Existe algum tipo de relação entre as variáveis?

p variáveis quantitativas: matriz de gráficos de dispersão.

p variáveis qualitativas: tabelas de contingência multidimensionais e gráficos de mosaico.

Utilizaremos os gráficos em grade (*trellis plots*) em R (pacote `lattice`).

Sintaxe baseada em fórmulas.

Exemplos. (1) **`var1 ~ var2 | var3 + var4 + var5`**
(2) **`~ var1 | var2 + var3`**

A barra vertical (|) indica condicionamento. O sinal "+" não é adição.

Em (1), **`var1`** é a variável dependente e **`var2`** é a variável independente.

Todas as combinações de (**`var3, var4, var5`**) são consideradas na relação **`var2`** $\rightarrow$ **`var1`**.

Em (2), não há variável dependente. Todas as combinações de (**`var2, var3`**) são consideradas.

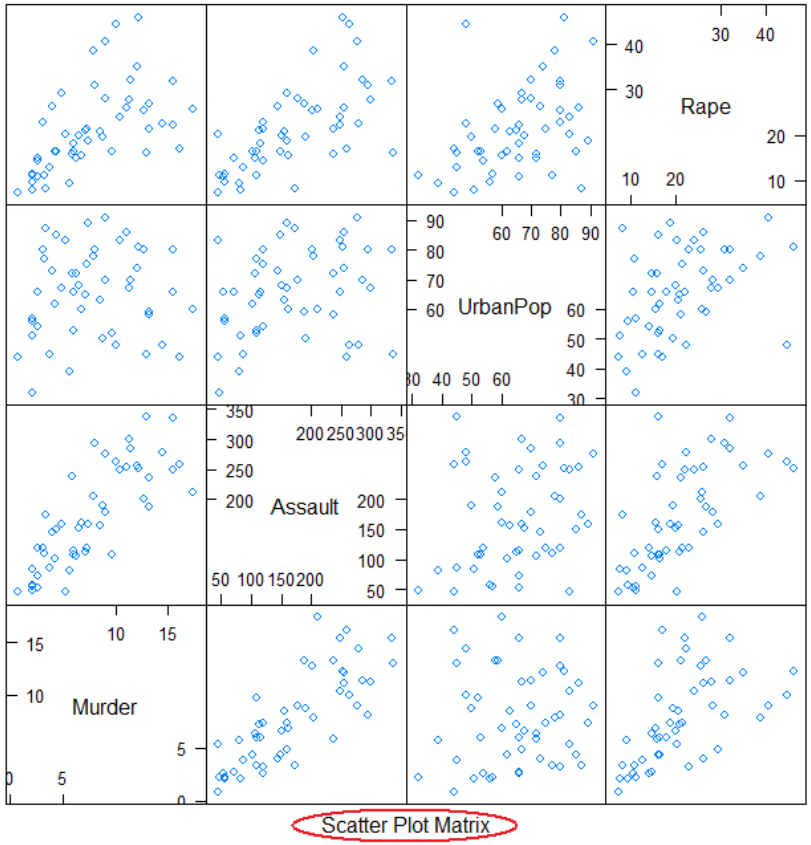
Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

# Variáveis quantitativas

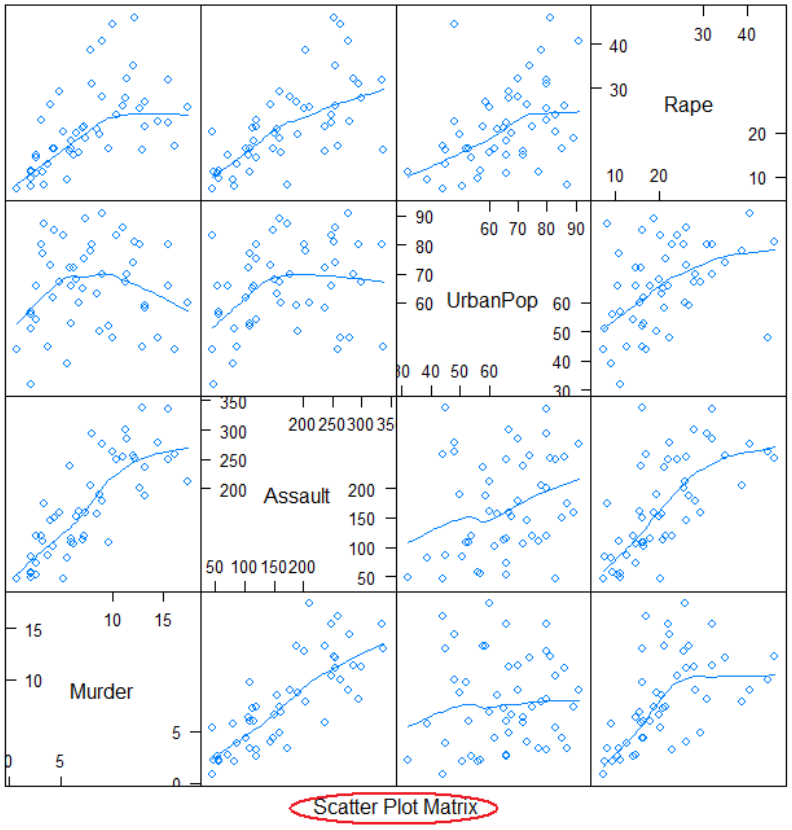
Função `splom` (`lattice`): matriz de gráficos de dispersão (scatter plot matrix).

Dados USArrests (Seção 8.1).

```
> library(lattice)
> splom(USArrests)
```




```
> splom(USArrests, type
= c("p", "smooth"))
```

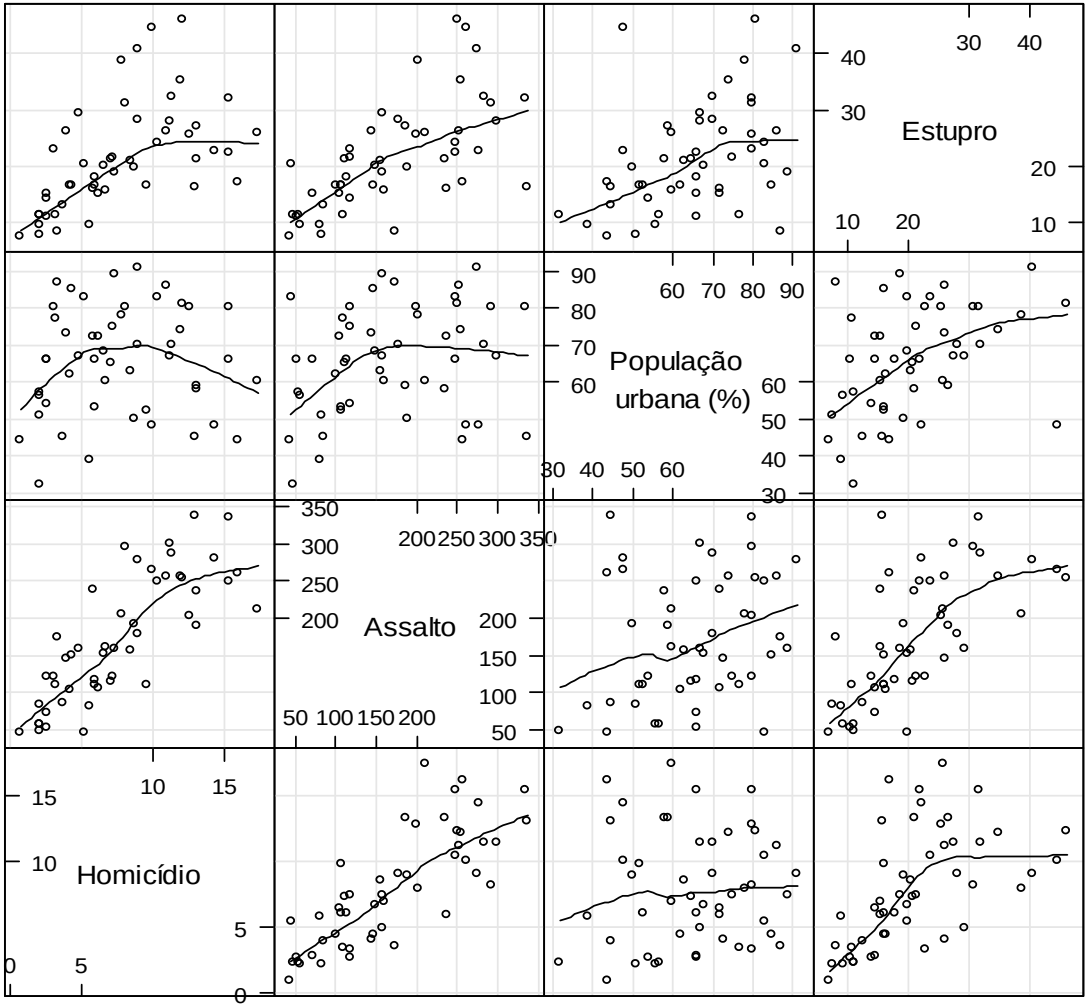



Gráficos com pontos (p) e linhas de tendência (`smooth`).

Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

# Variáveis quantitativas

```
> splom(USArrests, type = c("g", "p", "smooth"), col =
"black", xlab = "", varnames = c("Homicídio",
"Assalto", "População \n urbana (%)", "Estupro"))
```




Gráficos com reticulados (g), pontos (p) e linhas de tendência (`smooth`).

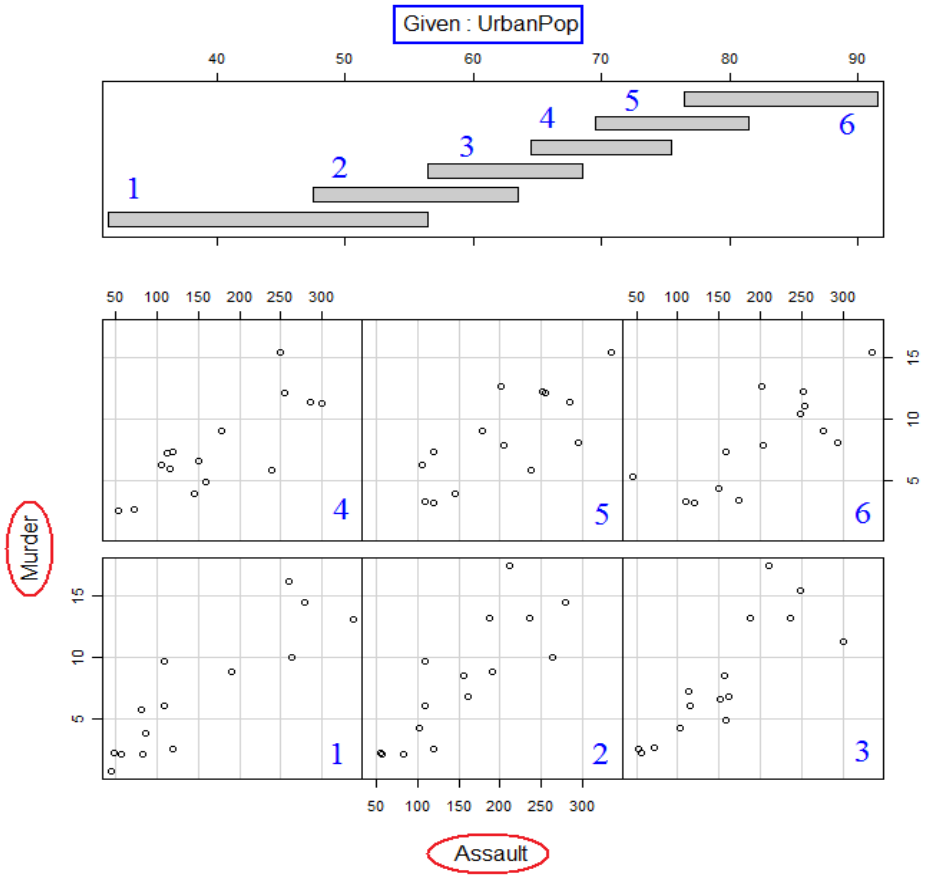
Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

# Variáveis quantitativas

Gráficos condicionais (*conditional plots*): gráfico de dispersão de ($x_1$, $x_2$) para faixas de valores de outras variáveis quantitativas.

Funções `coplot` (`graphics`) e `xyplot` (`lattice`).

```
> attach(USArrests)
> coplot(Murder ~ Assault | UrbanPop)
```




Por *default*, são criadas seis faixas com aproximadamente o mesmo número de observações da variável condicionante e com superposição (*overlap*) de 50% (estes argumentos podem ser mudados).

Ver

```
> co.intervals(UrbanPop,
number = 6, overlap = 0.5)
```

Os painéis são dispostos a partir do canto inferior esquerdo.

Permite avaliar se a relação entre $x_1$ e $x_2$ depende de valores de outra(s) variável(is).

Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

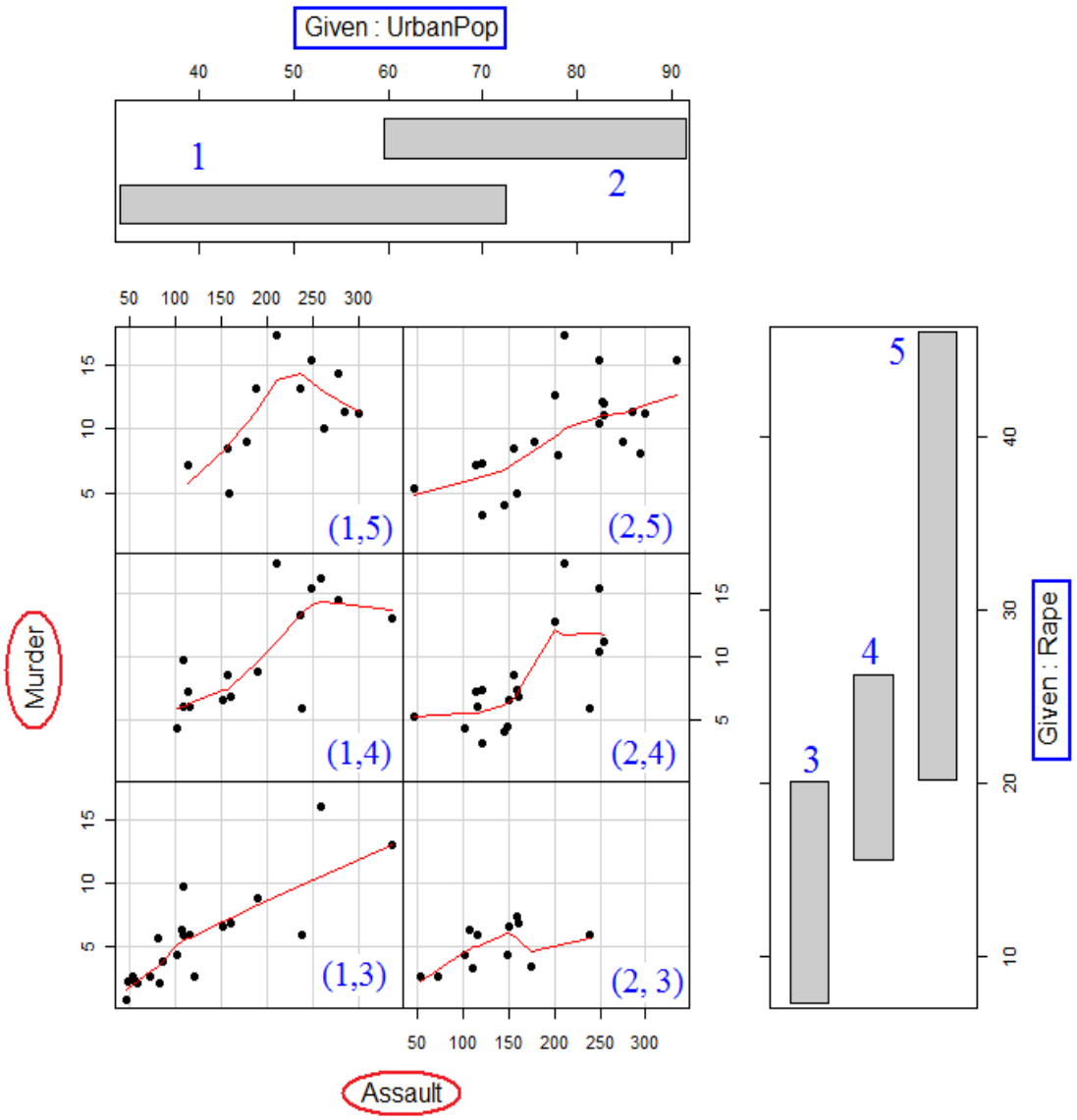
# Variáveis quantitativas

Duas variáveis condicionantes:

`UrbanPop` e `Rape`.

Número de intervalos (faixas) é diferente para cada variável condicionante.

```
> coplot(Murder ~ Assault |
UrbanPop * Rape, number =
c(2, 3), pch = 20, cex =
1.5, panel = panel.smooth)
```




Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

# Variáveis quantitativas

`UrbanPop` com três intervalos de igual comprimento.

```
> xyplot(Murder ~ Assault | cut(UrbanPop, 3))
```

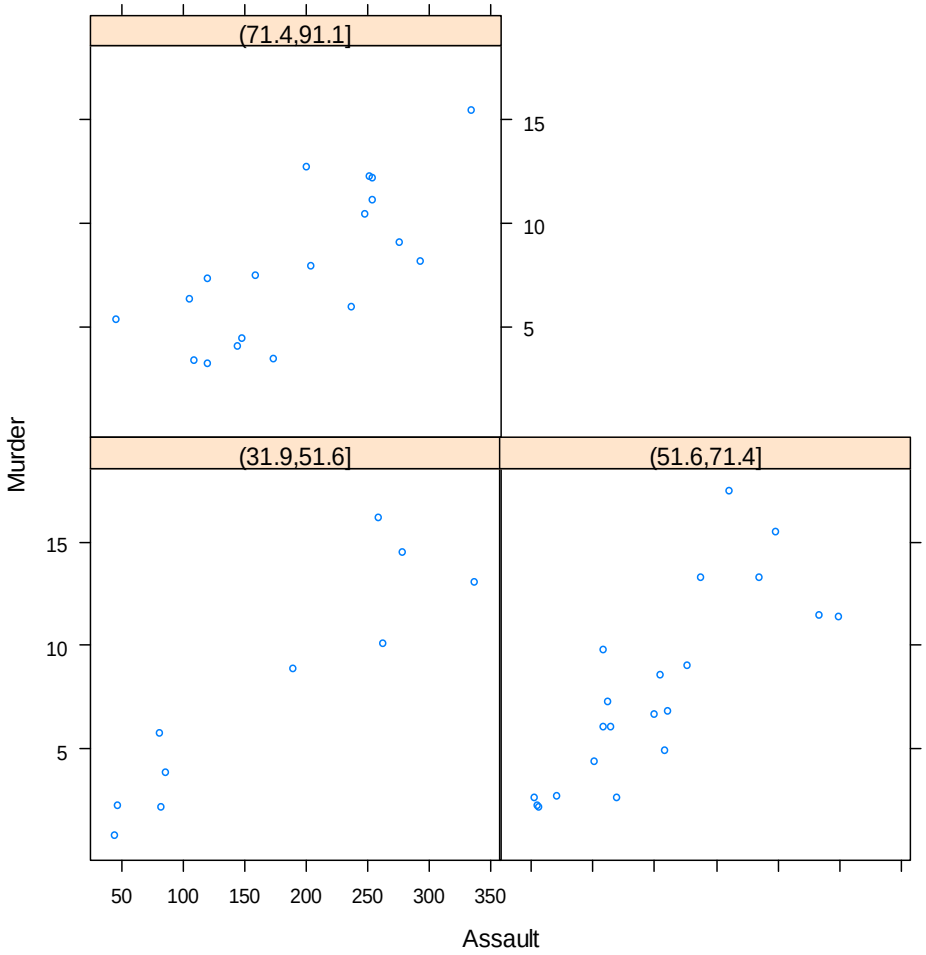
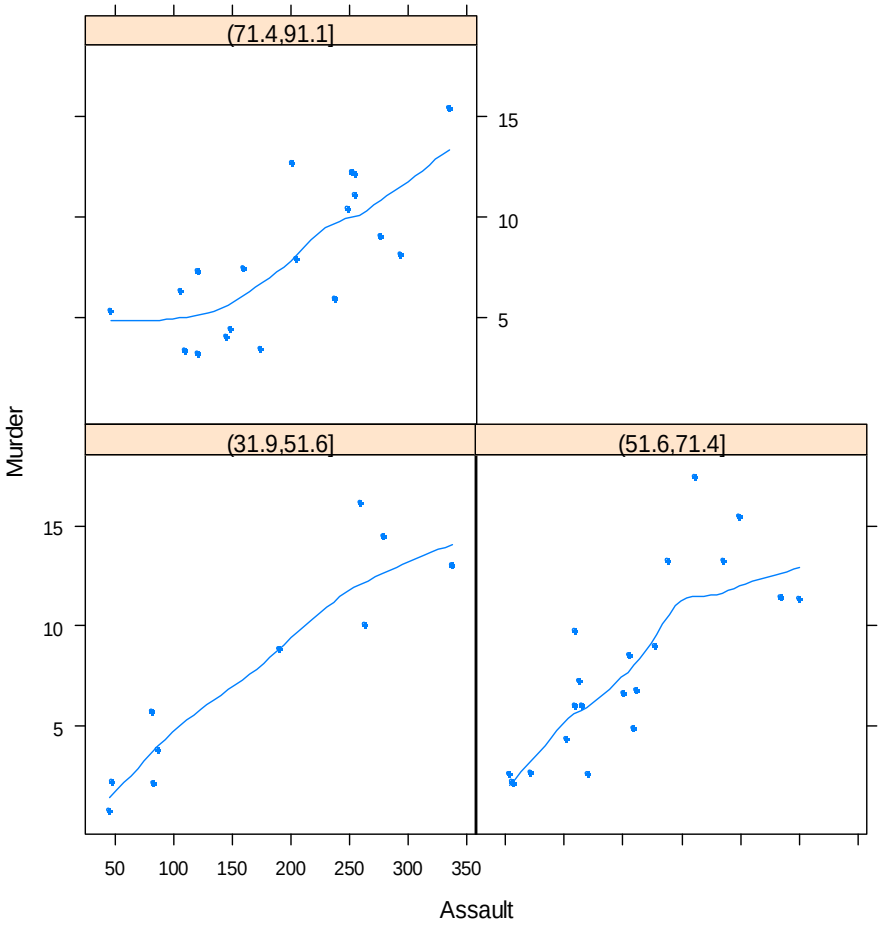



Gráfico com pontos (p) e linhas de tendência (`smooth`)

```
> xyplot(Murder ~ Assault | cut(UrbanPop, 3), type = c("p", "smooth"), pch = 20)
```

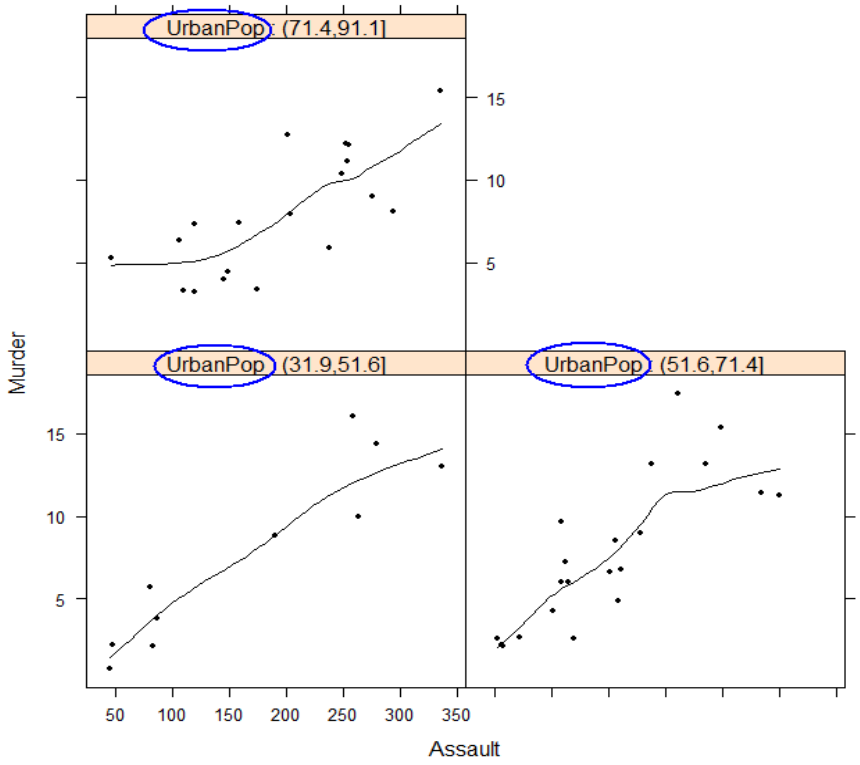



Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

# Variáveis quantitativas

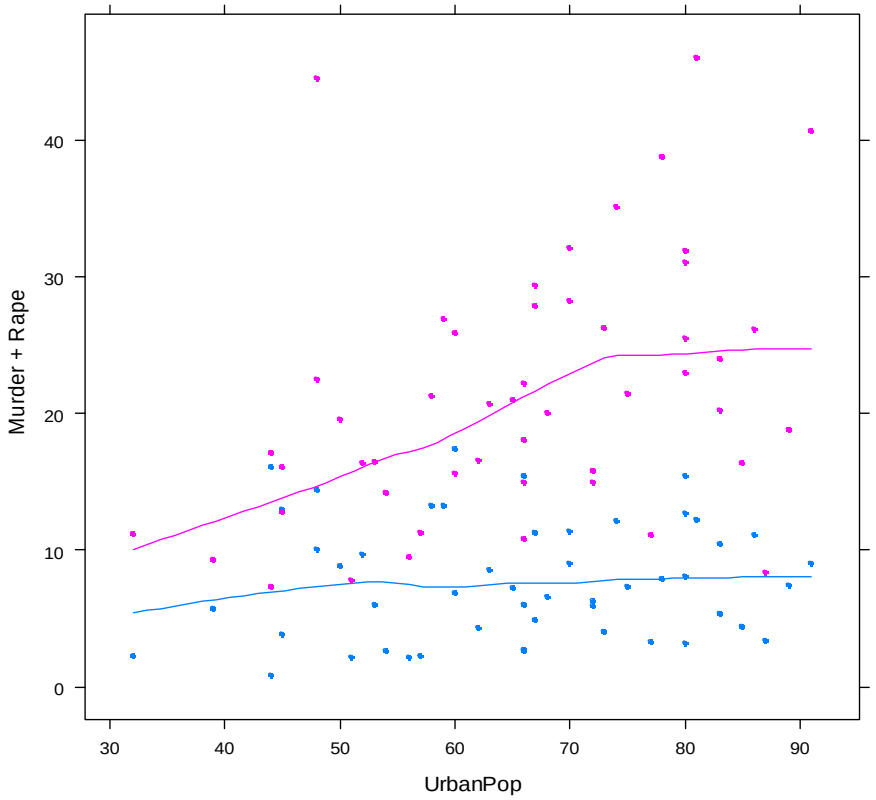
Inclusão do nome da variável condicionante nos painéis

```
> xyplot(Murder ~ Assault | cut(UrbanPop, 3), type = c("p", "smooth"), pch = 20, strip.custom(strip.names = TRUE, var.name = "UrbanPop"))
```




Duas variáveis dependentes, sem variável condicionante

```
> xyplot(Murder + Rape ~UrbanPop, type = c("p", "smooth"), pch =20)
```




Obs. “+” não significa adição.

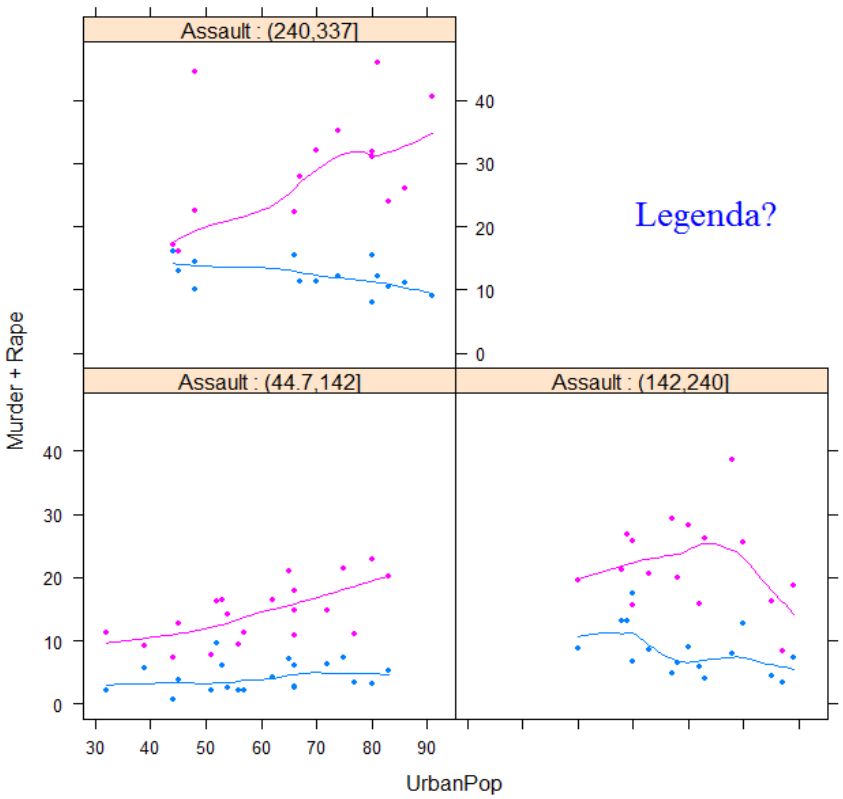
Exercício. Incluir uma legenda.

Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

# Variáveis quantitativas

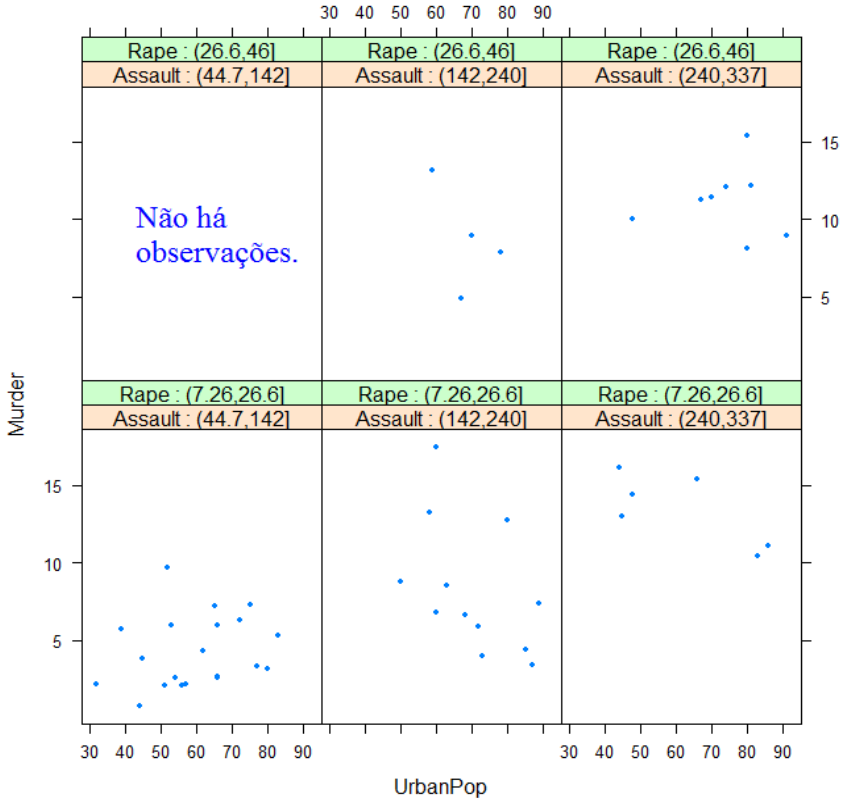
## Duas variáveis dependentes e uma variável condicionante

```
>  xyplot(Murder + Rape ~UrbanPop | cut(Assault, 3), type = c("p", "smooth"),   pch = 20, strip = strip.custom(strip.names = TRUE, var.name = "Assault"))
```




## Duas variáveis condicionantes

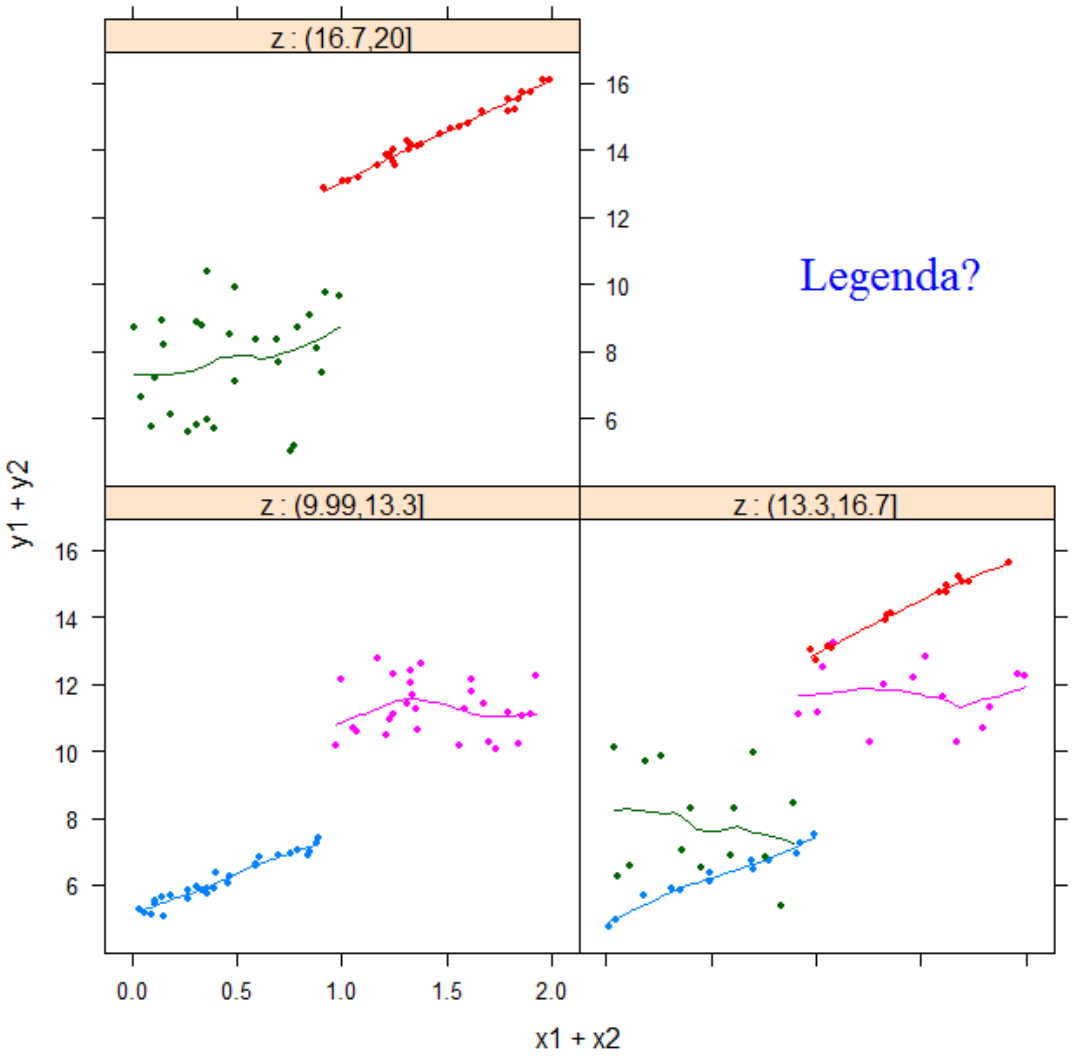
```
>  xyplot(Murder ~  UrbanPop | cut(Assault, 3) + cut(Rape, 2), pch = 20, strip = strip.custom(strip.names = TRUE, var.name = c("Assault", "Rape")))
```




Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

# Variáveis quantitativas

Duas variáveis dependentes, duas variáveis independentes e uma variável condicionante (cinco variáveis)




Obs. (1) Quatro cores correspondem aos quatro pares de variáveis (x, y).

Neste exemplo, em cada painel podemos ter até quatro gráficos de dispersão.

(2) Em uma fórmula, se quisermos somar variáveis (e se fizer sentido), utilizamos

`I(x1 + x2)` e/ou

`I(y1 + y2)`.

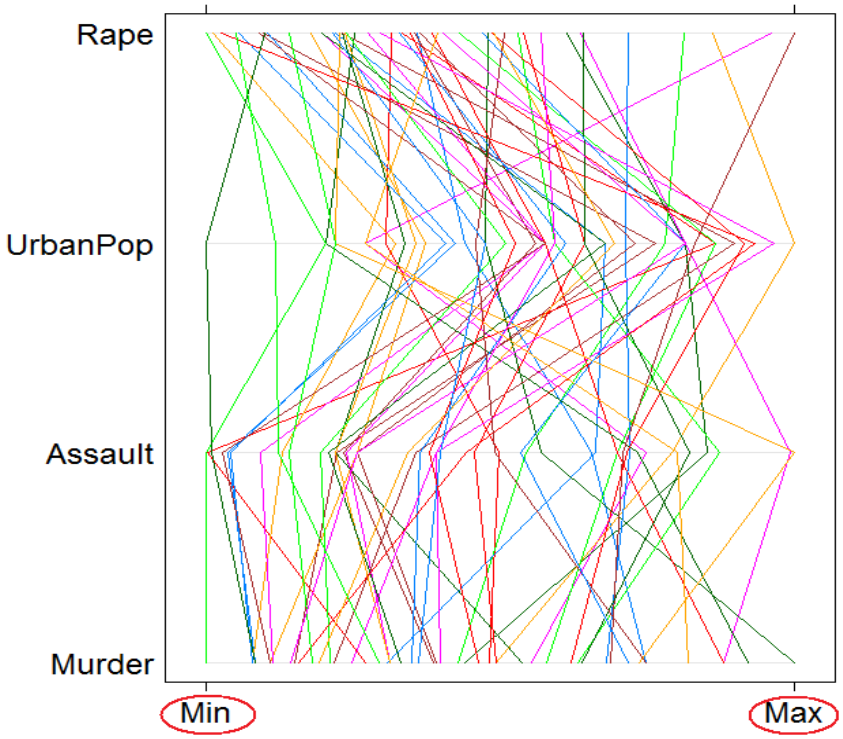
Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

# Variáveis quantitativas

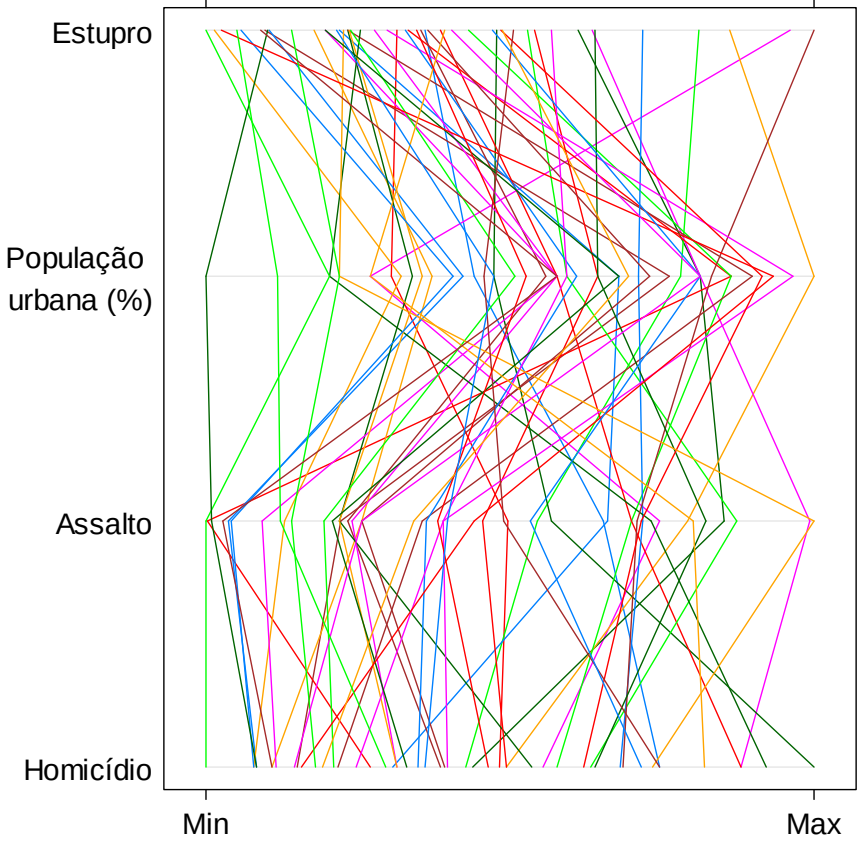
Função `parallel` (`lattice`): gráfico de coordenadas paralelas.

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

```
> parallel(USArrests)
```




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




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

Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

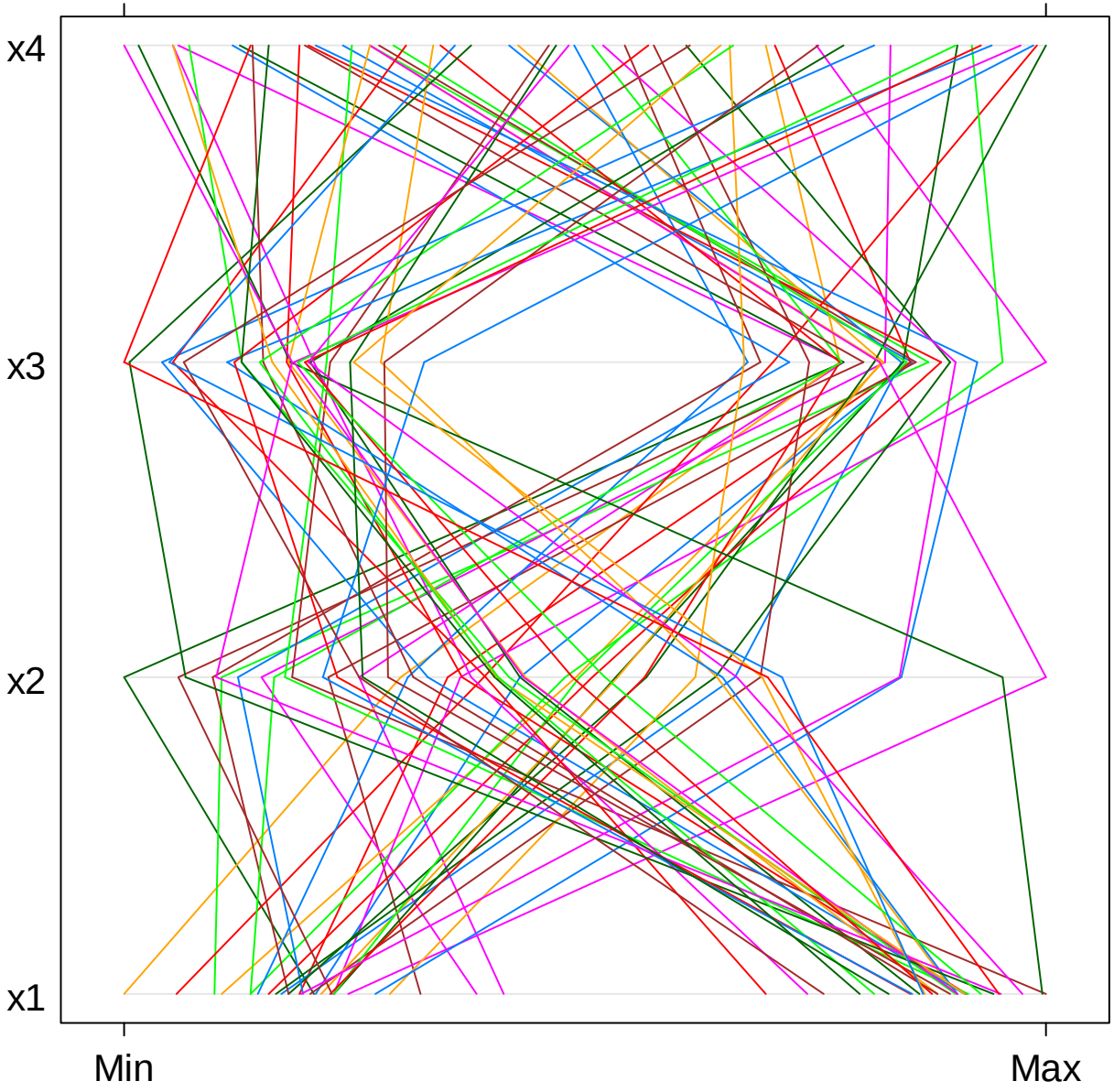
# Variáveis quantitativas

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 $x_3$.




Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

# Variáveis qualitativas

Dados `Ilocos` da Seção 8.2.

```
> library(ineq)
> data(Ilocos)
> dados = Ilocos
> attach(dados)
> names(dados)
```

```
"income" "sex" "family.size" "urbanity" "province" "AP.income"
"AP.family.size" "AP.weight"
```

Função `ftable`: tabela de contingências multidimensional.

```
> (tab3 =
ftable(urbanity,
province, sex))
```

tab3

| urbanity | province | sex female | male |
|---|---|---|---|
| rural | Ilocos Norte | 5 | 42 |
| | Ilocos Sur | 9 | 36 |
| | La Union | 9 | 62 |
| | Pangasinan | 18 | 120 |
| urban | Ilocos Norte | 3 | 15 |
| | Ilocos Sur | 9 | 14 |
| | La Union | 9 | 36 |
| | Pangasinan | 52 | 193 |

```
> tab3rel =
prop.table(tab3, margin
= 1)
```

```
> (tab3relp = tab3rel *
100)
```

tab3rel

| urbanity | province | sex female | male |
|---|---|---|---|
| rural | Ilocos Norte | 10.63830 | 89.36170 |
| | Ilocos Sur | 20.00000 | 80.00000 |
| | La Union | 12.67606 | 87.32394 |
| | Pangasinan | 13.04348 | 86.95652 |
| urban | Ilocos Norte | 16.66667 | 83.33333 |
| | Ilocos Sur | 39.13043 | 60.86957 |
| | La Union | 20.00000 | 80.00000 |
| | Pangasinan | 21.22449 | 78.77551 |

Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

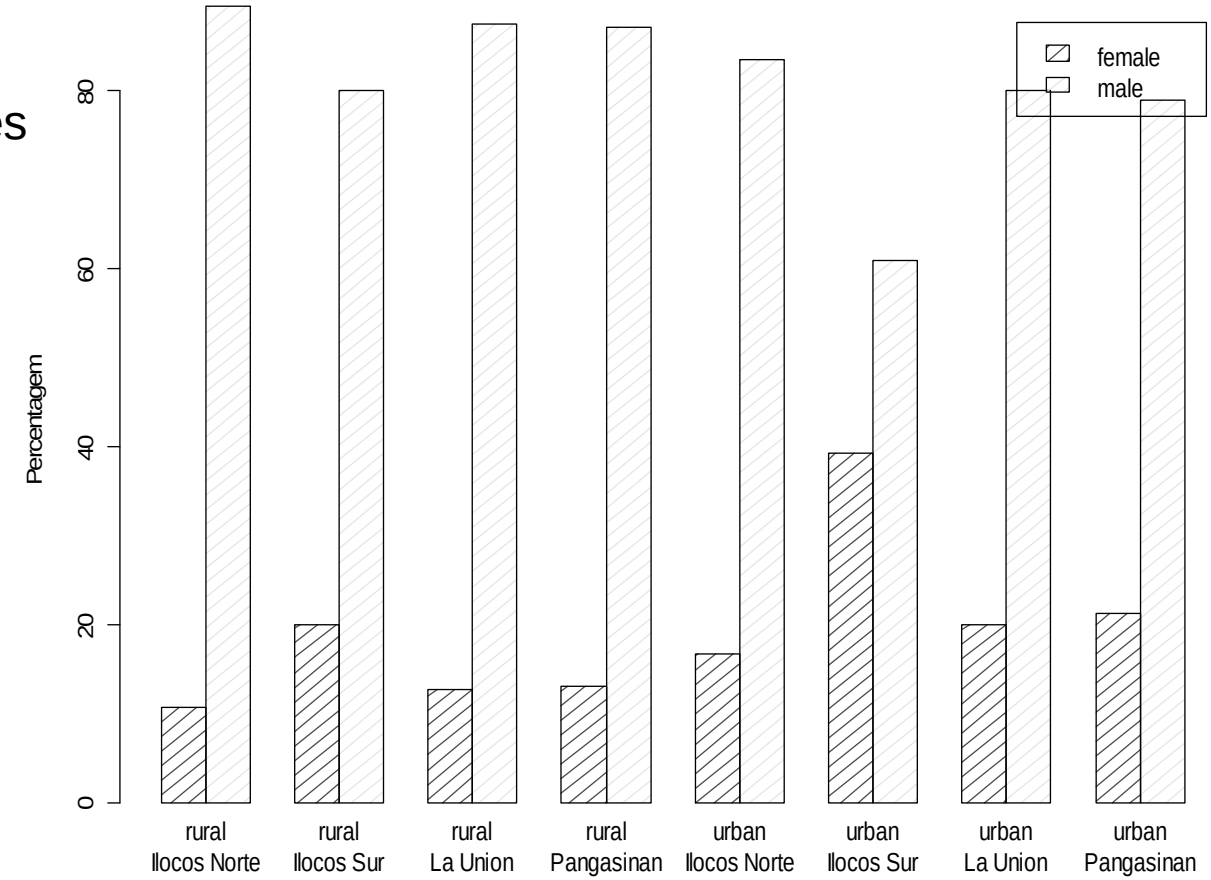
# Variáveis qualitativas

Gráfico de barras

```
> rownames(tab3relp) = paste(rep(levels(urbanity), each =
length(levels(province))), levels(province), sep = "\n")
> barplot(t(tab3relp), beside = TRUE, legend = levels(sex), density
= 15, ylab = "Percentagem")
> box()
```

Gráfico com as distribuições condicionais de `sex` | (`urbanity`, `province`).

Exercício. Apresentar os rótulos e a legenda em português.




Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

# Variáveis qualitativas

Função `xtabs`: tabelas multidimensionais utilizando uma `formula`.

```
> (tab3var = xtabs(~ urbanity + province + sex))
```

```
, , sex = female

          province
urbanity Ilocos Norte Ilocos Sur La Union Pangasinan
   rural           5          9        9         18
   urban           3          9        9         52
, , sex = male

          province
urbanity Ilocos Norte Ilocos Sur La Union Pangasinan
   rural          42         36       62        120
   urban          15         14       36        193
```

As duas vírgulas indicam as outras duas variáveis.

```
> class(tab3var)
[1] "xtabs" "table"
```

Tabela na forma de uma folha de dados (*data frame*)

```
> as.data.frame(tab3var)
```

| | urbanity | province | sex | Freq |
|---|---|---|---|---|
| 1 | rural | Ilocos Norte | female | 5 |
| 2 | urban | Ilocos Norte | female | 3 |
| 3 | rural | Ilocos Sur | female | 9 |
| 4 | urban | Ilocos Sur | female | 9 |
| 5 | rural | La Union | female | 9 |
| 6 | urban | La Union | female | 9 |
| 7 | rural | Pangasinan | female | 18 |
| 8 | urban | Pangasinan | female | 52 |
| 9 | rural | Ilocos Norte | male | 42 |
| 10 | urban | Ilocos Norte | male | 15 |
| 11 | rural | Ilocos Sur | male | 36 |
| 12 | urban | Ilocos Sur | male | 14 |
| 13 | rural | La Union | male | 62 |
| 14 | urban | La Union | male | 36 |
| 15 | rural | Pangasinan | male | 120 |
| 16 | urban | Pangasinan | male | 193 |

Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

# Variáveis qualitativas

Gráfico de barras de `sex` com frequencias relativas ao par (`urbanity`, `province`).

Função `barchart` (`lattice`).

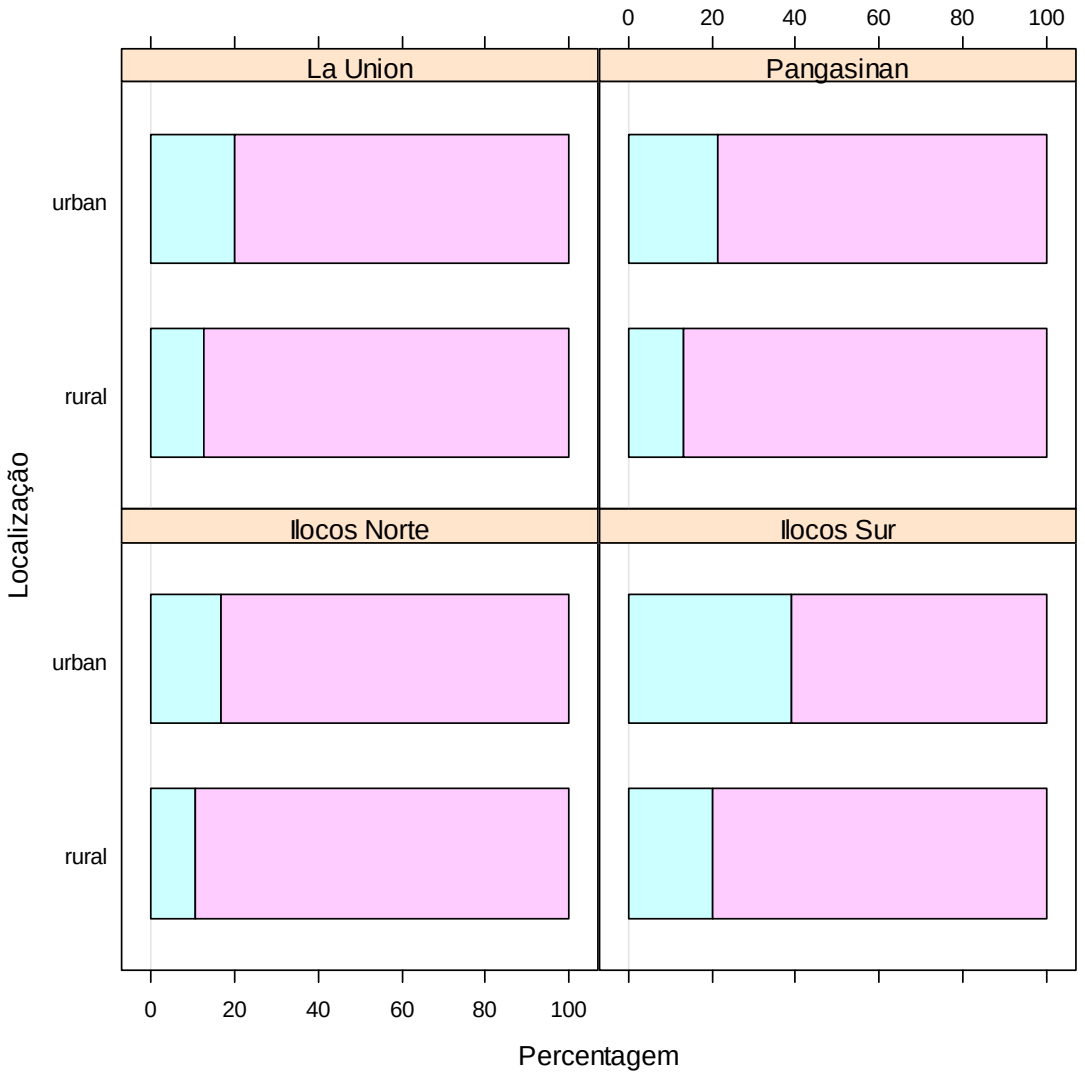
```
> barchart(prop.table( tab3var, margin = c(1, 2)) * 100, xlab ="Percentagem", ylab = "Localização")
```

Cada nível de `sex` com uma cor diferente.

Exercícios.

1. Mudar as cores e adicionar uma legenda.

2. Verificar o resultado da função `prop.table`.




Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

# Variáveis qualitativas

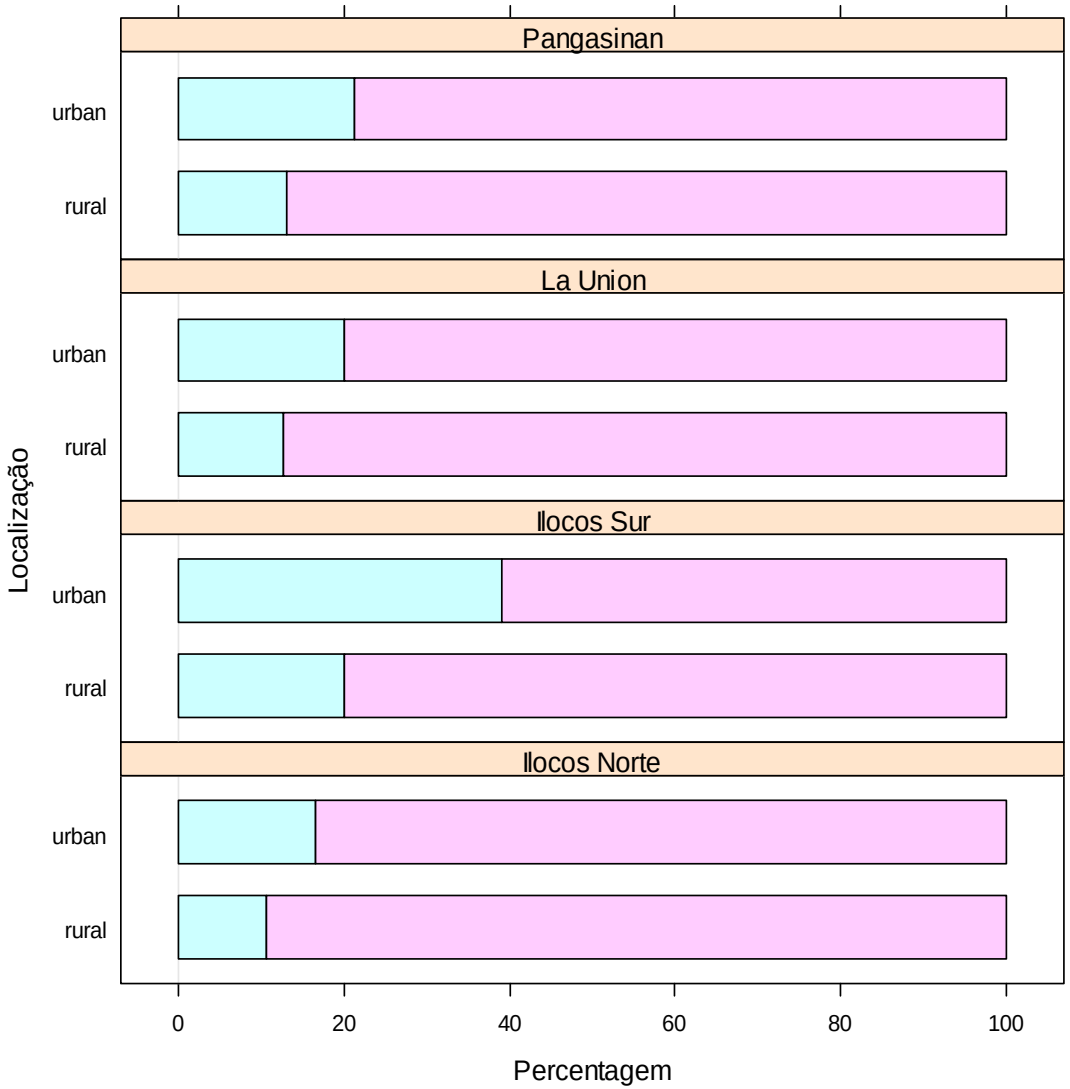
Gráfico de barras de `sex` com frequencias relativas ao par (`urbanity`, `province`).

Níveis de `province` empilhados.

```
> barchart(prop.table( tab3var, margin = c(1, 2)) * 100, xlab = Percentagem",  ylab = "Localização", layout = c(1, 4))
```

Exercício. Compare com o gráfico do slide anterior.

O que pode ser afirmado sobre a associação entre as variáveis?




Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

# Variáveis quantitativas e qualitativas

Dados `Ilocos`

```
> names(dados)
```

```
 y       x       y                x          x
"income" "sex" "family.size" "urbanity" "province" "AP.income"
"AP.family.size" "AP.weight"
```

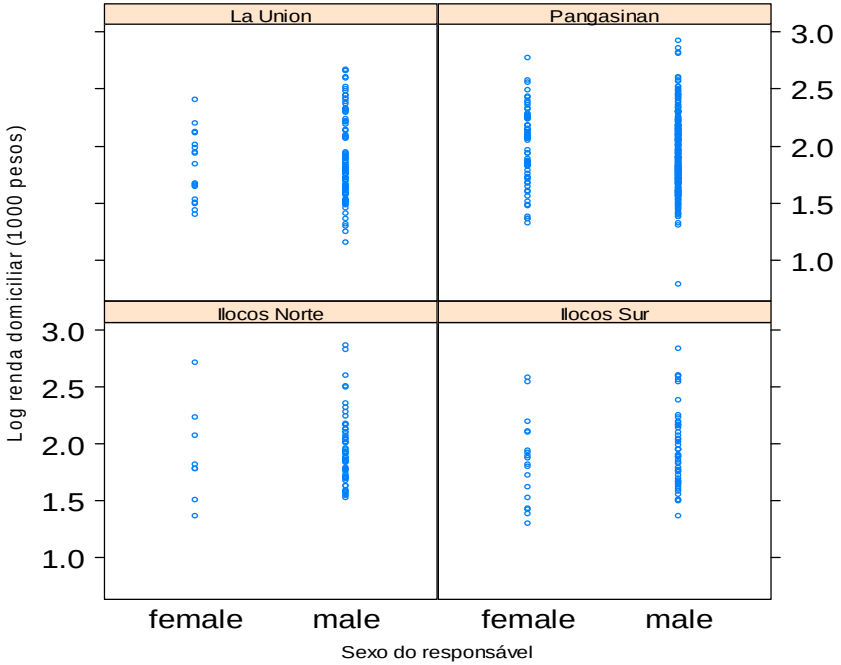
Gráfico de pontos

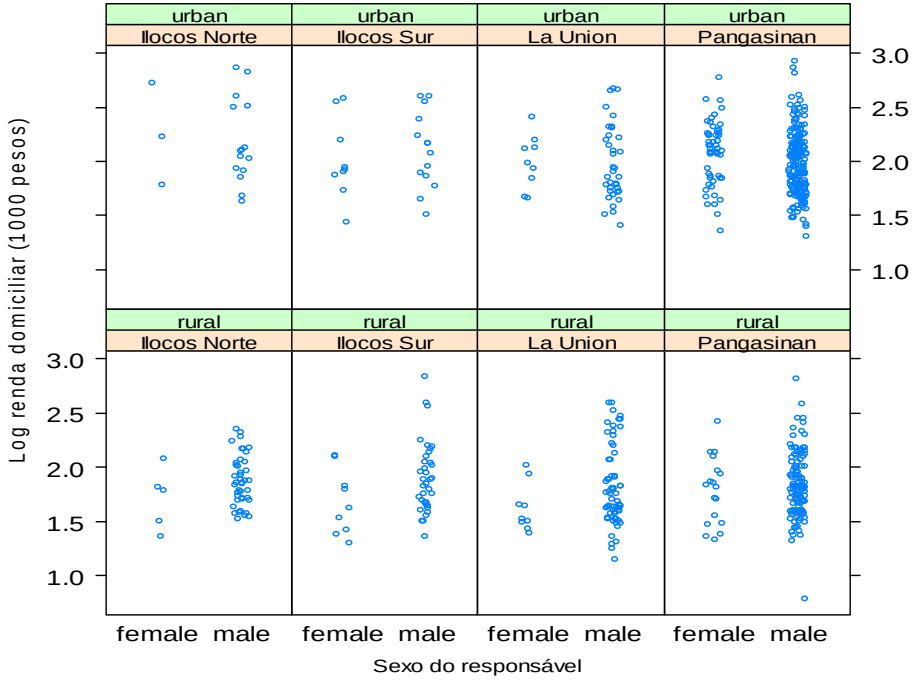
Função `stripplot` (`lattice`)

```
> stripplot(log(income / 1000, 10) ~ sex | province, xlab = "Sexo do responsável", ylab = "Log renda domiciliar (1000 pesos)")
```

Duas variáveis condicionantes e acréscimo de ruído

```
> stripplot(log(income / 1000, 10) ~ sex | province + urbanity, xlab = "Sexo do responsável", ylab = "Log renda domiciliar (1000 pesos)", jitter.data = TRUE)
```



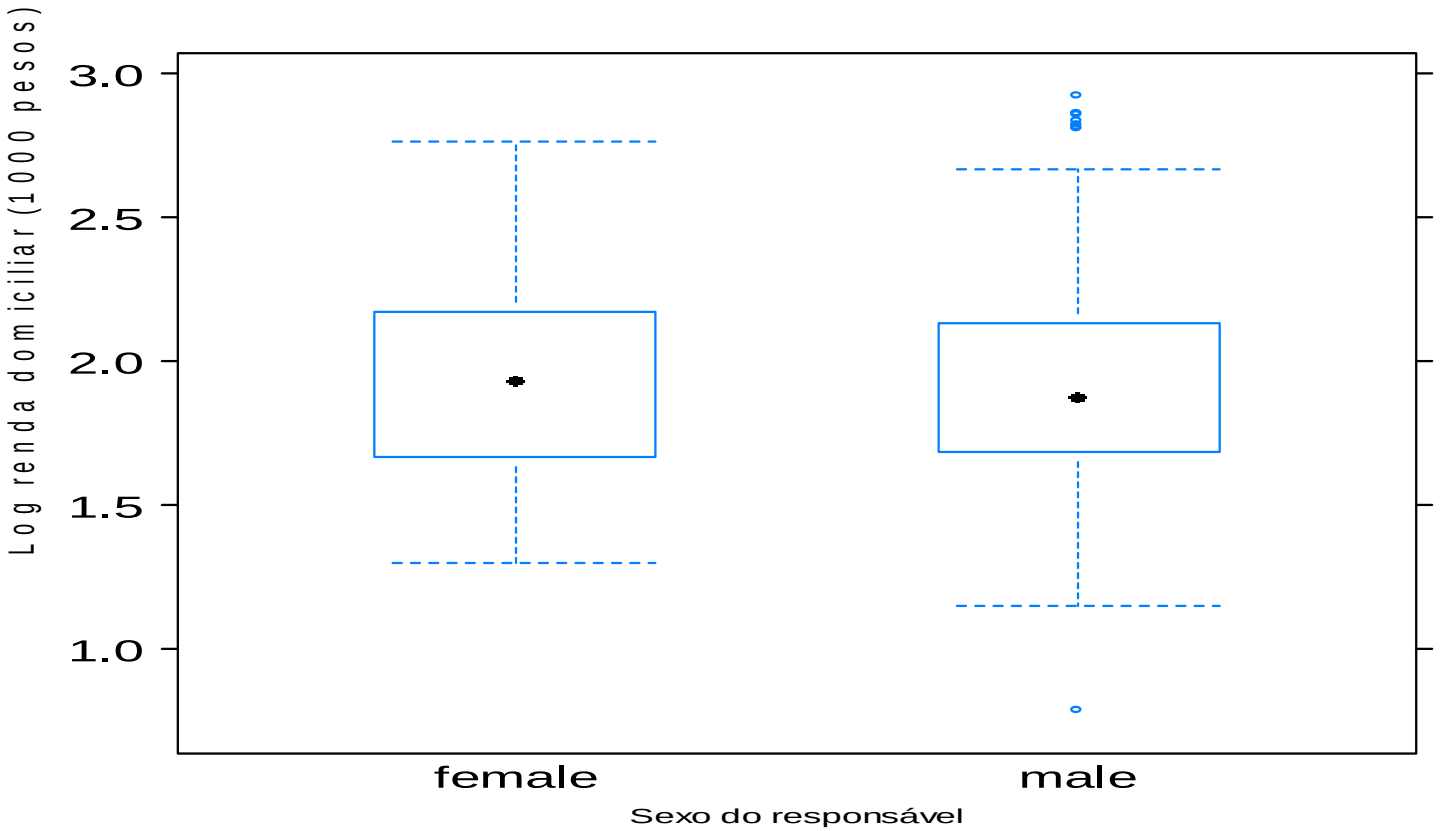

Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

# Variáveis quantitativas e qualitativas

Gráfico de caixas

Função `bwplot` (`lattice`)

```
> bwplot(log(income / 1000, 10) ~ sex, xlab = "Sexo do
responsável",  ylab = "Log renda domiciliar (1000
pesos)")
```

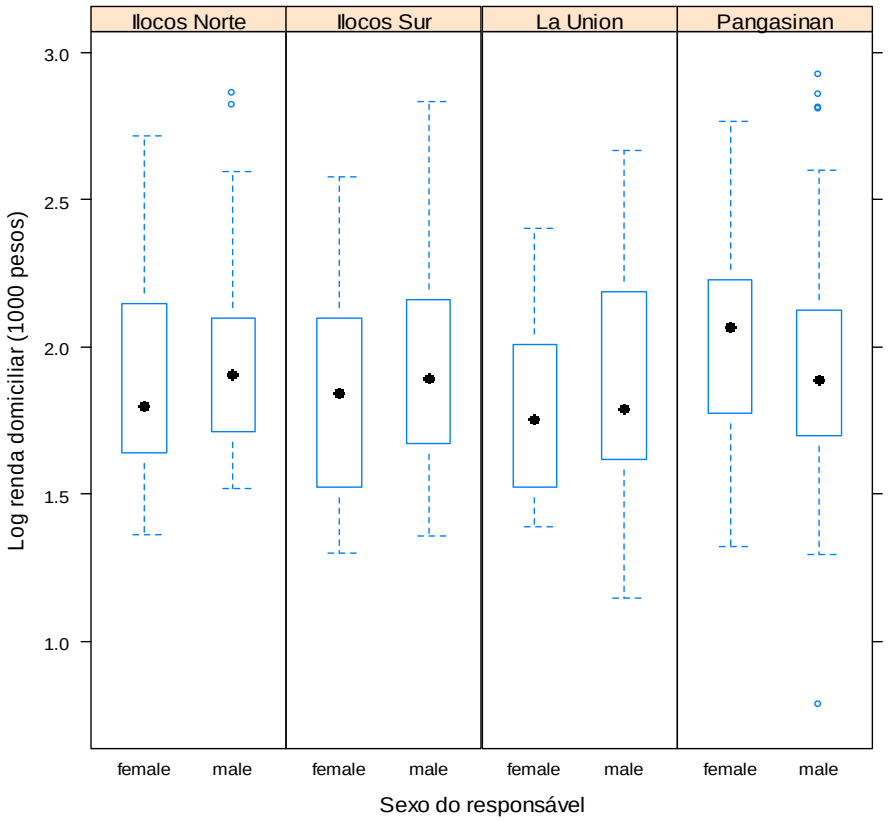



Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

# Variáveis quantitativas e qualitativas

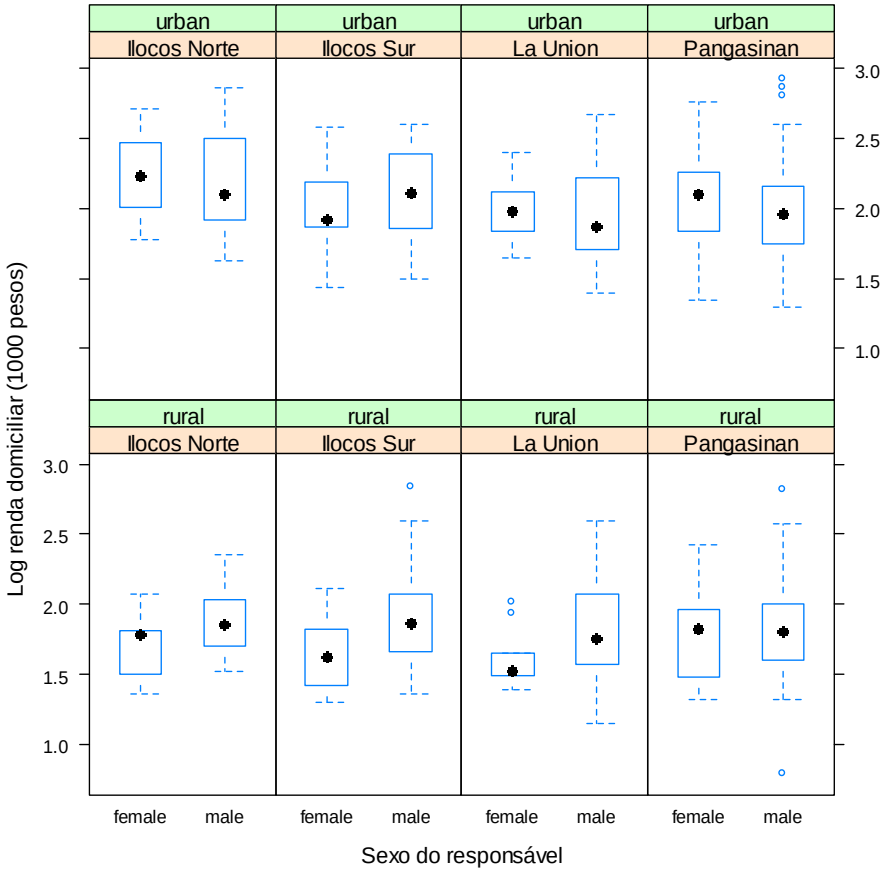
## Uma variável condicionante

```
> bwplot(log(income / 1000, 10)
~ sex | province, xlab = "Sexo
do responsável",   ylab = "Log
renda domiciliar (1000 pesos)",
layout = c(4, 1))
```




## Duas variáveis condicionantes

```
> bwplot(log(income / 1000, 10)
~ sex | province + urbanity,
xlab = "Sexo do responsável",
ylab = "Log renda domiciliar
(1000 pesos)")
```



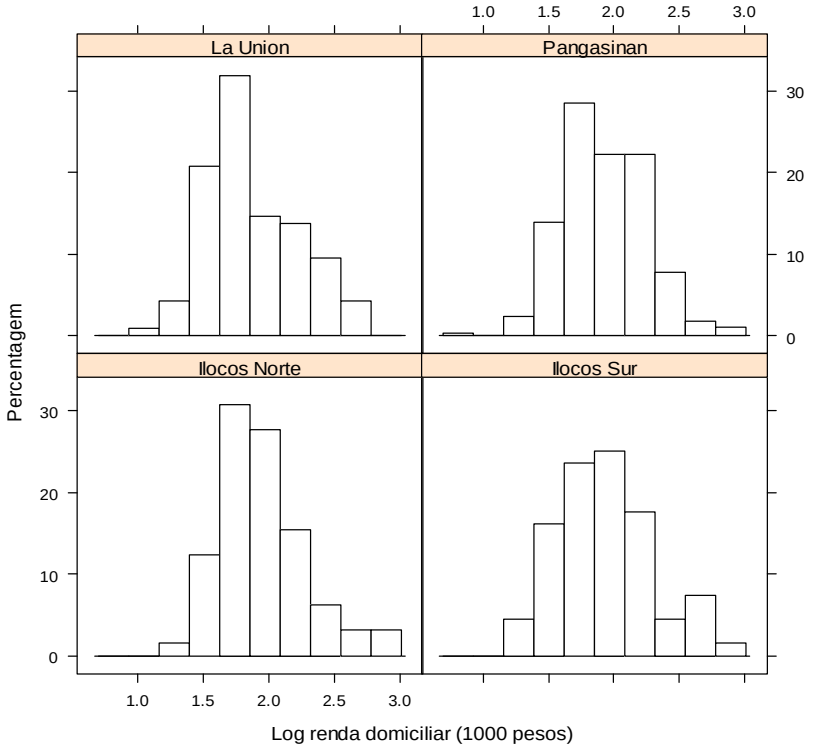

Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

# Variáveis quantitativas e qualitativas

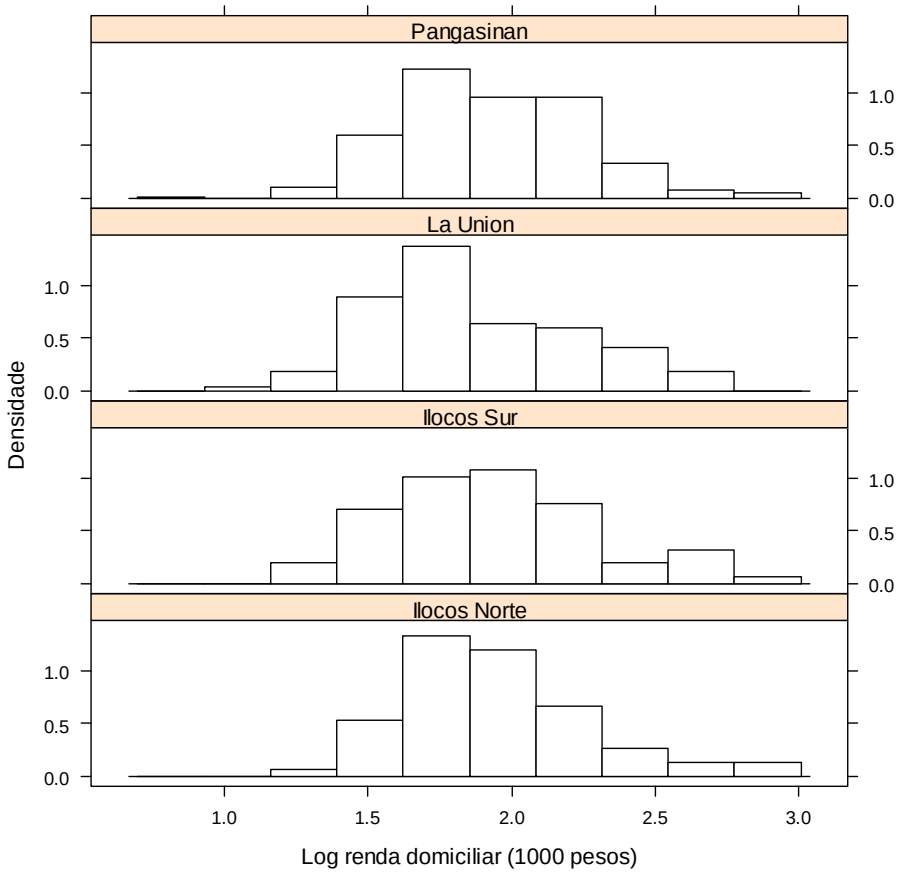
Histograma

Função `histogram` (lattice)

```
> histogram(~ log(income / 1000, 10) | province, type = "percent", ylab = "Percentagem", xlab = "Log renda domiciliar (1000 pesos)", col = "white")
```




```
> histogram(~ log(income / 1000, 10) | province, type = "density", layout = c(1, length(levels(province))), ylab = "Densidade", xlab = "Log renda domiciliar (1000 pesos)", col = "white")
```



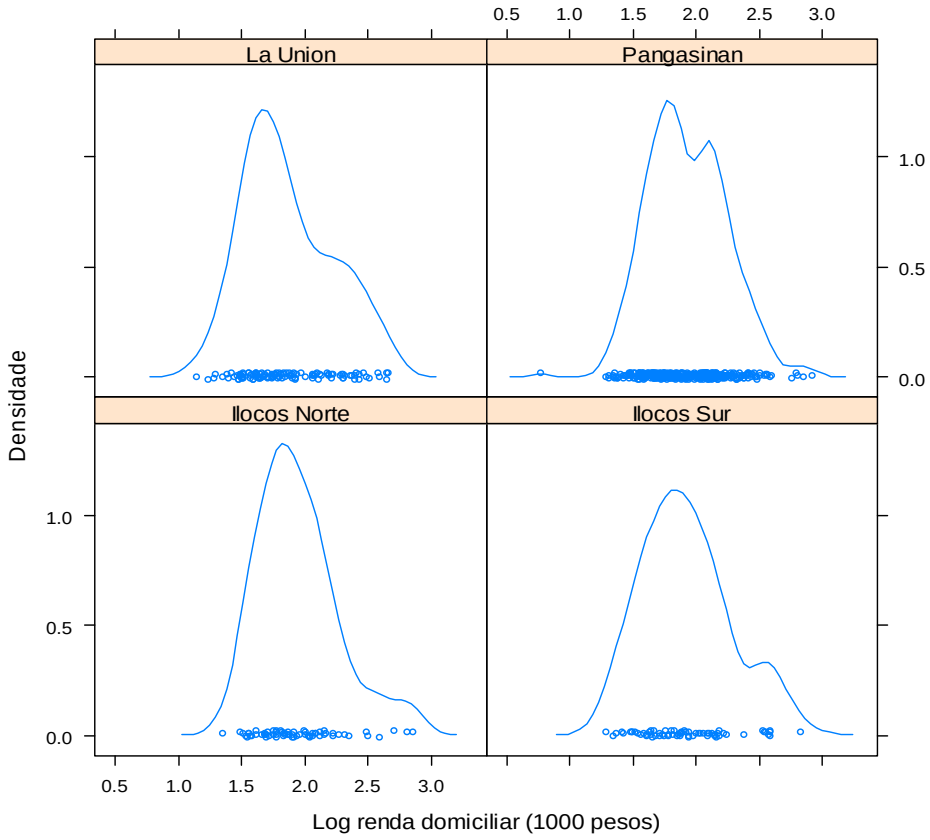

Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

# Variáveis quantitativas e qualitativas

## Gráfico de densidade

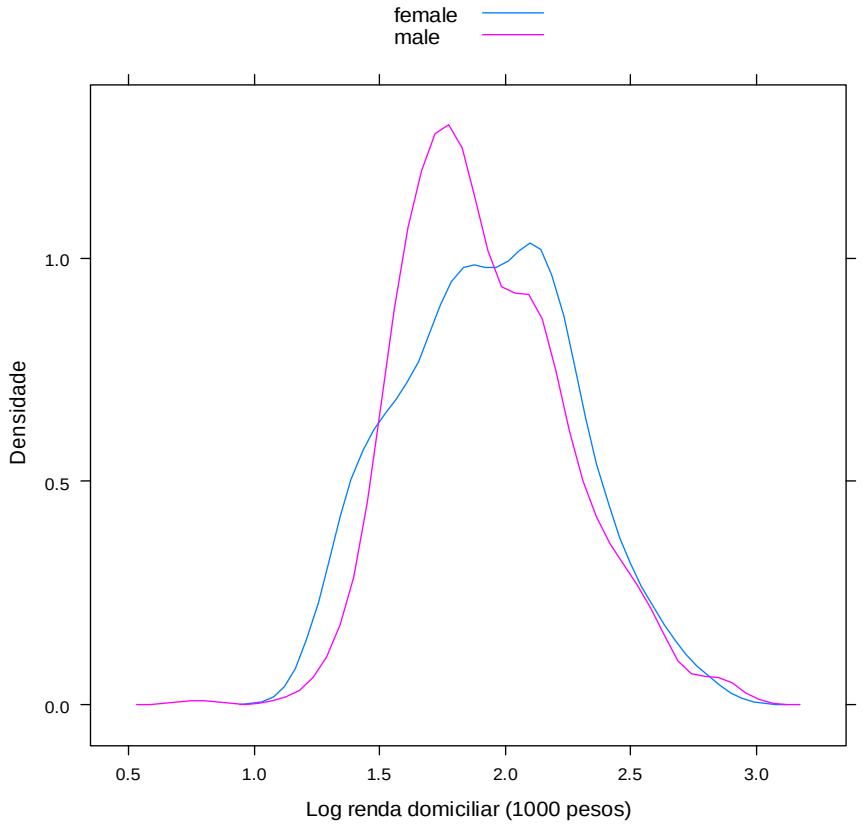
## Função `densityplot` (`lattice`)

```
> densityplot(~ log(income / 1000, 10) | province, ylab = "Densidade",  xlab = "Log renda domiciliar (1000 pesos)")
```




## Grupos em um só painel

```
> densityplot(~ log(income / 1000, 10), groups = sex, ylab = "Densidade",  xlab = "Log renda domiciliar (1000 pesos)", plot.points = FALSE, auto.key = TRUE)
```

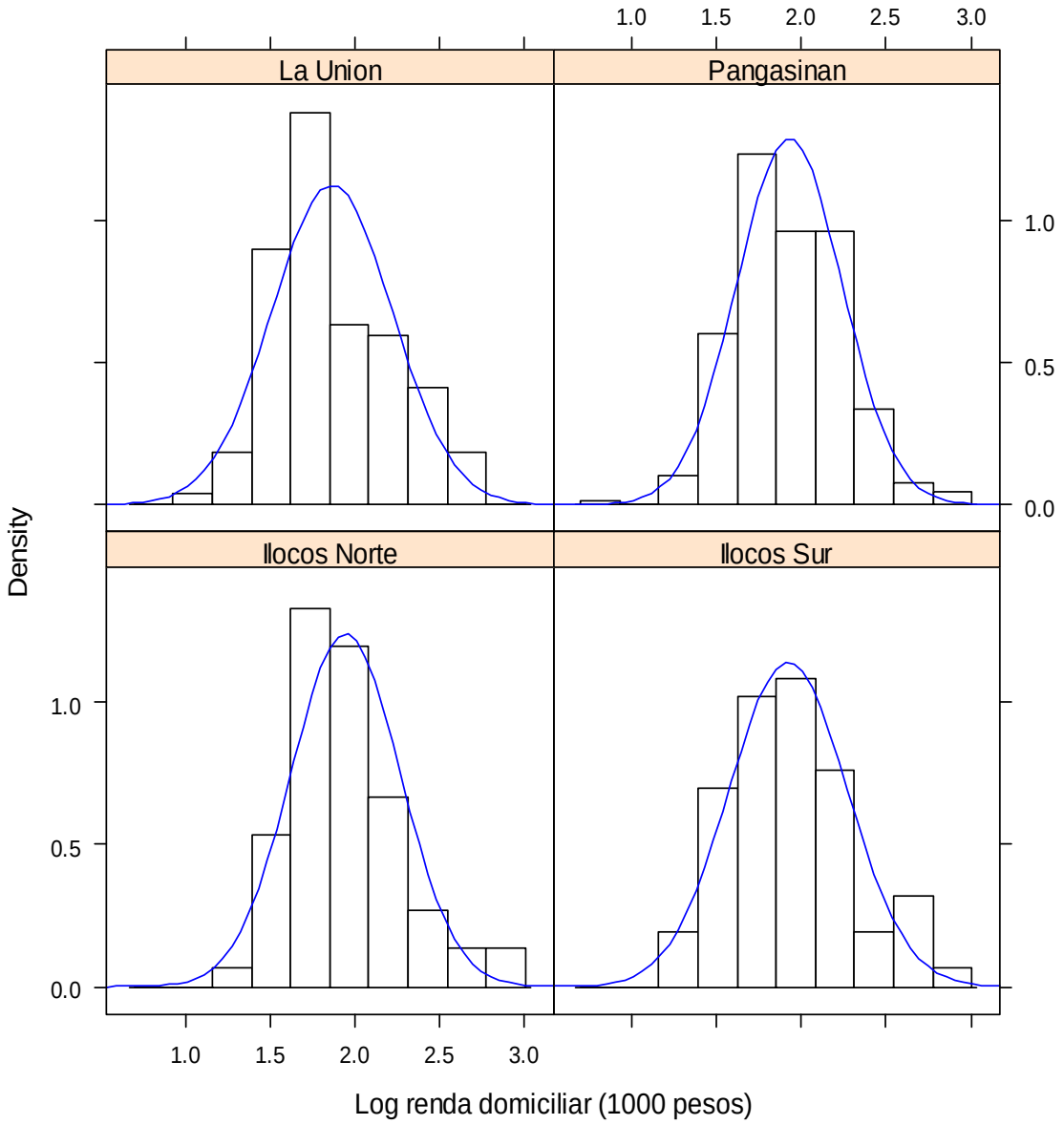



Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

# Variáveis quantitativas e qualitativas

Histograma e função densidade normal

```
> histogram(~ log(income / 1000, 10) | province, type = "density",
 ylab = "Densidade", xlab = "Log renda domiciliar (1000 pesos)", col = "white",
 panel = function(x, ...) {    panel.histogram(x, ...)
panel.mathdensity(dmath = dnorm, col = "blue", args = list(mean  = mean(x),sd = sd(x))) })
```




## Exercícios.

1. Substituir a função densidade normal pela densidade estimada.

2. Incluir os pontos no eixo horizontal.

Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

# Variáveis quantitativas e qualitativas

Gráfico de dispersão

Função `xyplot` (`lattice`)

```
> xyplot(log(income / 1000, 10) ~ family.size | province, xlab = "Tamanho da família", ylab = "Log renda domiciliar (1000 pesos)")
```

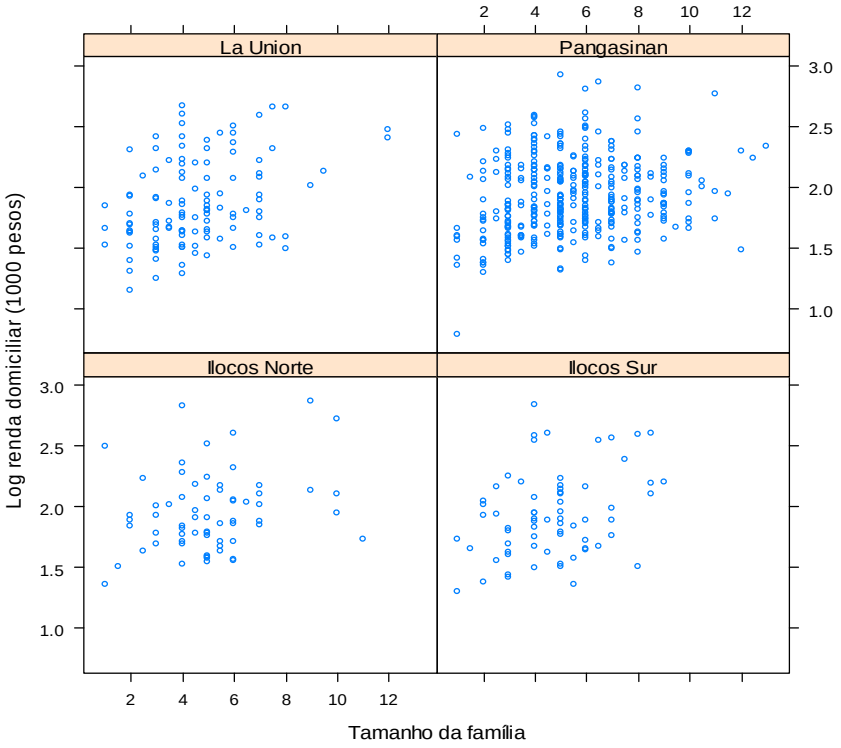
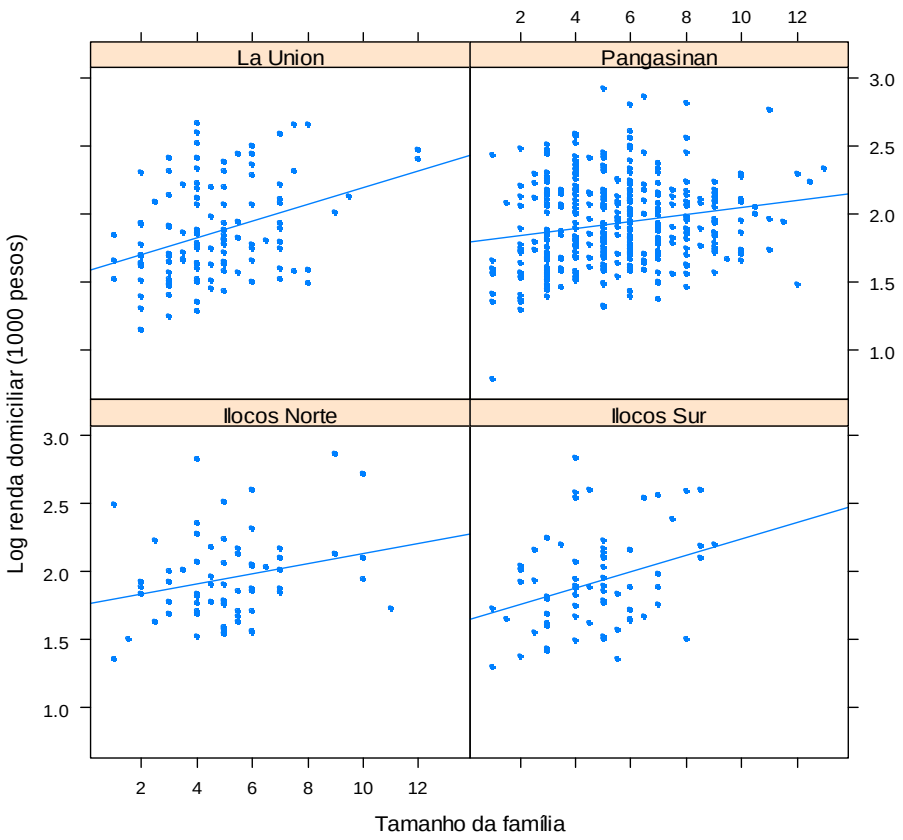



Gráfico com pontos (p) e reta ajustada (r)

```
> xyplot(log(income / 1000, 10) ~ family.size | province, xlab = "Tamanho da família", ylab = "Log renda domiciliar (1000 pesos)", pch = 20, type = c("p", "r"))
```



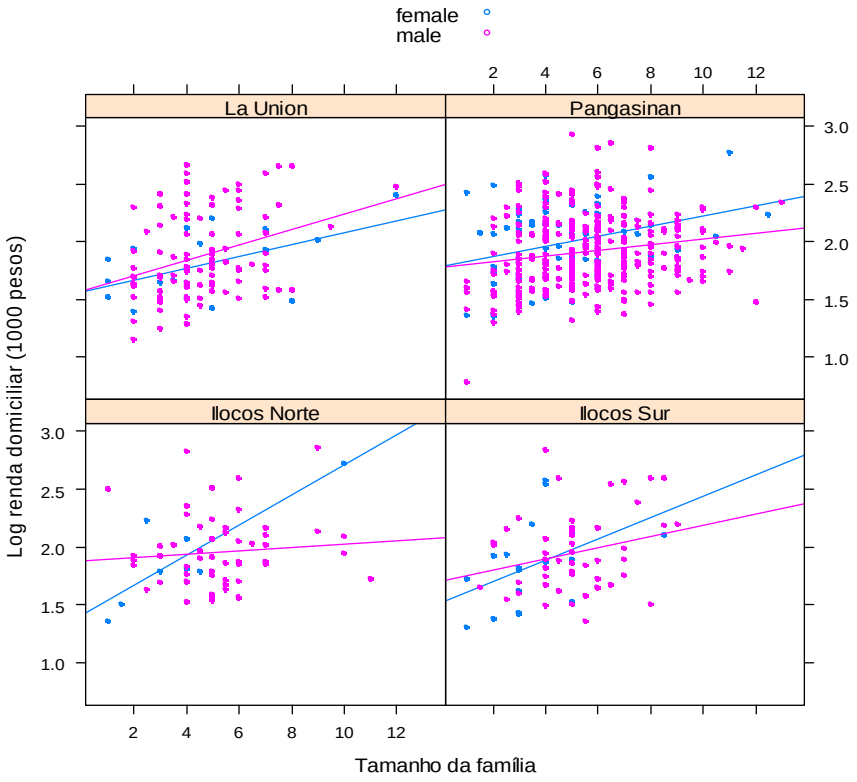

Exercício. Substituir as retas ajustadas por linhas de tendência.

Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

# Variáveis quantitativas e qualitativas

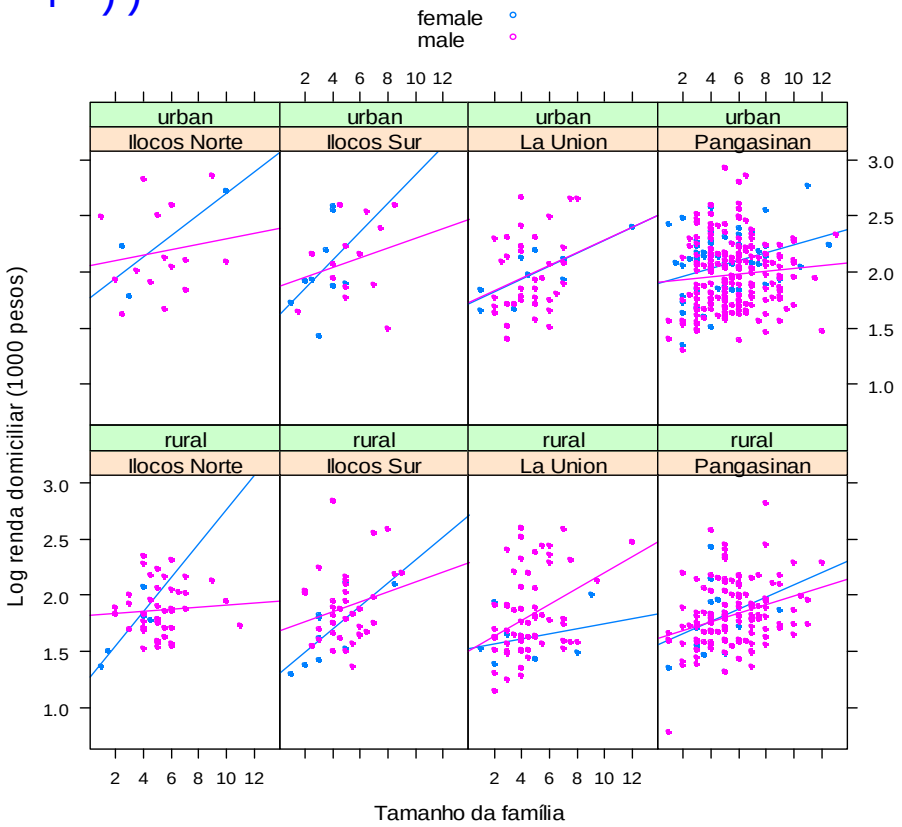
## Grupos de acordo com a variável sex

```
> xyplot(log(income / 1000, 10) ~ family.size | province, group = sex,  auto.key = TRUE, xlab = "Tamanho da família",  ylab = "Log renda domiciliar (1000 pesos)", pch = 20, type = c("p", "r"))
```




## Duas variáveis condicionantes

```
> xyplot(log(income / 1000, 10) ~ family.size | province + urbanity, group = sex,  auto.key = TRUE, xlab = "Tamanho da família",  ylab = "Log renda domiciliar (1000 pesos)", pch = 20, type = c("p", "r"))
```



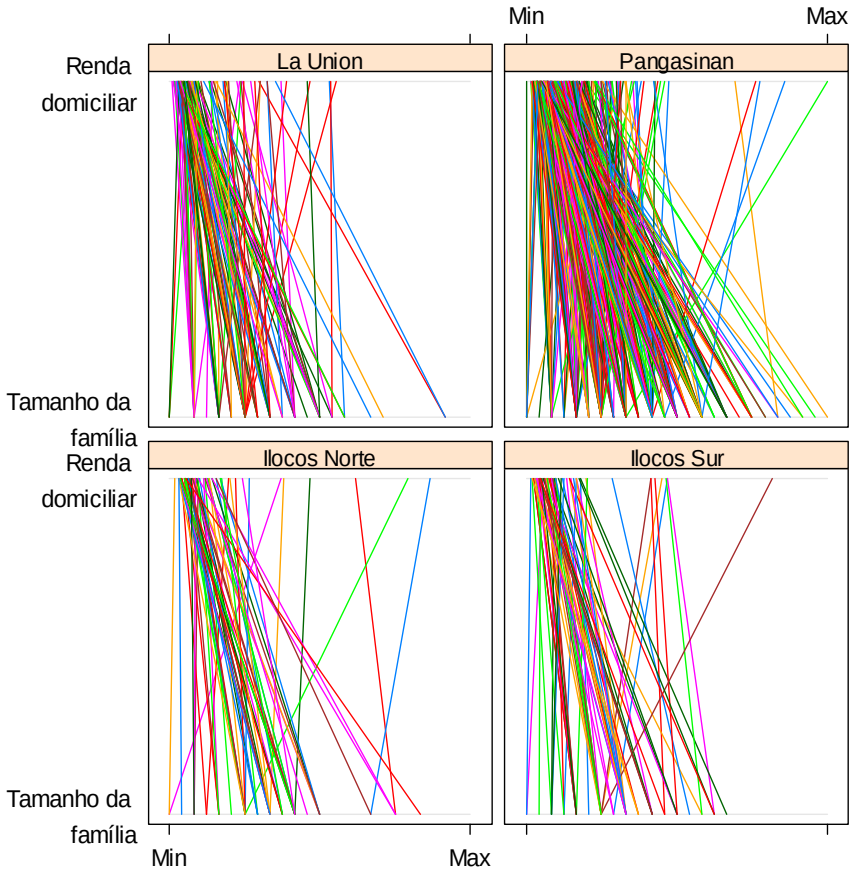

Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.

# Variáveis quantitativas e qualitativas

## Gráfico de coordenadas paralelas

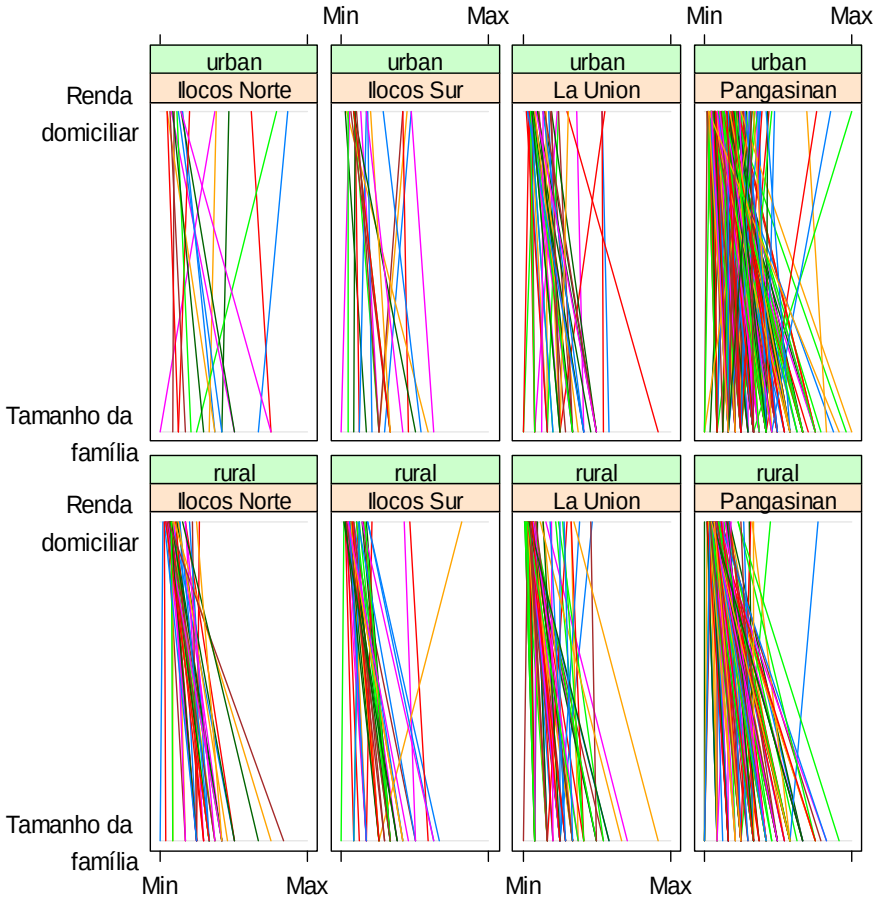
## Função `parallel` (`lattice`)

```
> parallel(~ cbind( family.size, income) | province,  varnames = c("Tamanho da \nfamília", "Renda\n domiciliar"))
```




## Duas variáveis condicionantes

```
> parallel(~ cbind(family.size, income) | province + urbanity, varnames =  c("Tamanho da \nfamília", "Renda domiciliar"))
```




Fonte: Mário de Castro, Notas de aula de Análise Exploratória de Dados. ICMC-USP, 2010.