SME 0803 - Análise Exploratória de Dados

Katiane S. Conceição

com modificações por Juliana Cobre
e colaboração do material cedido por Mário de Castro

SME/ICMC/USP

Parte VII - 2019

Medidas de associação

Medidas de associação

- É quando se analisa conjuntamente duas ou mais variáveis ao mesmo problema.
- Em muitos problemas práticos, para compreender melhor os dados as variáveis devem ser analisadas conjuntamente. A análise isolada de cada variável pode não ser suficiente.
- A análise bivariada é uma extensão da análise de uma única variável e é também um caso particular da análise multivariada.

Medidas de associação: Análise bivariada

Objetivo: Analisar como é a relação entre as observações de x e suas y que são observadas em pares $(x_i, y_i), i = 1, \ldots, n$.

Variáveis: Podem ser quantitativas ou qualitativas (ambas diferentes ou ambas com a mesma classificação).

Relação de sentido: negativa ou positiva.

Relação de força: fraca ou forte.

Ferramentas: Gráficos e medidas estatísticas.

Nota: Podemos ter mais de duas variáveis $(x_1, x_2, \ldots, x_k, y_1, y_2, \ldots, y_g) \Rightarrow$ Análise multivariada.

Medidas de associação: Análise bivariada

Tipos de relação

- **Causal unilateral:** y depende de x ou x depende de y.

$$\boxed{x} \rightarrow \boxed{y}$$

Exemplo: O preço de venda de um produto y depende da distância entre o local de armazenamento/produção e o local de entrega x.

- **Causal bilateral ou interdependência:** y depende de x e x depende de y.

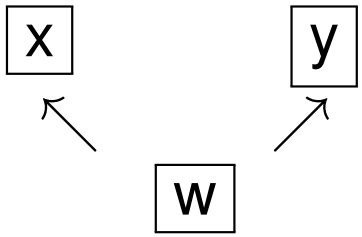
$$\boxed{x} \rightleftarrows \boxed{y}$$

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

Medidas de associação: Análise bivariada

Tipos de relação

- **Dependência indireta (ou associação espúria):** *x* e *y* são influenciadas por outras variáveis.



Exemplo:. Relação entre número de casos de queimadura de sol *y* número de mosquitos da dengue *x*. (Por Doris Fontes).

Medidas de associação: Análise bivariada

- Tabela de variáveis qualitativas $\times$ qualitativas.
- Tabela de variáveis quantitativas $\times$ qualitativas.
- Tabela de variáveis quantitativas $\times$ quantitativas.
- Covariância e coeficiente de correlação.

Qualitativas × Qualitativas

Exemplo: Tipo de aplicativo (Entreternimento, Informação, Cosumo) e o *gênero* do usuário.

Residência	Feminino	Masculino	Total
Entreternimento	165	264	429
Informação	149	251	400
Consumo	399	772	1171
Total	713	1287	2000

Nota: Esta tabela recebe o nome de **Tabela de contingência**.

Distribuições conjunta e marginais:

- A tabela de contingência apresenta a distribuição conjunta das variáveis.
  - Qual o número de clientes que preferem aplicativos de informações e são homens?
- A Tabela permite encontrar as distribuições da cada variável que são chamadas **distribuições marginais**.
  - Quantos clientes preferem aplicativos de compras?
  - Quantos clientes são mulheres?

Pergunta: Existe associação entre o tipo de aplicativo e o gênero do usuário? Como quantificar?

Independência: x e y são independentes, se somente se,

$$f_{ij} = \frac{f_{i.} f_{.j}}{n}, i = 1, \ldots, k \text{ e } j = 1, \ldots, m,$$

ou

$$\frac{f_{ij}}{n} = \frac{f_{i.}}{n} \frac{f_{.j}}{n}, i = 1, \ldots, k \text{ e } j = 1, \ldots, m,$$

Justificativa: Adaptação do conceito de independência entre as v.a. discretas X e Y.

Medidas de associação — Qualitativas $\times$ Qualitativas

Coeficientes de associação: ideia é comparar as diferenças entre as frequências absolutas observadas e as frequências calculadas supondo independência entre x e y.

Sob suposição de independência $f_{ij}^{\text{ind}} = \dfrac{f_{i\cdot} f_{\cdot j}}{n}$.

Coeficiente qui-quadrado de Pearson

$$Q^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(f_{ij} - f_{ij}^{\text{ind}})^2}{f_{ij}^{\text{ind}}}$$

Interpretação

- Q^2 maior do que o quantil da $\chi^2_{(k-1)(m-1)} \Rightarrow$ existe de associação.
- Q^2 menor do que o quantil da $\chi^2_{(k-1)(m-1)} \Rightarrow$ não existe de associação.

Nota: Quanto maior o valor do coeficiente qui-quadro de Pearson, mais forte é a associação.

Coeficiente de contingência:

$$C = \sqrt{\frac{Q^2}{Q^2 + n}}.$$

Nota: O valor máximo de C depende de k e m.

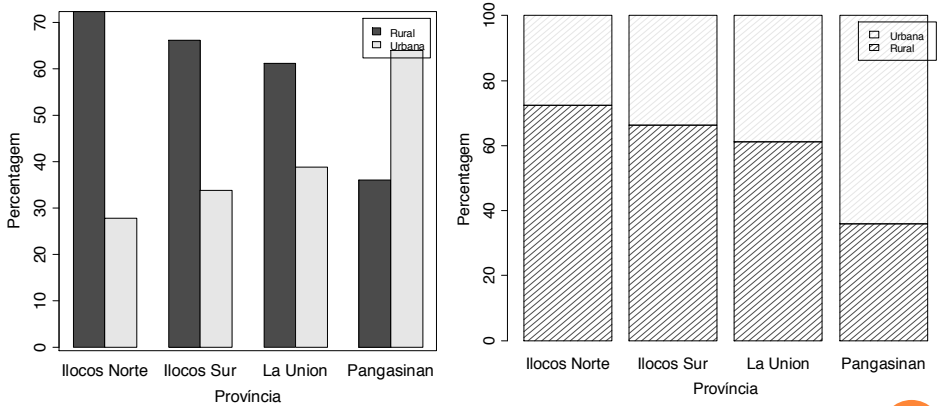
Interpretação: não existe uma única interpretação (ou valores).

- $C \approx 0 \Rightarrow$ associação fraca.
- $C \approx 0,5 \Rightarrow$ associação moderada.
- $C \approx 1 \Rightarrow$ associação forte.

Medidas de associação

Qualitativas $\times$ Qualitativas

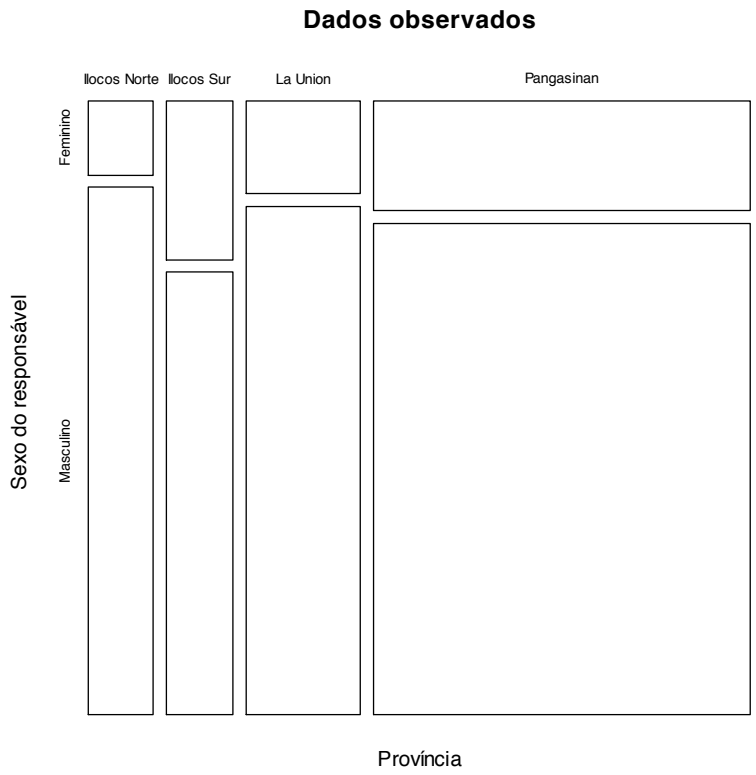
Graficamente

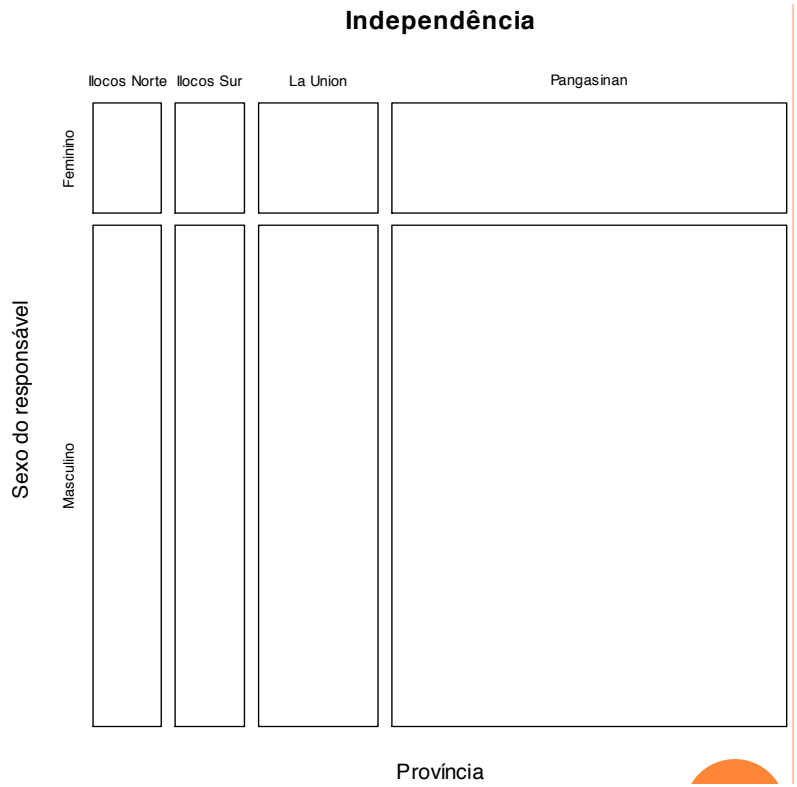


Medidas de associação

Qualitativas $\times$ Qualitativas

Graficamente





Qualitativas $\times$ Quantitativas

Variáveis: (x_i, y_i) sendo x qualitativa e y quantitativa.

Interesse: Relação causal unilateral $x \rightarrow y$.

Procedimento: medidas resumo e gráfico de y para cada classe de x.

Coeficiente R^2 : Sejam f_i as frequências observadas de cada classe x_i e s_i^2 a variância dos valores y_j para cada valor x_i, ou seja, para cada classe. E

$$\overline{s^2} = \frac{\sum_{i=1}^{k} f_i s_i^2}{\sum_{i=1}^{k} f_i} \text{ e } s^2 = \frac{1}{n-1} \sum_{j=1}^{n} (y_j - \overline{y})^2.$$

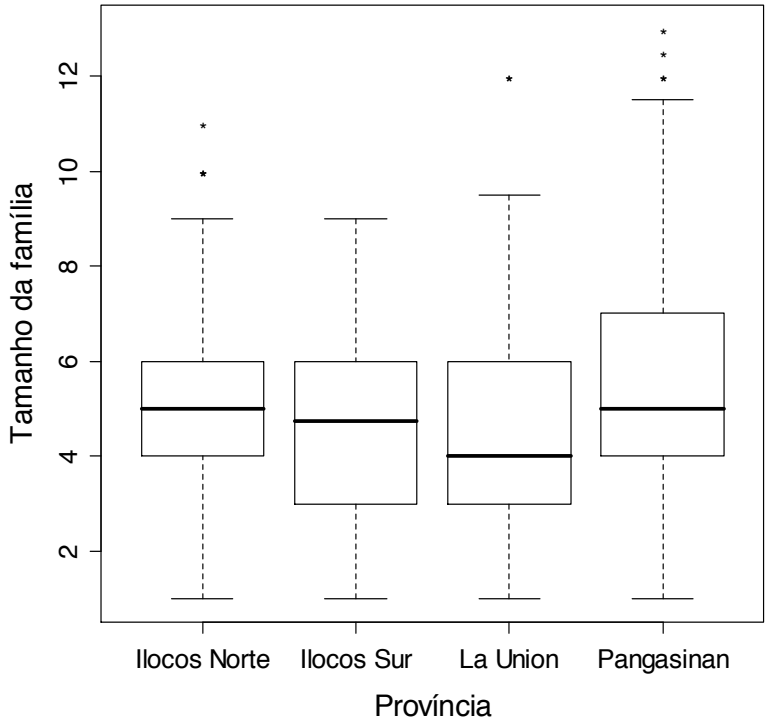
$$R^2 = \frac{s^2 - \overline{s^2}}{s^2}, 0 \leq R^2 \leq 1,$$

em que

Interpretação: Quanto maior R^2, mais forte a associação entre x e y, ou seja, maior o poder de explicação de x para y (em termos de variabilidade).

Medidas de associação — Quantitativas $\times$ Qualitativas

Exemplo



$R^2 = 0,02429767 \approx 2,4\%$: A variável *province* explica cerca de 2,4% da variabilidade do tamanho da família.

Quantitativas × Quantitativas

- Quando as duas variáveis a serem analisadas são de mesma natureza sendo ambas quantitativas, mas são agrupadas em classes.
- **Exemplo:** Anos de experiência e salário (R$) de funcionários de um banco.

Tempo (anos)	800-1000	1000-2000	2000 - 40000	> 4000	Total
1 - 5	10	8	2	0	20
6 - 8	0	6	4	1	11
9 -12	0	3	5	8	16
> 12	0	3	6	8	17
Total	10	20	17	17	64

- **Questão:** Dado que um funcionário tem entre 6 e 8 anos de experiência qual a risco de ele ter um salário superior a R$ 4000,00?

Medidas de associação — Quantitativas $\times$ Quantitativas

- Quando as duas variáveis a serem analisadas são de mesma natureza sendo ambas quantitativas, mas são apresentadas em valores brutos.
- **Exemplo:** Tempo em que o cliente está empregado e seu salário.

Tempo (anos)	Salário
1	1200
2	2000
3	2800
4	3200
5	4000

- Neste caso a tabela apresenta a relação entre as variáveis.

Coeficiente de correlação linear

- Quando as duas variáveis a serem analisadas são de mesma natureza sendo **ambas quantitativas**, a representação gráfica entre as variáveis deve ser feita com gráfico de dispersão.
- Uma medida de **relação linear entre as variáveis quantitativas** pode ser calculada com o **coeficiente de correlação linear**.

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)\,Var(Y)}},$$

em que $Cov(X, Y)$ representa a covariância entre as variáveis quantitativas X e Y; e os termos $Var(X)$ e $Var(Y)$ são as respectivas variâncias.

- A covariância $Cov(X, Y)$ pode ser estimada a partir dos dados $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ por:

$$Cov_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}.$$

em que $\overline{x}$ e $\overline{y}$ são as respectivas médias dos dados.

- Estimando-se as variâncias respectivamente por:

$$s_X^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \quad \text{e} \quad s_Y^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}.$$

- O coeficiente de correlação linear pode ser estimado por:

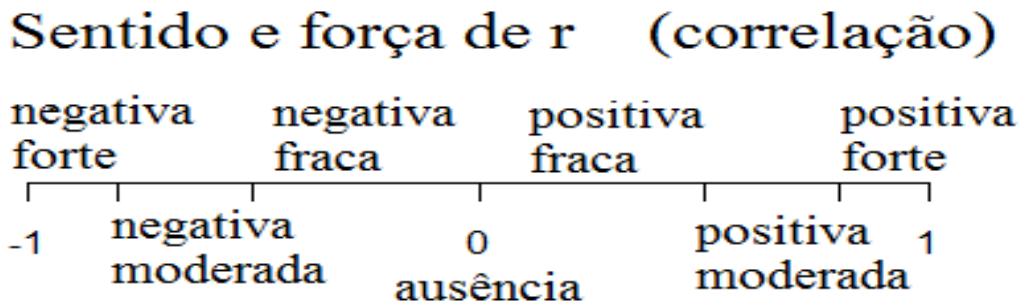
$$\widehat{\rho} = \frac{Cov_{X,Y}}{s_X s_Y},$$

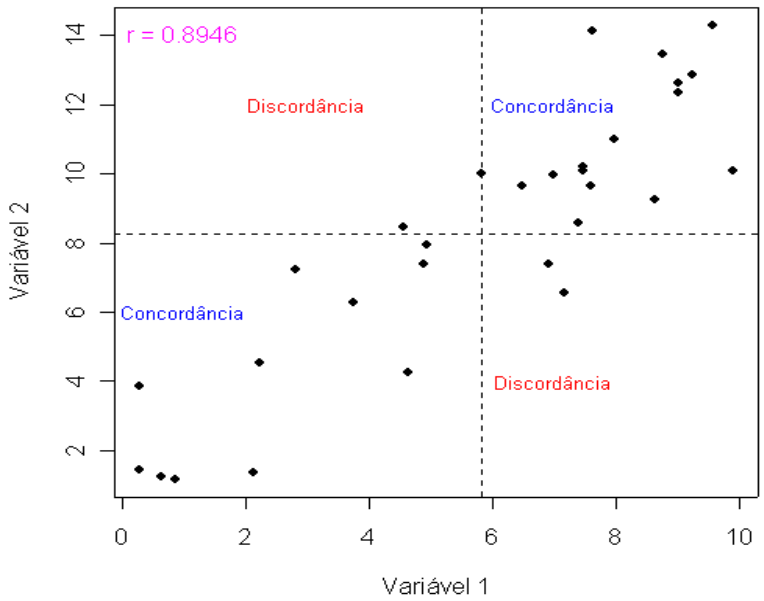
que é um número entre -1 e 1, ou seja $-1 \leq \rho \leq 1$.

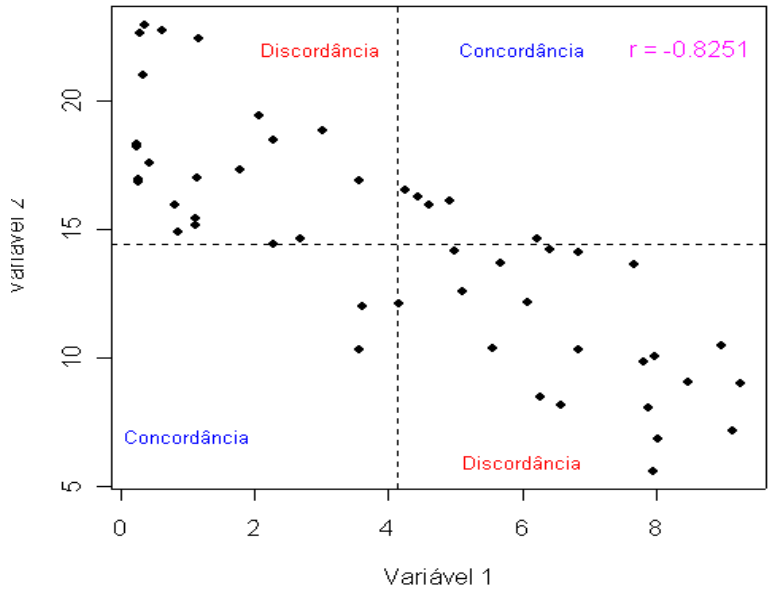
  - Se $\rho < 0$ indica uma relação linear decrescente, $X \uparrow \Leftrightarrow Y \downarrow$.
  - Se $\rho > 0$ indica uma relação linear crescente, $X \uparrow \Leftrightarrow Y \uparrow$.
  - Se $\rho \approx 0$ indica falta de relação linear, ou as variáveis são não correlacionadas.

Medidas de associação

Coeficiente de correlação linear







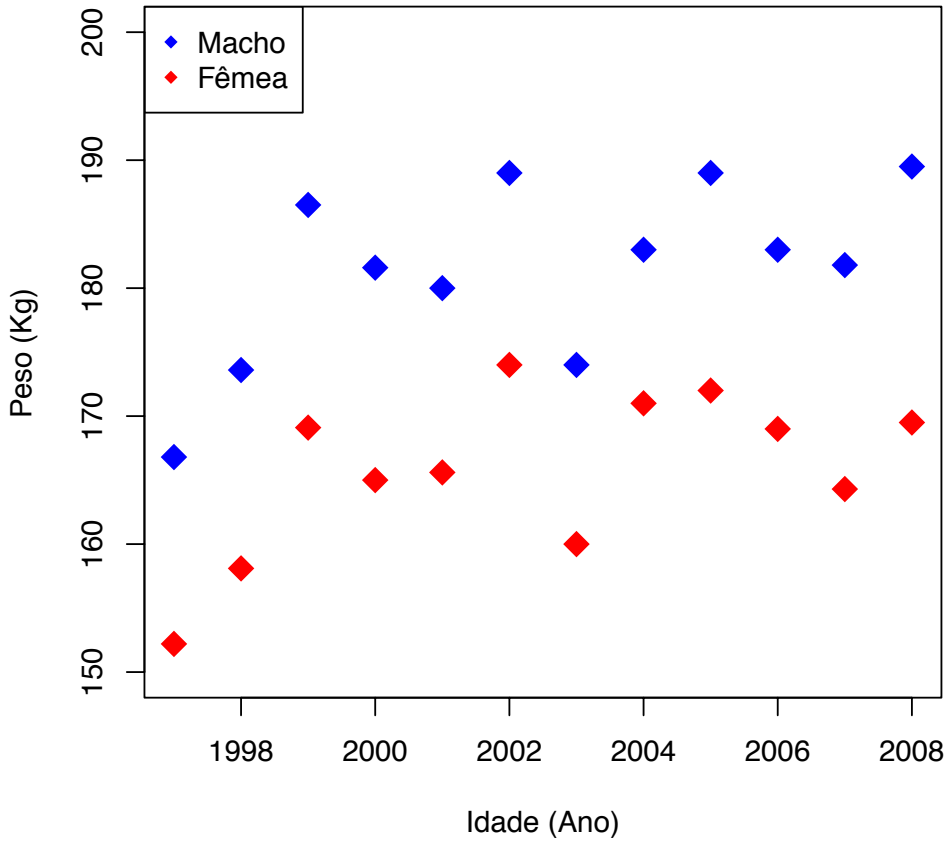
Exemplo: Peso de um boi (kg) e idade (ano), no eixo *x* temos idade e no eixo *y* pesos.

Table: Pesos de bovinos (kg) e idade (ano).

Idade	Macho	Fêmea
12	189,5	169,5
11	181,8	164,3
10	183,0	169,0
9	189,0	172,0
8	183,0	171,0
7	174,0	160,0
6	189,0	174,0
5	180,0	165,6
4	181,6	165,0
3	186,5	169,1
2	173,6	158,1
1	166,8	152,2

Medidas de associação

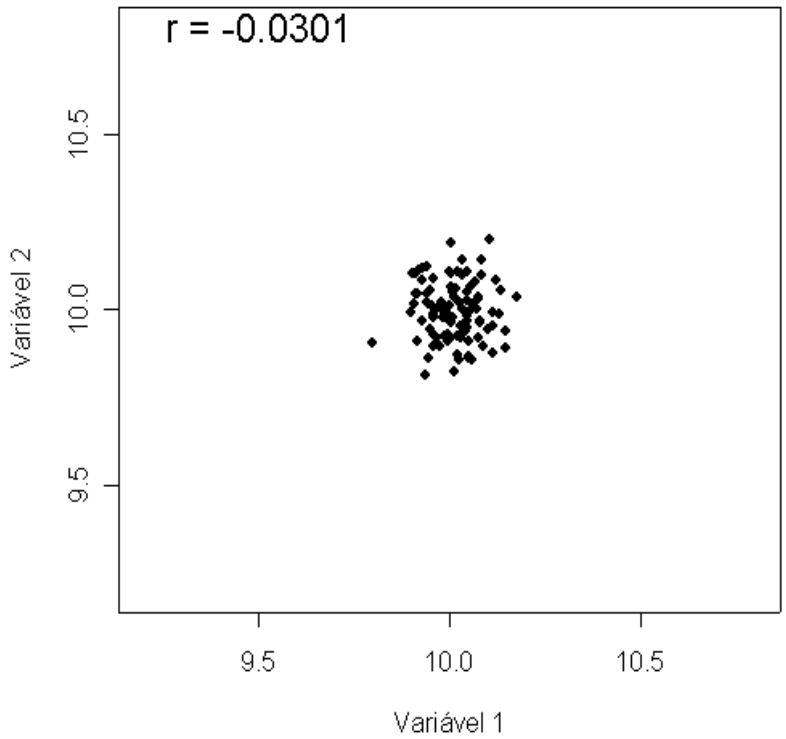
Coeficiente de correlação linear

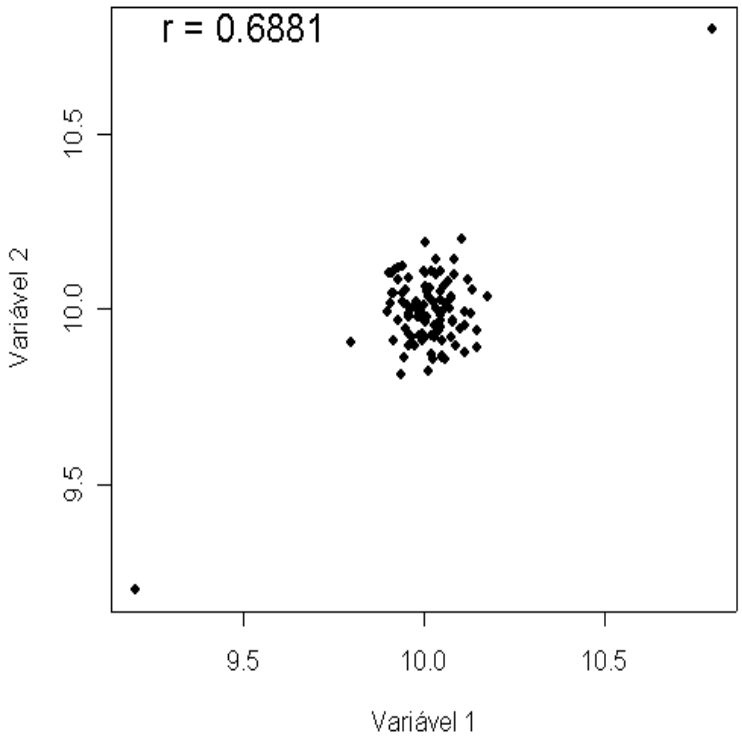


- Coeficiente de correlação entre idade e peso dos machos $\widehat{\rho} = 0,5921$.
- Coeficiente de correlação entre idade e peso das fêmeas $\widehat{\rho} = 0,5580$.
- Coeficiente de correlação entre os peso dos bois machos e das fêmeas $\widehat{\rho} = 0,9539$.

Medidas de associação

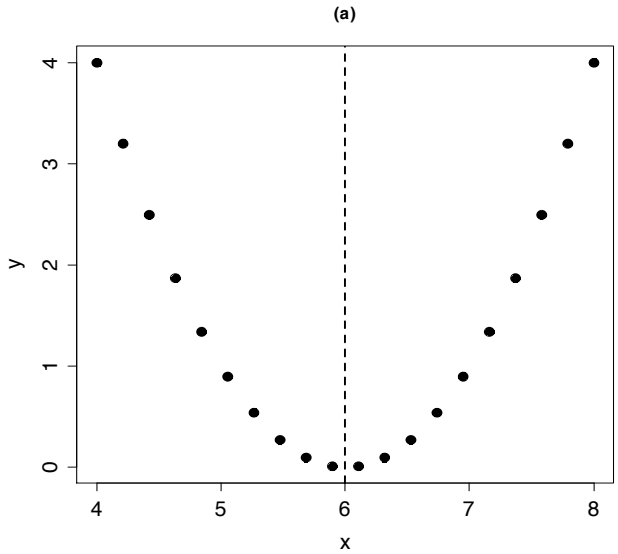
Coeficiente de correlação linear

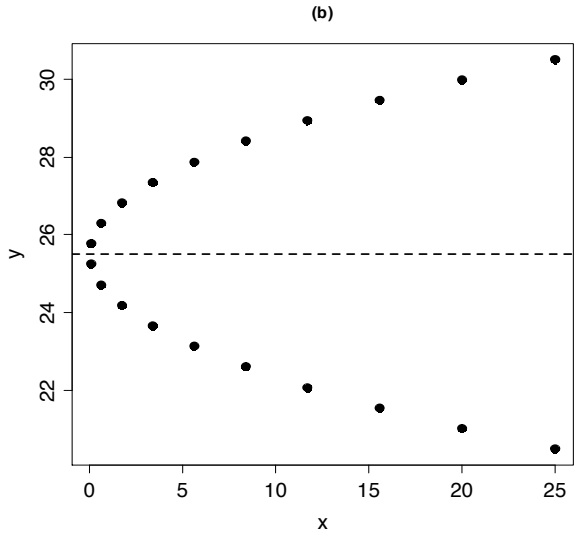


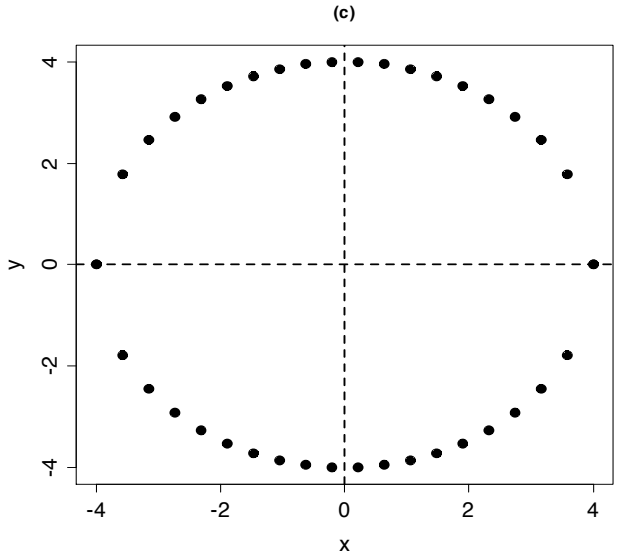


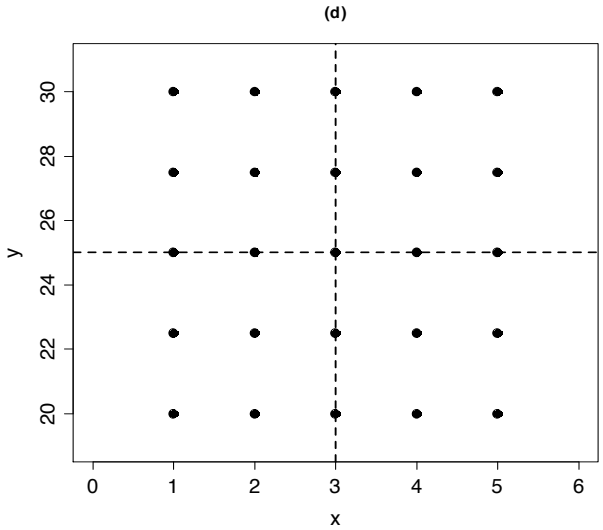
Medidas de associação

Coeficiente de correlação linear







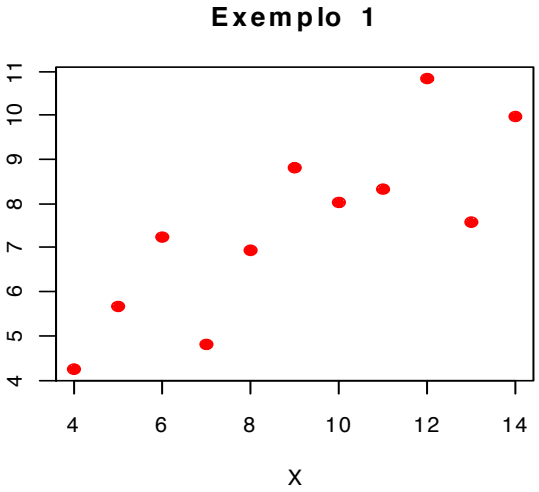


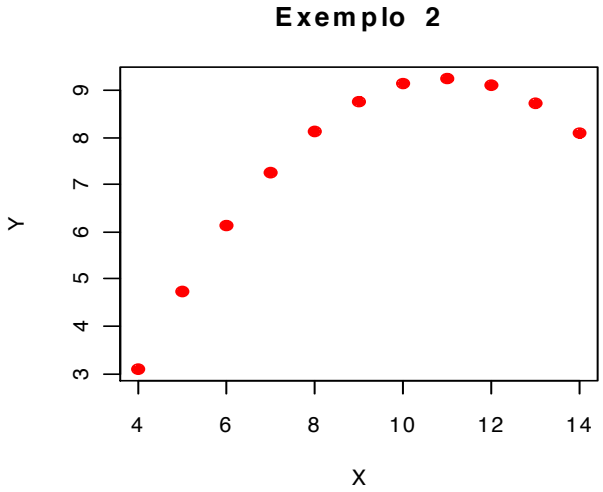
Exercício. Prove que se houver simetria em x e/ou y, então r = 0.

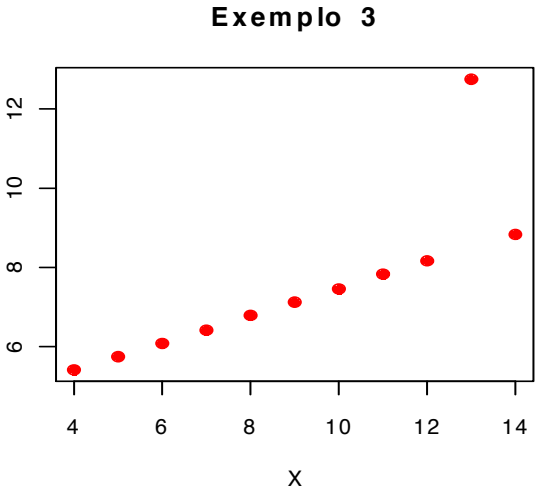
Obs. r = 0 não significa ausência de associação.

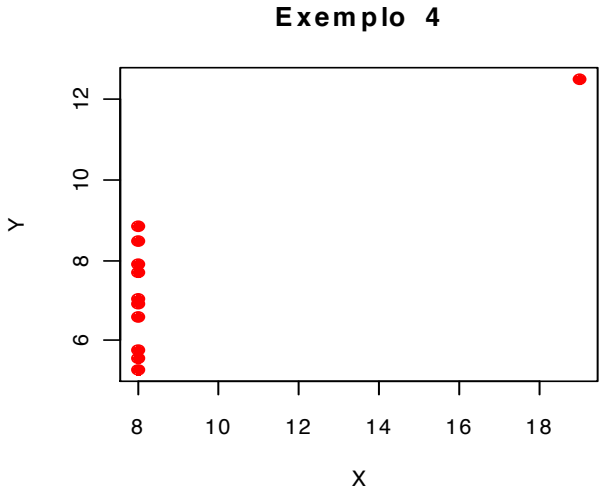
Medidas de associação

Coeficiente de correlação linear









Dados anscombe em R

```
> ?anscombe
```

Valores de r:

Exemplo 1: 0,8164

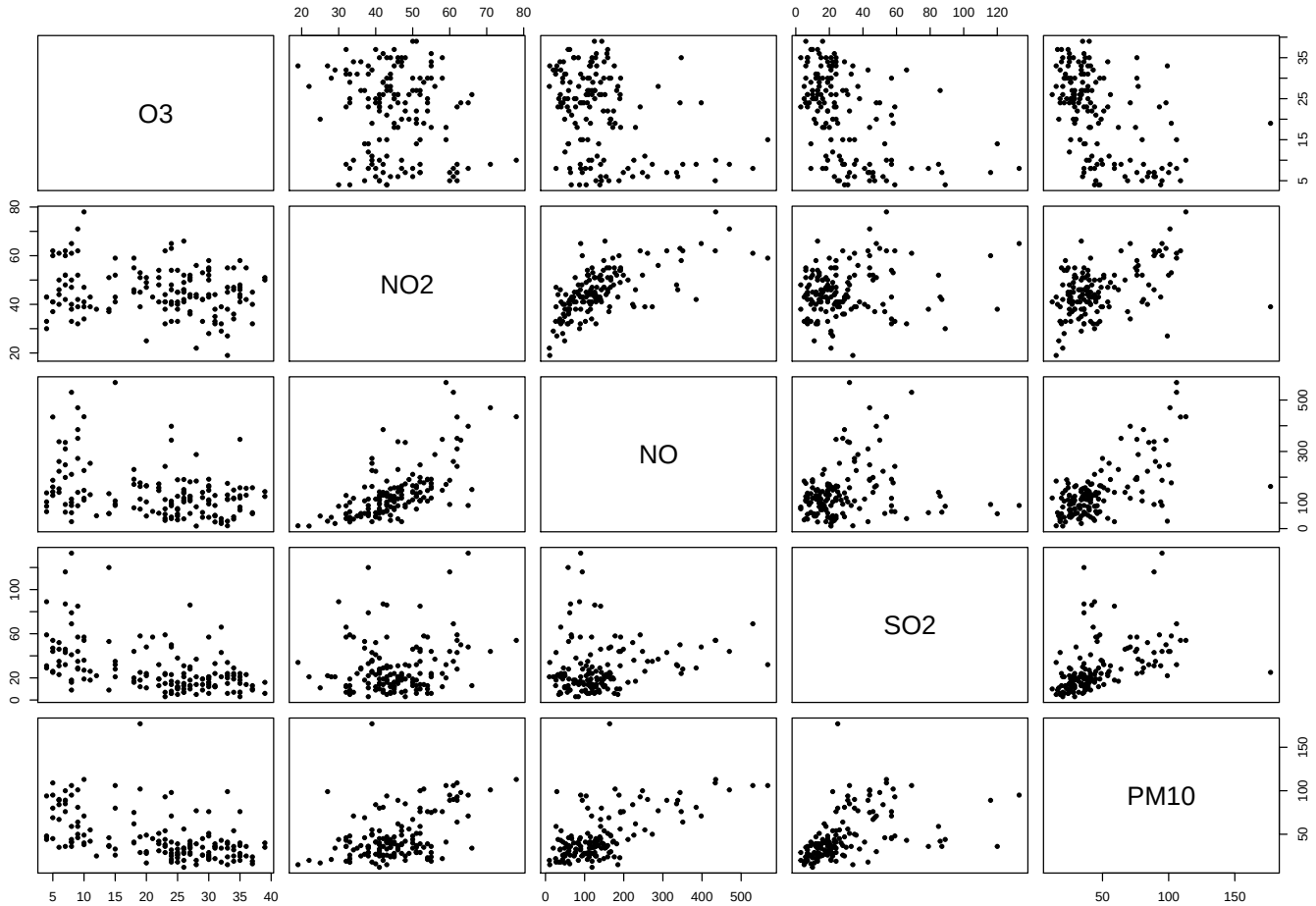
Exemplo 2: 0,8162

Exemplo 3: 0,8163

Exemplo 4: 0,8165

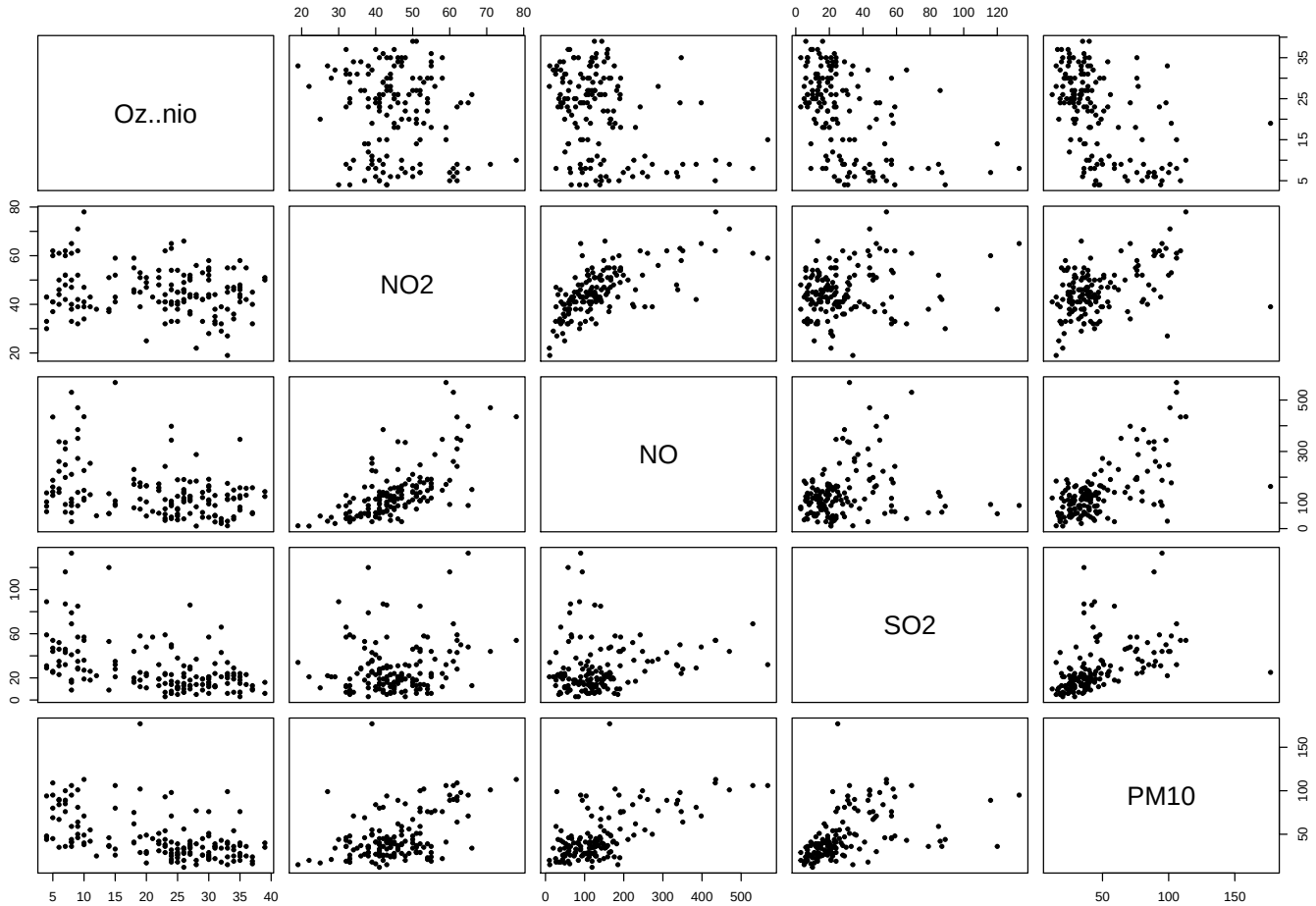
Medidas de associação Coeficiente de correlação linear

Exemplo: Dados `winter` do pacote `texmex` do R.



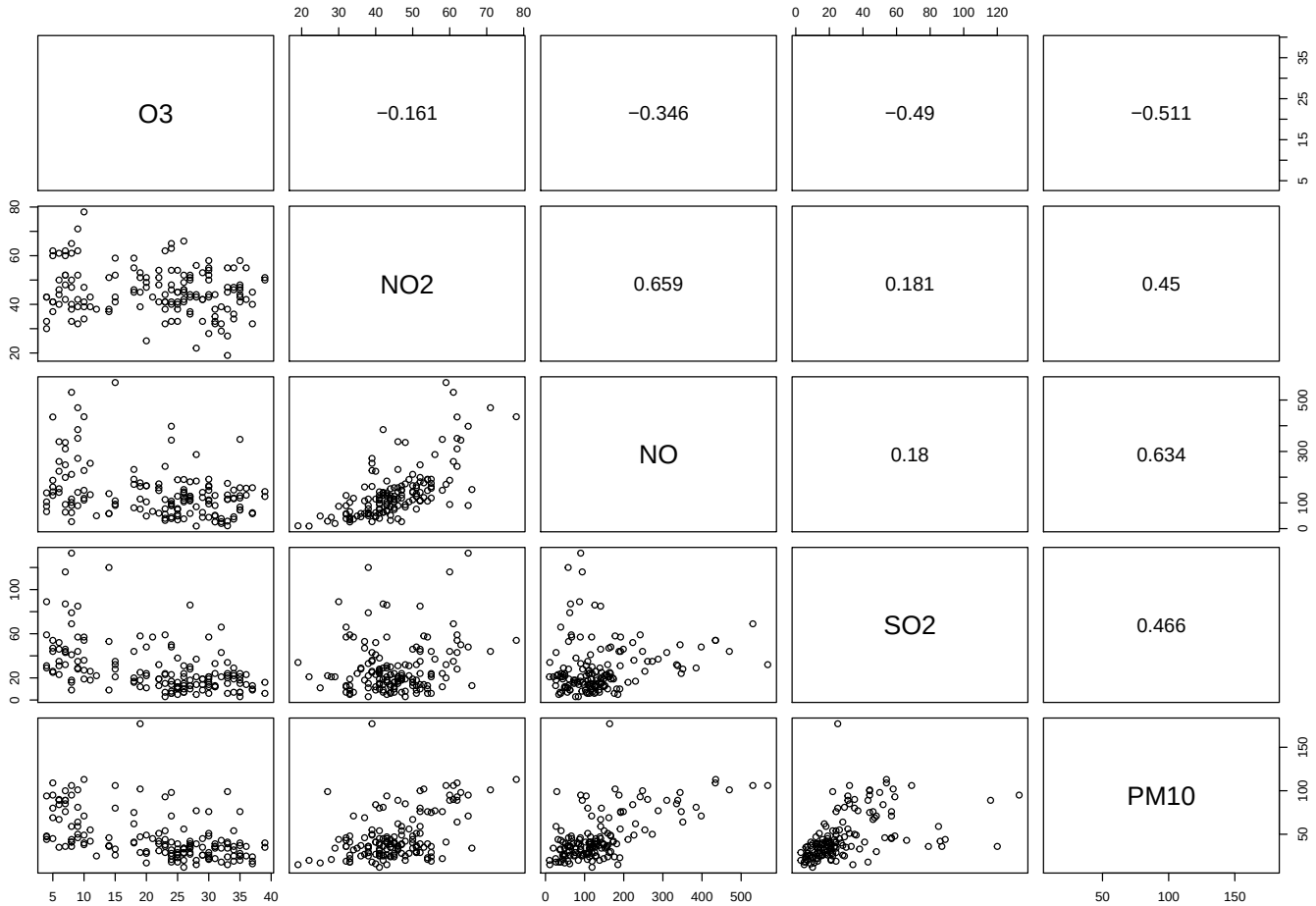
Medidas de associação Coeficiente de correlação linear

Exemplo: Dados `winter` do pacote `texmex` do R.



Medidas de associação Coeficiente de correlação linear

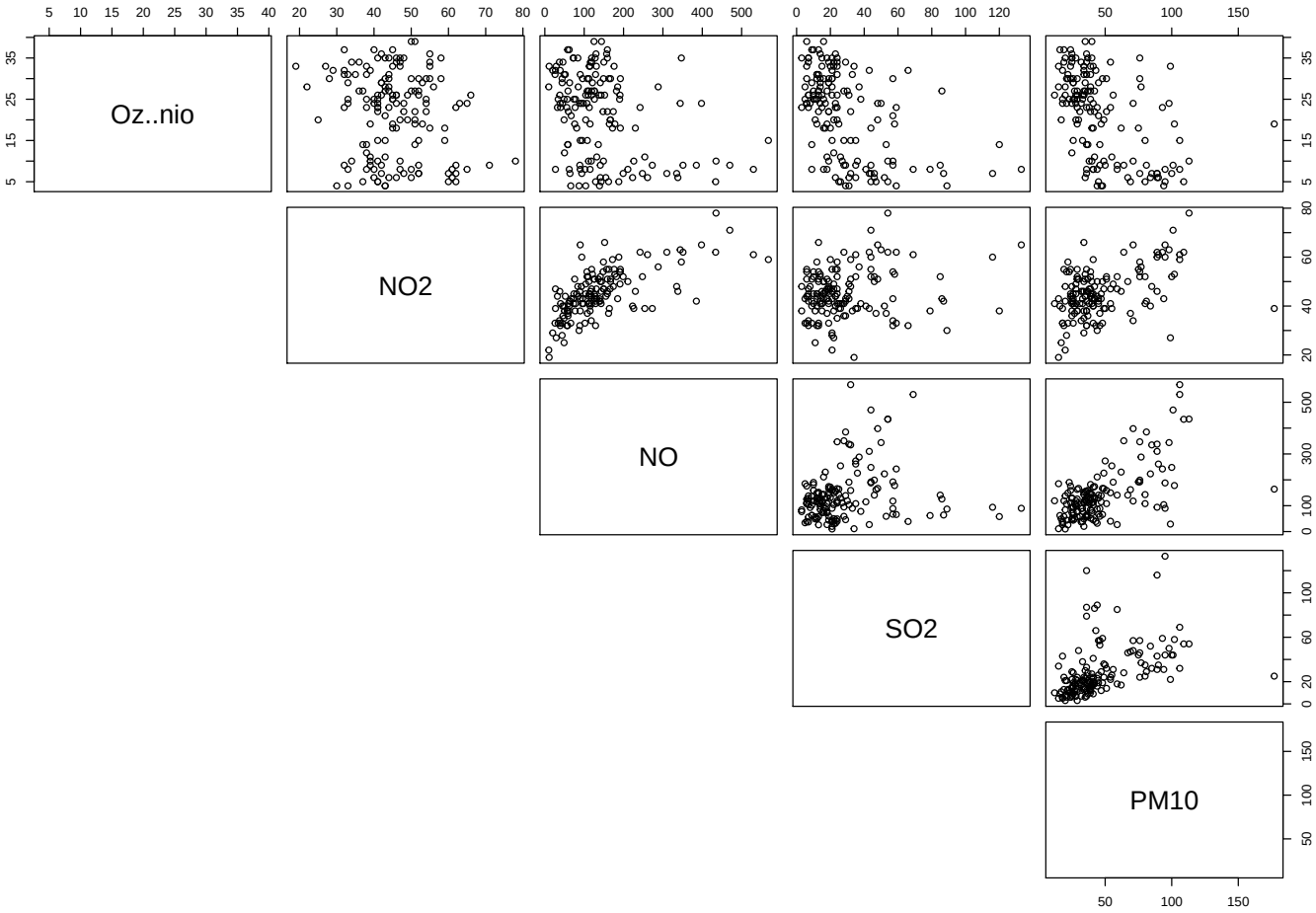
Exemplo: Dados `winter` do pacote `texmex` do R.



Medidas de associação

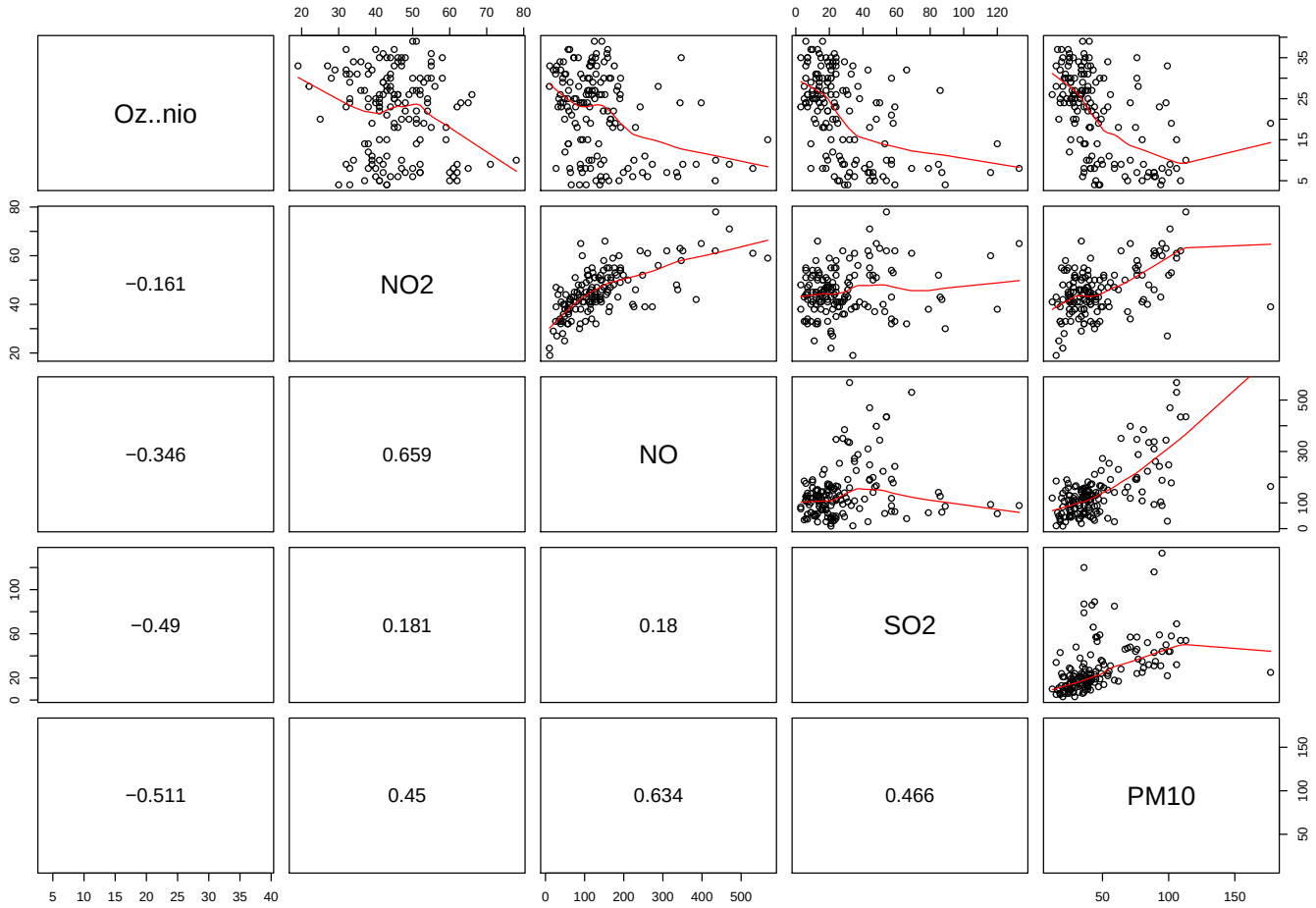
Coeficiente de correlação linear

Exemplo: Dados `winter` do pacote `texmex` do R.



Medidas de associação Coeficiente de correlação linear

Exemplo: Dados `winter` do pacote `texmex` do R.



Medidas de associação Coeficiente de correlação linear

Exemplo: Dados `winter` do pacote `texmex` do R.

