Estatística Aplicada a Ciências Ambientais

Correlação e Regressão

Daniel Detzel
detzel@ufpr.br

Agenda

Correlação

- definições
- medidas: Pearson e Spearman

Regressão

- regressão linear simples
- regressão linear múltipla

Aplicações em R

Correlação

definições

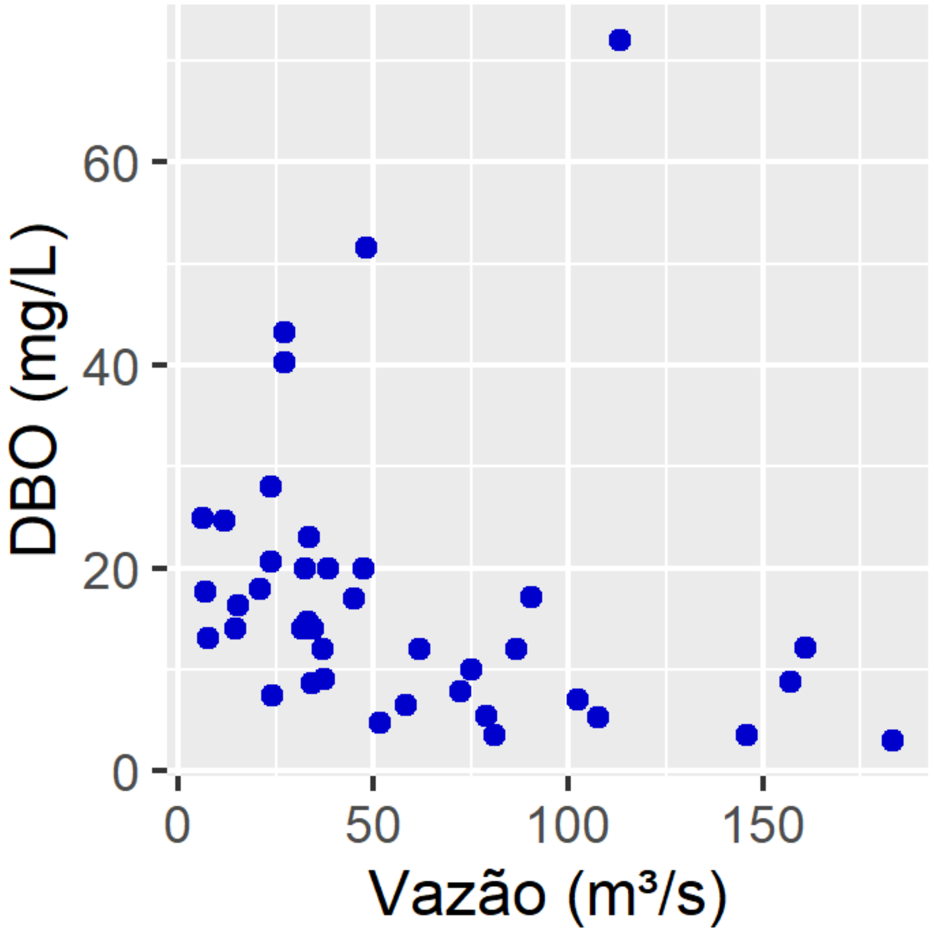
Correlação | definições

Análises de correlação buscam estudar a associação entre duas variáveis aleatórias

Ex.: DBO vs. Vazão no posto IG2 (rio Iguaçu)
(este é um gráfico de dispersão)

Pontos a avaliar:

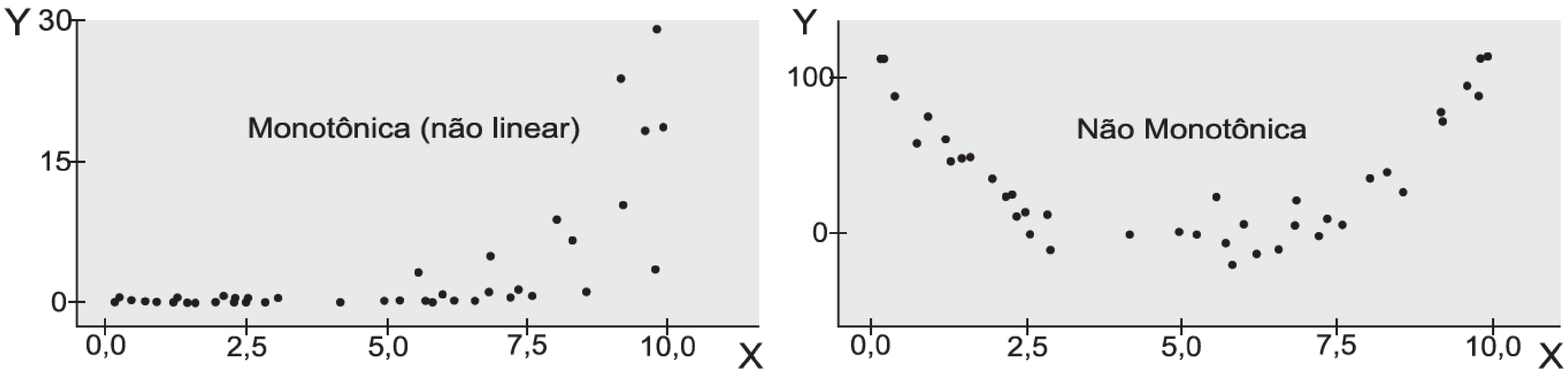
há um padrão de comportamento conjunto?
a variação positiva de uma variável está associada à variação positiva da outra?
e a variação negativa?



Correlação | definições

A correlação é uma medida numérica do grau de associação entre as variáveis em estudo

antes de apresentar as medidas, é preciso entender o conceito das relações monotônicas e não monotônicas



Naghettini e Pinto (2007, p. 356)

O estudo de correlações é tipicamente feito para relações monotônicas

Correlação | definições

Além disso, é imperativo entender que correlações não indicam necessariamente em relações causa-efeito

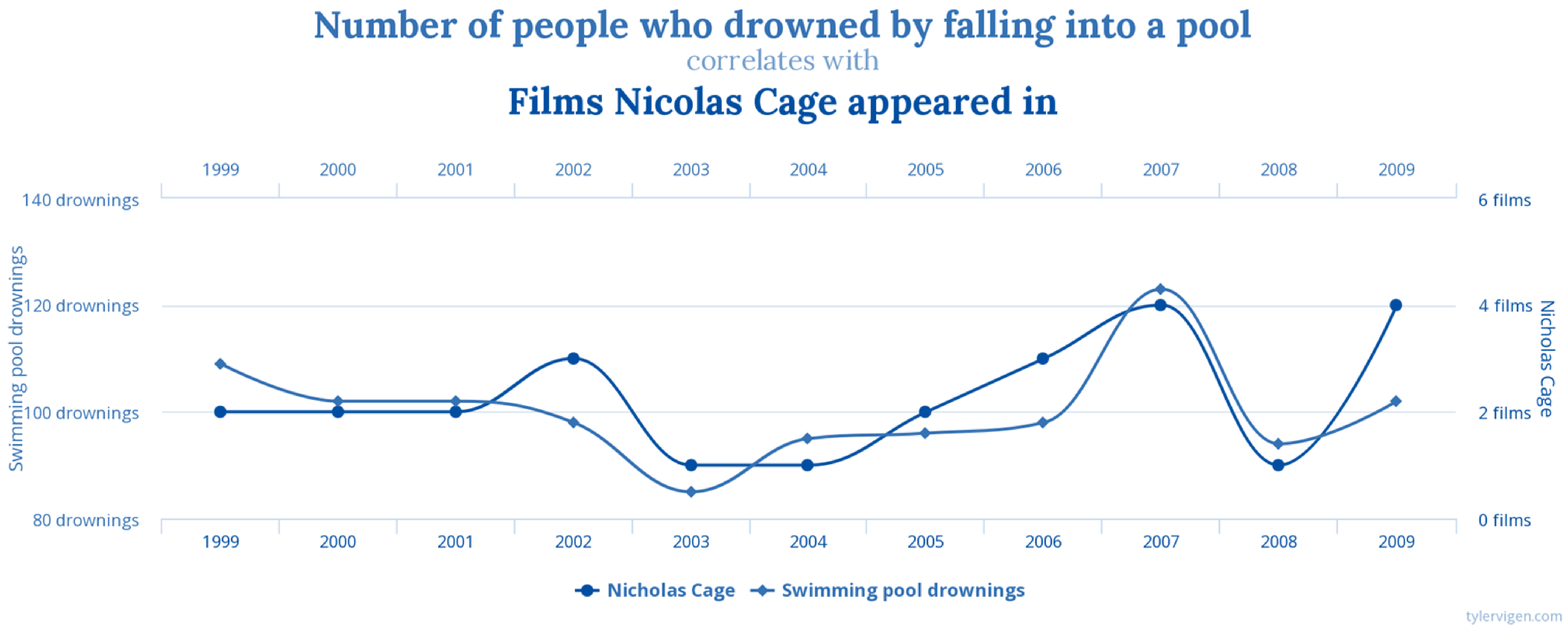
As variáveis podem ter associações positivas: uma aumenta, enquanto a outra também aumenta

Porém, não se pode afirmar que uma aumenta porque a outra aumenta

Fortes correlações sem significado físico podem existir

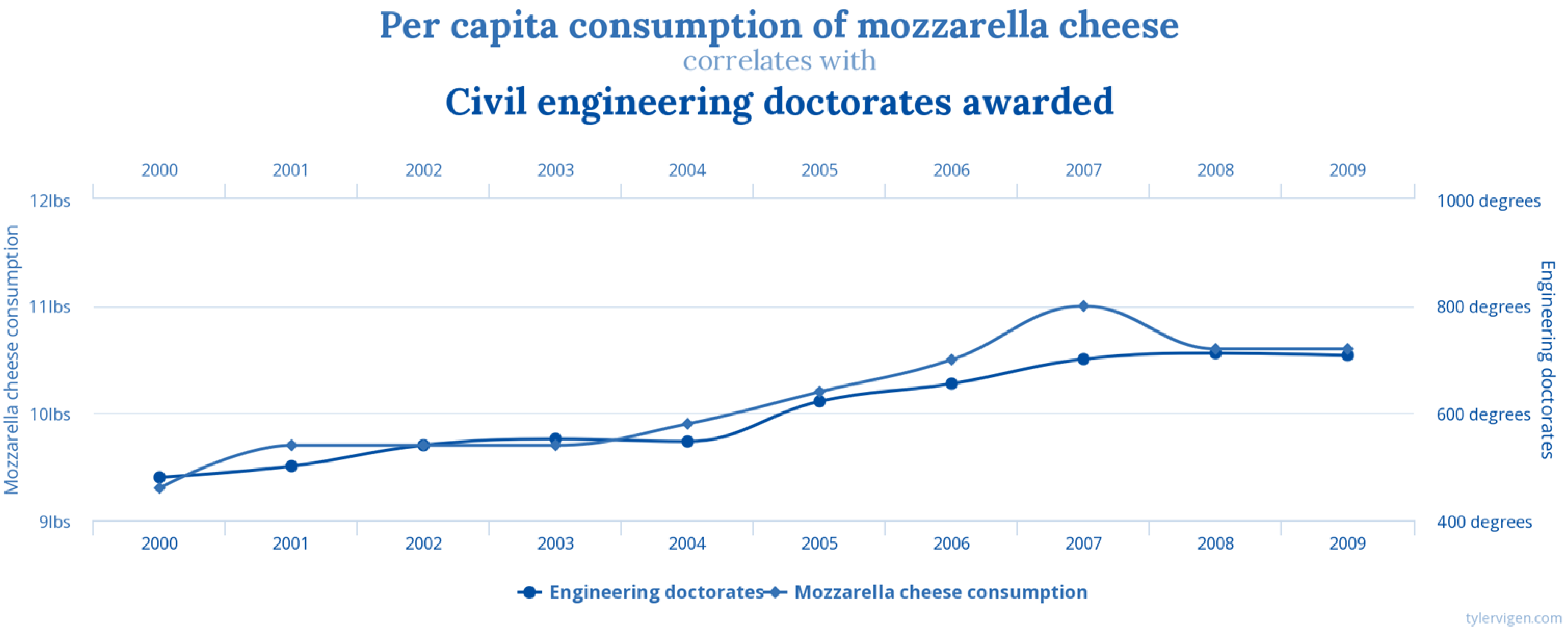
a elas dá-se o nome de correlações espúrias

Correlação | definições



Adaptado de: https://tylervigen.com/spurious-correlations

Correlação | definições



Adaptado de: https://tylervigen.com/spurious-correlations

Correlação

medidas: Pearson e Spearman

Correlação | medidas

As medidas de correlação são feitas por coeficientes adimensionais, definidos para variar entre -1 e 1

é preciso que as amostras tenham o mesmo tamanho

Quanto ao sinal do coeficiente

valores positivos indicam que as duas variáveis aumentam conjuntamente

valores negativos indicam que uma variável aumenta enquanto a outra diminui, ou vice-versa

Quanto ao valor do coeficiente

valores mais próximos de 1 (ou -1) indicam correlações mais fortes

valores próximos de zero indicam não associação (ou independência)

Correlação | coeficiente de Pearson

Coeficiente de correlação de Pearson: a mais bem conhecida medida de correlação

Mede a associação linear entre as variáveis

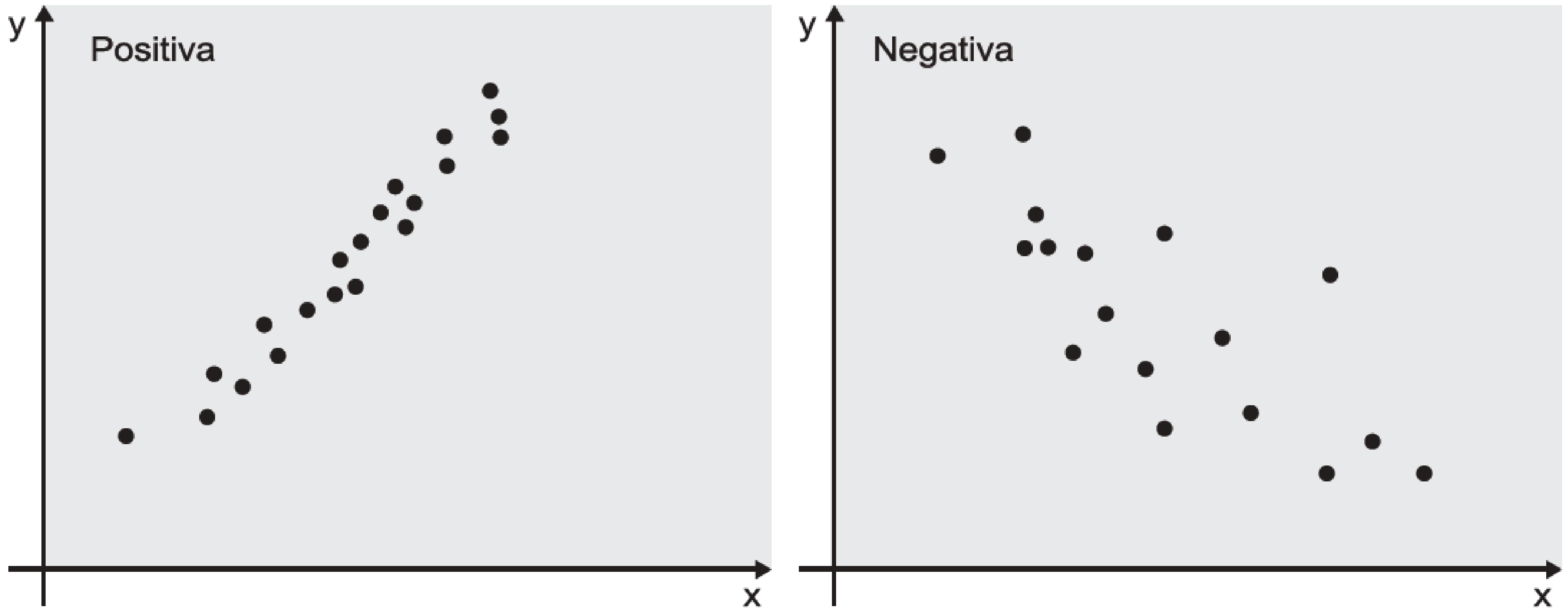
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

onde

$\bar{x}$ e $\bar{y}$	médias de x e y, respectivamente
s_x e s_y	desvios padrão de x e y, respectivamente
n	tamanho da amostra

Correlação | coeficiente de Pearson

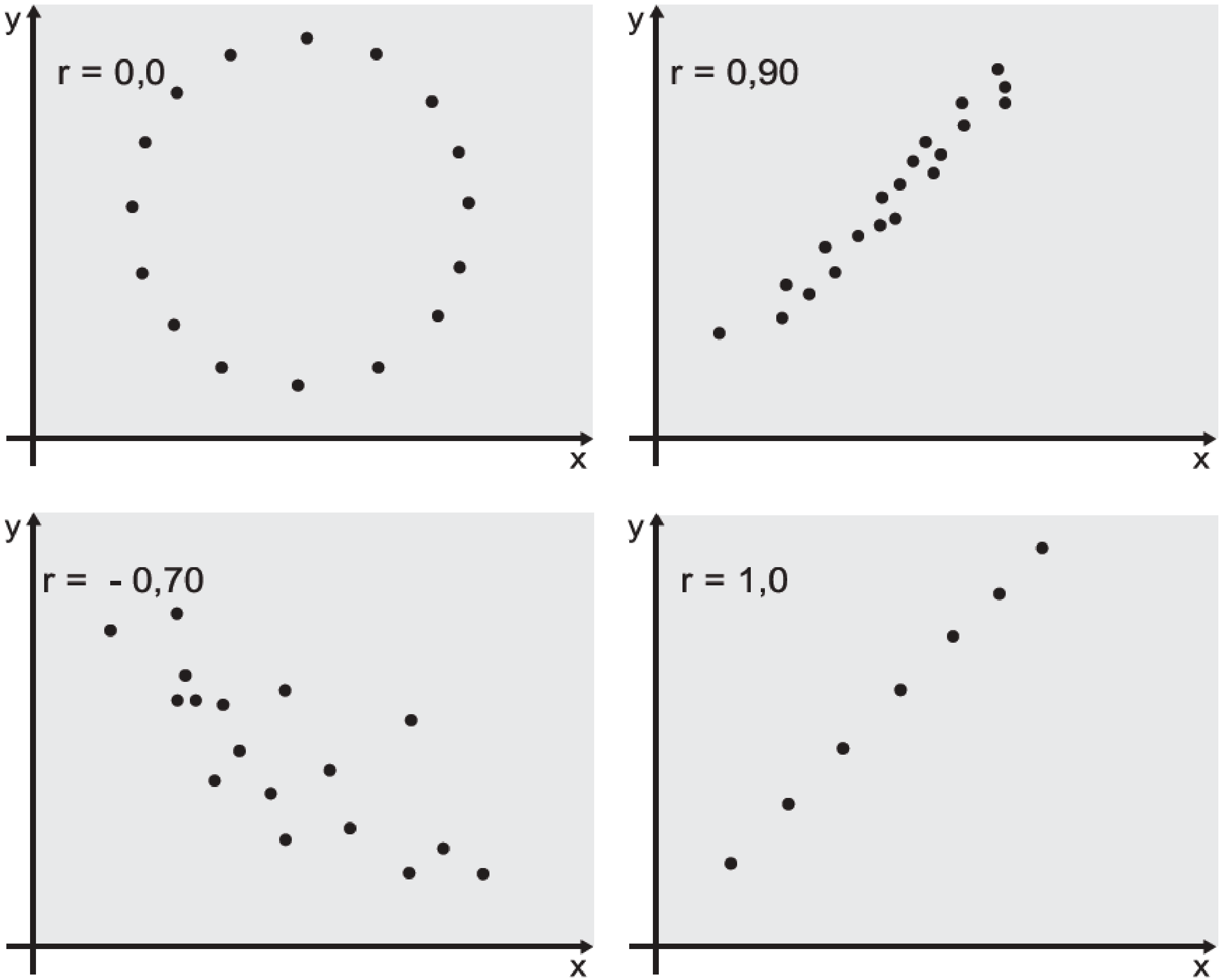
Ex.: correlações lineares positivas e negativas



Naghettini e Pinto (2007, p. 357)

Correlação | coeficiente de Pearson

Ex.: correlações lineares com diferentes valores



Naghettini e Pinto (2007, p. 359)

Correlação | coeficiente de Pearson

Teste de hipótese para r

verifica a existência, ou não, de associação linear

não verifica o grau de associação (ou seja, se r é grande ou pequeno)

Hipóteses do teste:

$H_0: r = 0$ | não há associação linear

$H_A: r \neq 0$ | há associação linear

A estatística do teste é calculada por:

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Correlação | coeficiente de Pearson

Rejeita-se H_0, se $|t_0| > t_{\alpha/2,n-1}$ (teste bilateral)

No R:

Cálculo da correlação

```
cor(amostra1,amostra2,method = "pearson")
```

Teste de hipótese sobre o coeficiente

```
cor.test(amostra1,amostra2,method = "pearson")
```

Correlação | coeficiente de Pearson

Ex.: correlação entre DBO e vazão no posto IG2 (rio Iguaçu)

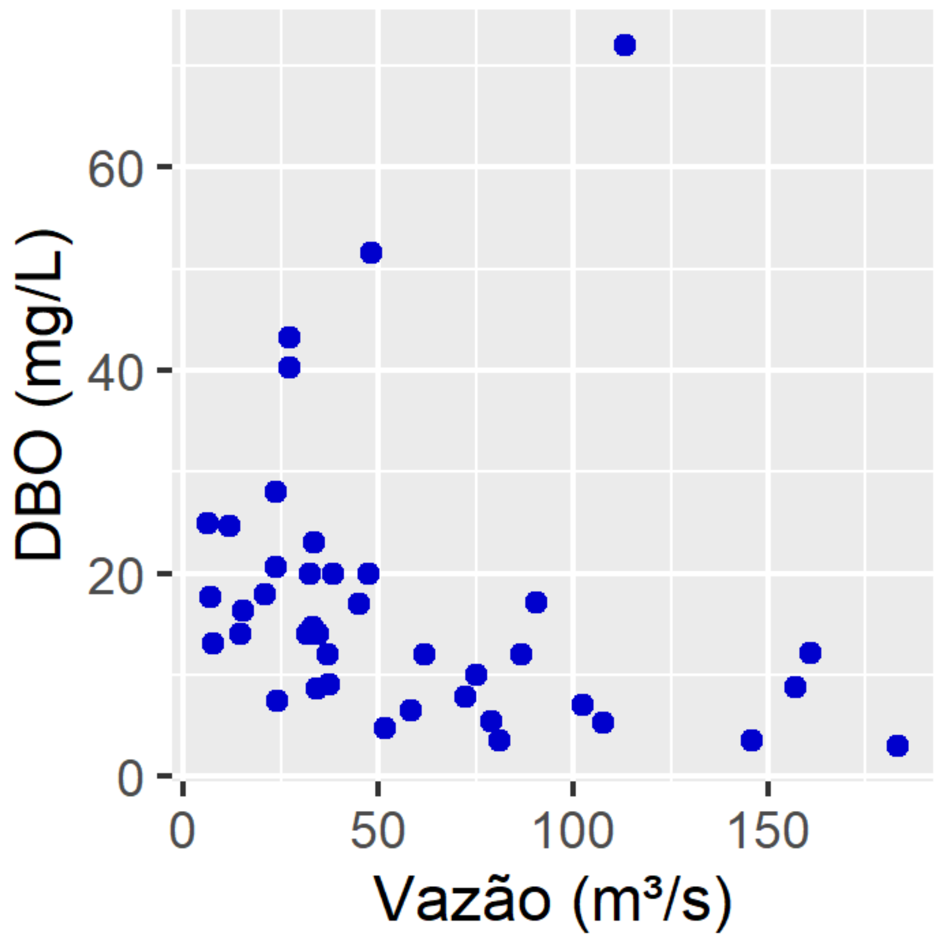
Valor do coeficiente

```
cor(DBO,Q,method = "pearson")
[1] -0.211435
```

Teste de hipótese sobre o coeficiente

```
cor.test(DBO,Q,method = "pearson")
p-value = 0.1845
```

Para $\alpha = 5\%$, não se rejeita H_0 (portanto, não há associação linear entre as variáveis)



Correlação | coeficiente de Speaman

Coeficiente de correlação de Spearman: uma medida não paramétrica de correlação

Mede a associação monotônica entre as variáveis (linear ou não)

O cálculo é feito sobre ranques associados aos elementos das amostras

portanto, o primeiro passo para sua determinação é ranquear as amostras de forma independente

Correlação | coeficiente de Spearman

Na sequência, aplica-se:

$$\rho = \frac{\sum_{i=1}^{n}(Rx_i Ry_i) - n\left(\frac{n+1}{2}\right)^2}{\frac{n(n^2-1)}{12}}$$

onde

Rx_i e Ry_i ranques das amostras x e y, respectivamente

Alternativamente, é possível aplicar a própria equação do coef. de Pearson substituindo x e y pelos seus ranques Rx_i e Ry_i

Correlação | coeficiente de Spearman

Teste de hipótese para ρ

verifica a existência, ou não, de associação monotônica

não verifica o grau de associação (ou seja, se ρ é grande ou pequeno)

Hipóteses do teste:

$H_0: \rho = 0$ | não há associação monotônica (linear ou não)

$H_A: \rho \neq 0$ | há associação monotônica (linear ou não)

A estatística do teste é calculada por:

$$S = \sum_{i=1}^{n} (Rx_i - Ry_i)^2$$

Correlação | coeficiente de Spearman

Rejeita-se H_0, se $|S| > t_{\alpha/2,n-2}$ (teste bilateral)

No R:

Cálculo da correlação

```
cor(amostra1,amostra2,method = “spearman")
```

Teste de hipótese sobre o coeficiente

```
cor.test(amostra1,amostra2,method = "spearman“, exact =
FALSE, continuity = TRUE)
```

Correlação | coeficiente de Spearman

Ex.: correlação de Spearman entre DBO e vazão no posto IG2 (rio Iguaçu)

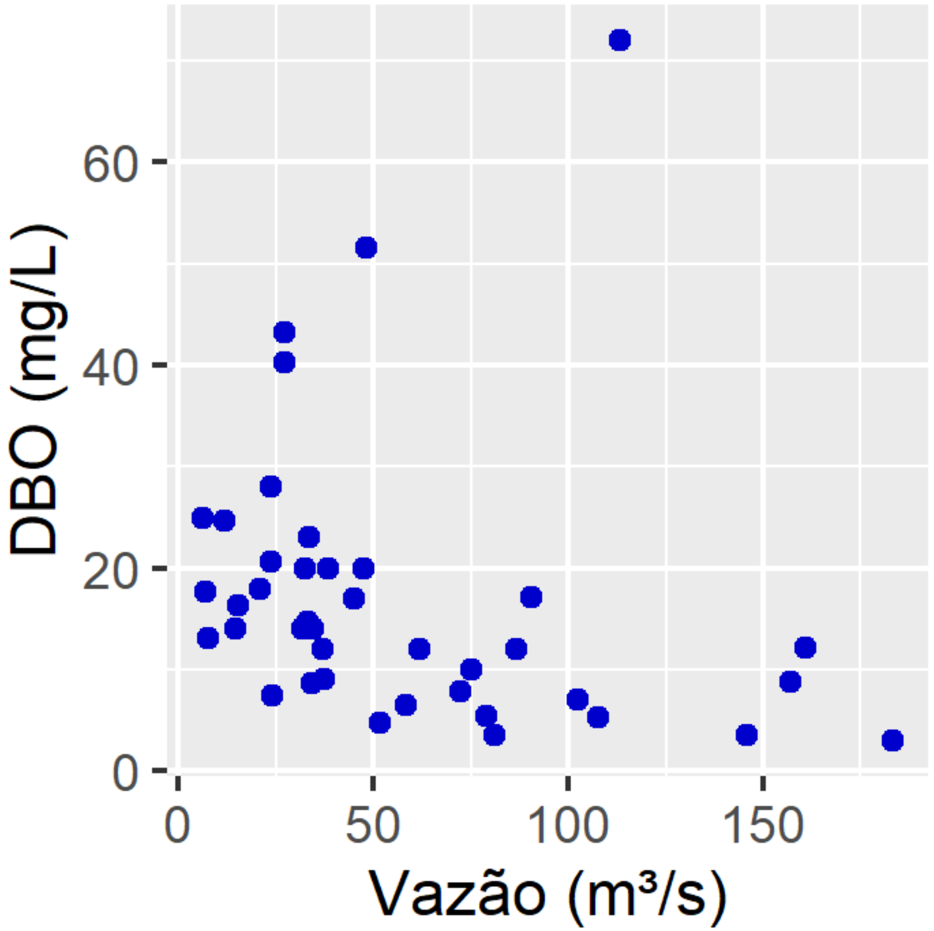
Valor do coeficiente

```
cor(DBO,Q,method = "spearman")
[1] -0.5391223
```

Teste de hipótese sobre o coeficiente

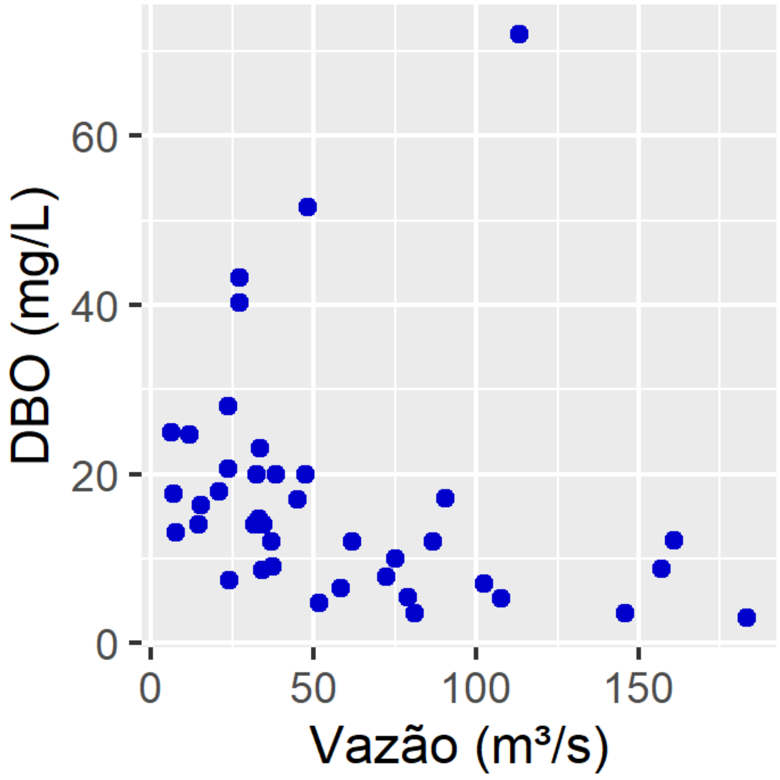
```
cor.test(DBO,Q,method = "spearman",
    exact = FALSE, continuity = TRUE)
p-value = 0.0002771
```

Para $\alpha = 5\%$, rejeita-se H_0 (portanto, há associação monotônica entre as variáveis)

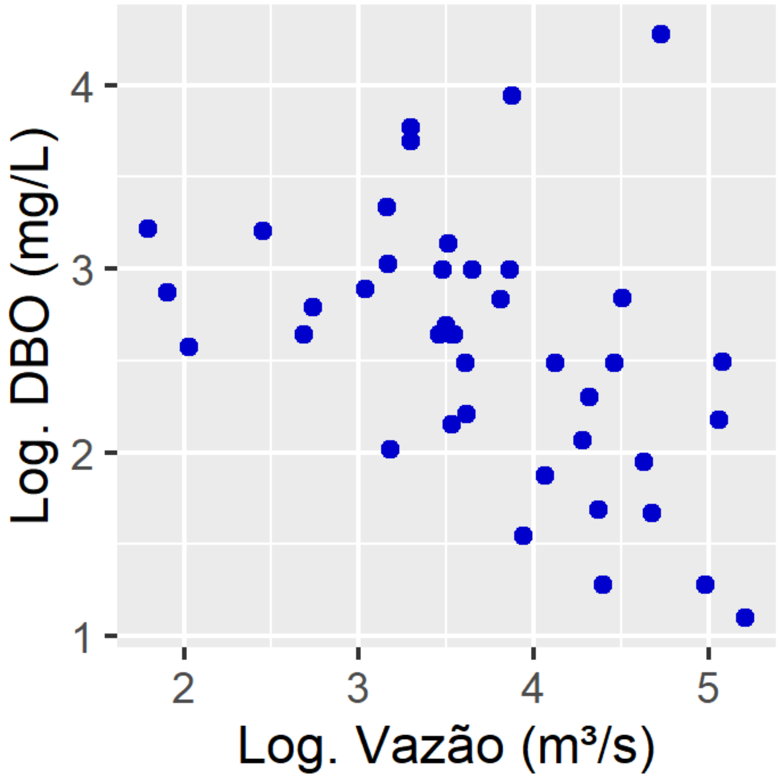


Correlação | Pearson vs. Spearman

As diferenças entre os coeficientes ficam ainda mais claras se aplica uma transformação não linear sobre os dados



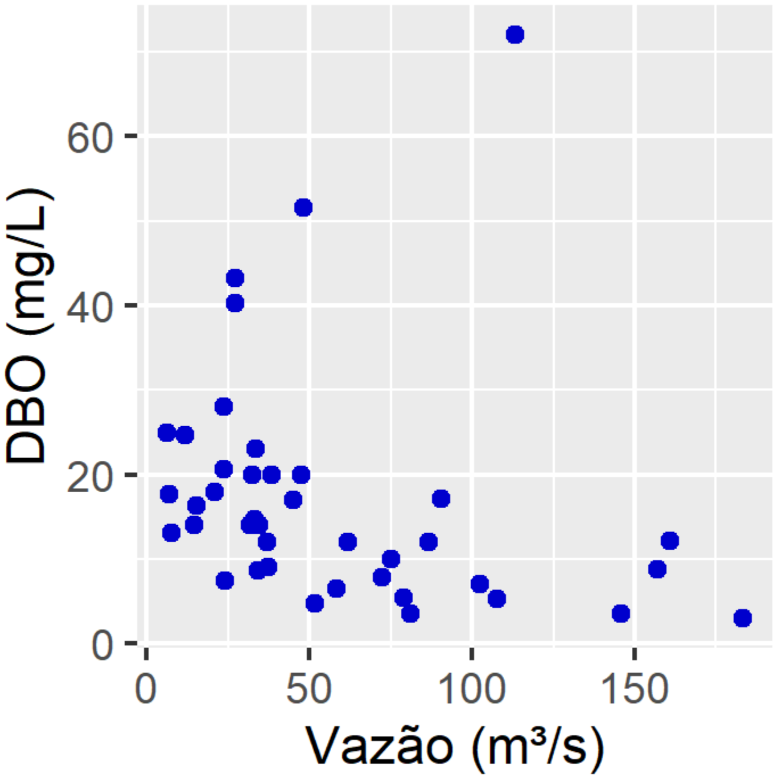
Escala original



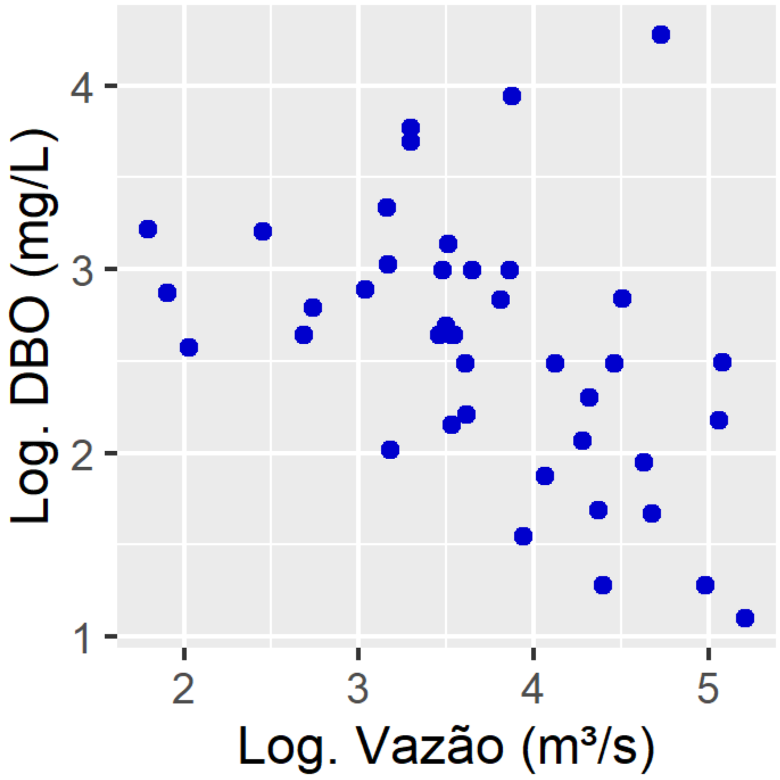
Escala logarítmica

Correlação | Pearson vs. Spearman

O coeficiente de Spearman não é afetado pela transformação



$$r = -0{,}211$$
$$\rho = -0{,}539$$



$$r = -0{,}441$$
$$\rho = -0{,}539$$

REGRESSÃO

regressão linear simples

Regressão | regressão linear simples

Técnica estatística que visa avaliar e modelar a relação entre variáveis

- admite-se a existência de uma função que explica o comportamento médio da variação de uma variável em relação a outra
- a porção não explicada (fora do comportamento médio) é atribuída a variações residuais

Regressão linear simples (RLS)

- lida com uma variável a ser explicada (dependente) e apenas uma variável explicativa (ou independente)
- a relação funcional entre as variáveis é representada por uma reta

Regressão | regressão linear simples

O modelo de RLS é dado por:

$$\hat{y}_i = a + bx_i + e_i$$

$$i = 1,2, \dots, n$$

onde

y_i	observação i da variável dependente
x_i	observação i da variável independente
a	parâmetro de intercepto da reta
b	parâmetro do ângulo da reta
e_i	erro, ou resíduo, da observação i
n	tamanho da amostra

Regressão | regressão linear simples

Premissas do modelo de RLS

- assume-se que y é linearmente relacionado com x
- a variância da série de resíduos é constante (homocedástica)
- os resíduos são independentes entre si
- os resíduos são normalmente distribuídos

Em conjunto, as premissas fazem com que a RLS seja uma técnica paramétrica

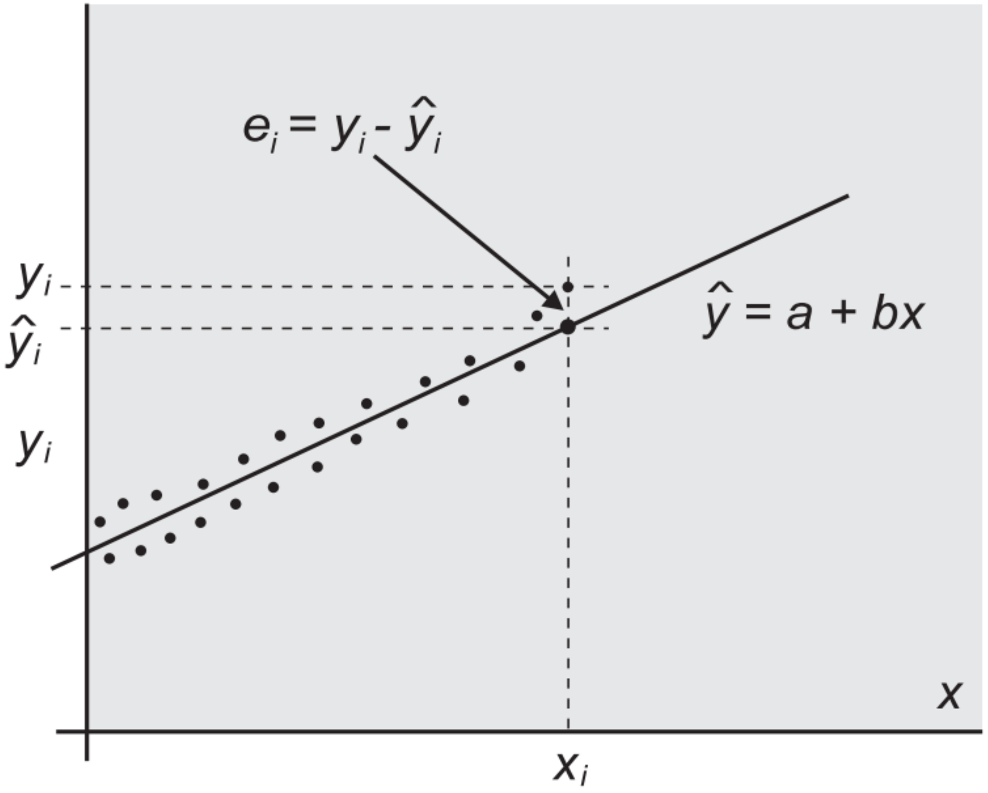
- porém não requer que nem y nem x sejam normalmente distribuídos!

Estimação dos coeficientes: Método dos Mínimos Quadrados

determina os coeficientes que produzem valor mínimo da função de soma dos quadrados dos resíduos

$$e_i = y_i - \hat{y}_i$$
$$e_i = y_i - a + bx_i$$

$$\min \sum_{i=1}^{n} e_i^2 = \min \sum_{i=1}^{n} (y_i - a + bx_i)^2$$



Naghettini e Pinto (2007, p. 364)

Regressão | estimadores da regressão

Estatística	Equação
Soma dos quadrados de y (soma total – SS)	$SS_y = \sum_{i=1}^{n} y_i^2 - n\mu_y^2$
Soma dos quadrados de x	$SS_x = \sum_{i=1}^{n} x_i^2 - n\mu_x^2$
Soma dos produtos cruzados entre x e y	$SS_{xy} = \sum_{i=1}^{n} x_i y_i - n\mu_x\mu_y$

Regressão | estimadores da regressão

Estatística	Equação
Estimativa de α ($\hat{\alpha} = a$)	$a = \mu_y - b\mu_x$
Estimativa de β ($\hat{\beta} = b$)	$b = S_{xy}/SS_x$
Erro médio quadrático (MSE – variância dos resíduos)	$\sigma_e^2 = \sum_{i=1}^{n} \frac{e_i^2}{n-2} = \frac{SS_y - bS_{xy}}{n-2}$

Regressão | estimadores da regressão

Estatística	Equação
Erro padrão da regressão (desvio padrão dos resíduos)	$\sigma_e = \sqrt{\sigma_e^2}$
Erro padrão de a	$SE_a = \sigma_e \sqrt{\frac{1}{n} + \frac{\mu_x^2}{SS_x}}$
Erro padrão de b	$SE_b = \frac{\sigma_e}{\sqrt{SS_x}}$

Regressão | estimadores da regressão

Estatística	**Equação**
Coeficiente de correlação	$r = \frac{S_{xy}}{\sqrt{SS_x SS_y}}$
Coeficiente de determinação	$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{SS_y}$
Coeficiente de determinação ajustado (p é o número de variáveis explicativas)	$R_a^2 = 1(1 - R^2)\frac{n-1}{n-p-1}$

O coeficiente de determinação (ajustado ou não) exprime a parcela da variância de y que foi explicada pela reta da regressão ($-\infty \leq R^2 \leq 1$)

Regressão | estimadores da regressão

Intervalos de confiança para os estimadores a e b

$$a - t_{1-\frac{\alpha}{2},n-2} SE_a \le \alpha \le a + t_{1-\frac{\alpha}{2},n-2} SE_a$$

$$b - t_{1-\frac{\alpha}{2},n-2} SE_b \le \beta \le b + t_{1-\frac{\alpha}{2},n-2} SE_b$$

onde

$t_{1-\frac{\alpha}{2},n-2}$ valor da variável *t*-Student para $(1-\alpha/2)$ e $n-2$ graus de liberdade

Teste de hipótese para a

verifica a existência, ou não, de relação linear

Hipóteses do teste:

$H_0: b = 0$ | não há relação linear
$H_A: b \neq 0$ | há relação linear

A estatística do teste é calculada por:

$$t_0 = \frac{b}{SE_b}$$

Regressão | estimadores da regressão

Rejeita-se H_0, se $|t_0| > t_{\alpha/2,n-2}$ (teste bilateral)

No R:

Ajuste por meio da função “lm”

```
reg <- lm(amostra1 ~ amostra2, data = dados)
```

Exibir os resultados do ajuste

```
Summary(reg)
```

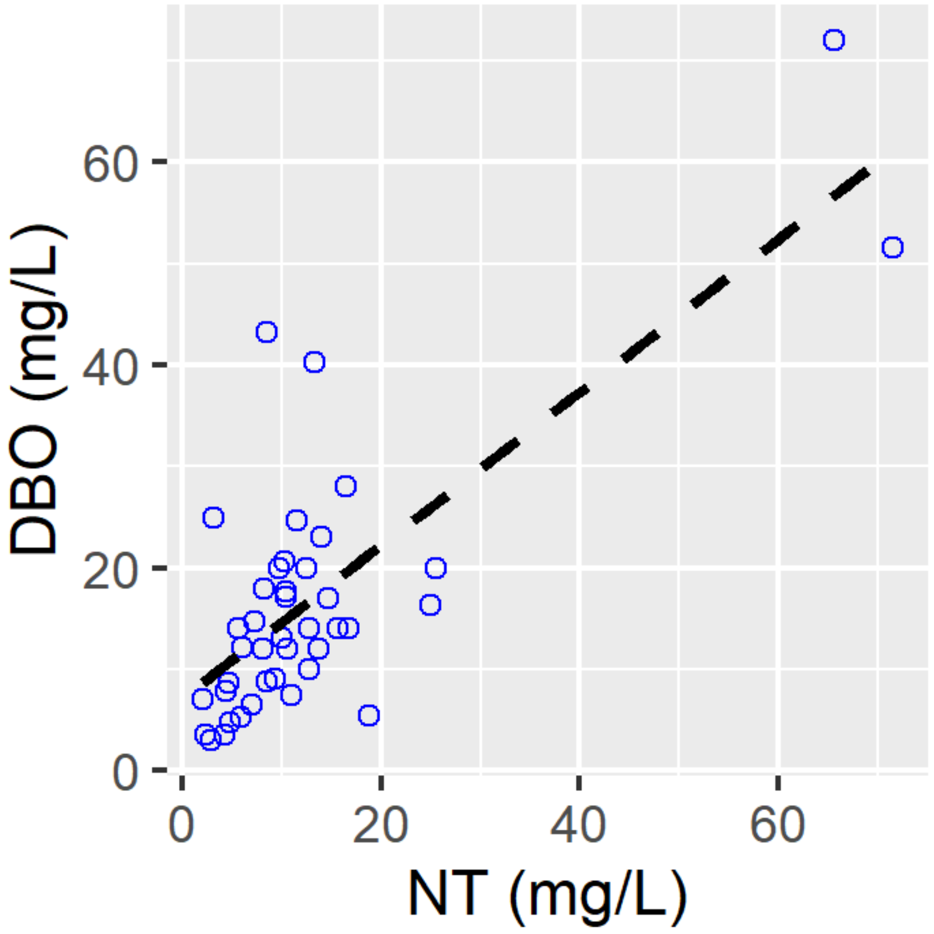
<u>Exemplo</u>: Obtenção do modelo de regressão que explica a DBO utilizando o NT (nitrogênio total)

Ajuste gráfico para avaliação inicial

importante saber quem é x e quem é y!

DBO: variável dependente → y

NT: variável independente → x



Regressão | DBO vs. NT

Em sendo `dadosReg` o `data.frame` que contém das variáveis, a sintaxe é:

```
reg <- lm(DBO ~ NT, data = dadosReg)
```

Os resultados obtidos são:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.1320     1.9448   3.667  0.00073 ***
NT            0.7548     0.1027   7.347 7.19e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.001 on 39 degrees of freedom
Multiple R-squared:  0.5805,  Adjusted R-squared:  0.5698
F-statistic: 53.98 on 1 and 39 DF,  p-value: 7.192e-09
```

Regressão | DBO vs. NT

Por partes.... (destaques em azul)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.1320     1.9448   3.667  0.00073 ***
NT            0.7548     0.1027   7.347 7.19e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.001 on 39 degrees of freedom
Multiple R-squared:  0.5805,  Adjusted R-squared:  0.5698
F-statistic: 53.98 on 1 and 39 DF,  p-value: 7.192e-09
```

Expressam os coeficientes da regressão. Portanto:

$$DBO = 7{,}1320 + 0{,}7548 \cdot NT$$

Regressão | DBO vs. NT

Por partes.... (destaques em azul)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.1320     1.9448   3.667  0.00073 ***
NT            0.7548     0.1027   7.347 7.19e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.001 on 39 degrees of freedom
Multiple R-squared:  0.5805,  Adjusted R-squared:  0.5698
F-statistic: 53.98 on 1 and 39 DF,  p-value: 7.192e-09
```

Expressam os erros padrão dos coeficientes da regressão (SE_a e SE_b)

Regressão | DBO vs. NT

Por partes.... (destaques em azul)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.1320     1.9448   3.667  0.00073 ***
NT            0.7548     0.1027   7.347 7.19e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.001 on 39 degrees of freedom
Multiple R-squared:  0.5805,  Adjusted R-squared:  0.5698
F-statistic: 53.98 on 1 and 39 DF,  p-value: 7.192e-09
```

Expressam as estatísticas relativas ao teste de hipótese para o coeficiente b

lembrando que $H_0\colon b = 0$ | não há relação linear

como em ambos os p-valores foram baixos ($< 5\%$), rejeita-se H_0

Regressão | DBO vs. NT

Por partes.... (destaques em azul)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.1320     1.9448   3.667  0.00073 ***
NT            0.7548     0.1027   7.347 7.19e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.001 on 39 degrees of freedom
Multiple R-squared:  0.5805,  Adjusted R-squared:  0.5698
F-statistic: 53.98 on 1 and 39 DF,  p-value: 7.192e-09
```

Expressa o erro padrão dos resíduos

portanto, a regressão resultou em um erro de ~9 mg/L na DBO

Regressão | DBO vs. NT

Por partes.... (destaques em azul)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.1320     1.9448   3.667  0.00073 ***
NT            0.7548     0.1027   7.347 7.19e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.001 on 39 degrees of freedom
Multiple R-squared:  0.5805,  Adjusted R-squared:  0.5698
F-statistic: 53.98 on 1 and 39 DF,  p-value: 7.192e-09
```

Expressam os coeficientes de determinação

são valores regulares (nem muito bom, nem muito ruim)

Regressão | DBO vs. NT

Por partes.... (destaques em azul)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.1320     1.9448   3.667  0.00073 ***
NT            0.7548     0.1027   7.347 7.19e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.001 on 39 degrees of freedom
Multiple R-squared:  0.5805,  Adjusted R-squared:  0.5698
F-statistic: 53.98 on 1 and 39 DF,  p-value: 7.192e-09
```

Expressa a significância geral do modelo de regressão

com o p-valor baixo ($< 5\%$), o modelo é estatisticamente significativo

Regressão | DBO vs. NT

Os intervalos de confiança dos parâmetros são obtidos pela função:

```
confint(reg)
```

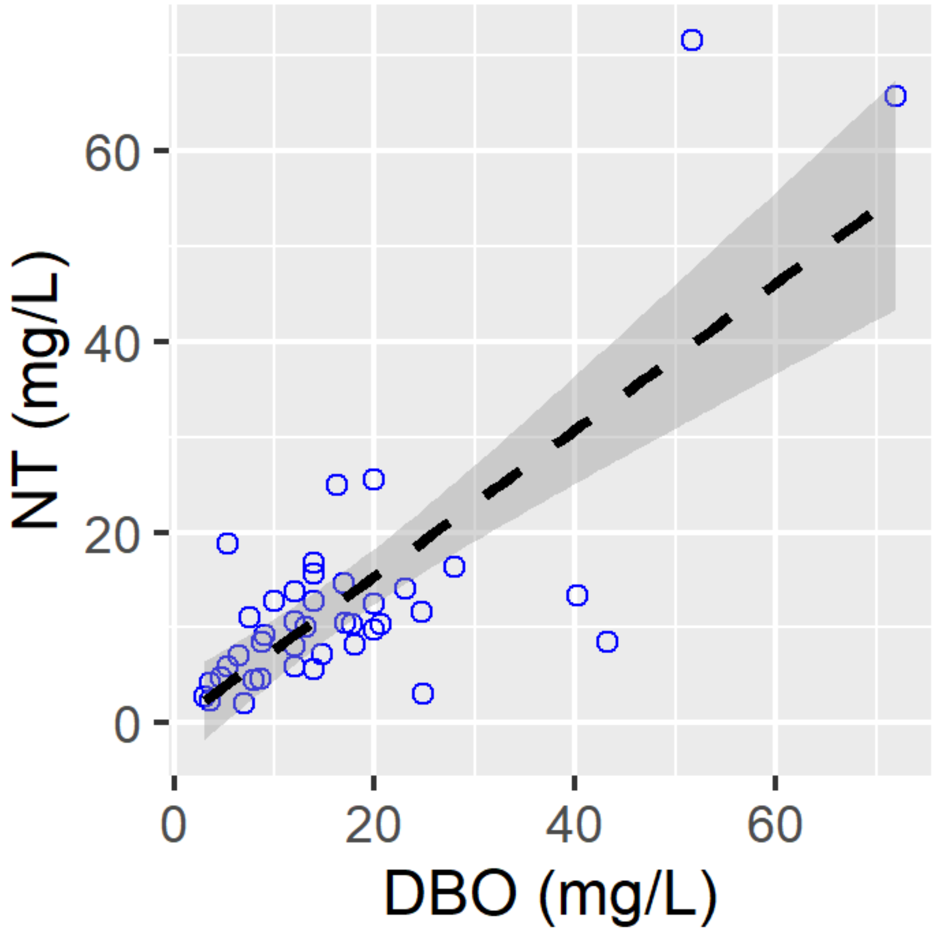
Que fornece como resultados:

```
                2.5 %     97.5 %
(Intercept) 3.1983461 11.0656312
NT          0.5469887  0.9625992
```

(lembrando que os parâmetros resultaram em 7,1320 e 0,7548)

Regressão | DBO vs. NT

Graficamente, os intervalos de confiança resultam em:



REGRESSÃO

regressão linear múltipla

Regressão linear múltipla (RLM)

- lida com uma variável a ser explicada (dependente) e mais de uma variável explicativa (ou independente)
- a relação funcional entre as variáveis continua sendo representada por uma reta

A escolha em passar para uma análise multivariada pode ser feita com base em:

- experiência do analista
- resultados insuficientes com a RLS

Regressão | regressão linear múltipla

O modelo de RLM é dado por:

$$\hat{y}_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \cdots + b_k x_{i,k} + e_i$$

$$i = 1,2,\ldots,n$$

onde

y_i	observação i da variável dependente
$x_{i,k}$	observação i da variável explicativa k
b_0	parâmetro de intercepto da reta
b_k	parâmetro do ângulo da reta para a variável explicativa k
e_i	erro, ou resíduo, da observação i
n	tamanho da amostra

Regressão | regressão linear múltipla

Para a RLM, evita-se se referir às variáveis explicativas como variáveis dependentes

- as variáveis explicativas podem ser dependentes entre si
- ainda assim, a correlação entre essas variáveis pode ser um problema

Variáveis explicativas fortemente correlacionadas entre si causam multicolinearidade

- variáveis colineares não agregam nenhuma informação nova ao modelo
- podem dificultar a interpretação dos coeficientes da RL
- a avaliação pode ser feita por meio de correlações entre as variáveis

Regressão | regressão linear múltipla

<u>Exemplo</u>: deseja-se obter um modelo de RLM para explicar a DBO do posto IG2 no rio Iguaçu. Avaliar a multicolinearidade das variáveis disponíveis.

A matriz de correlações é mostrada →

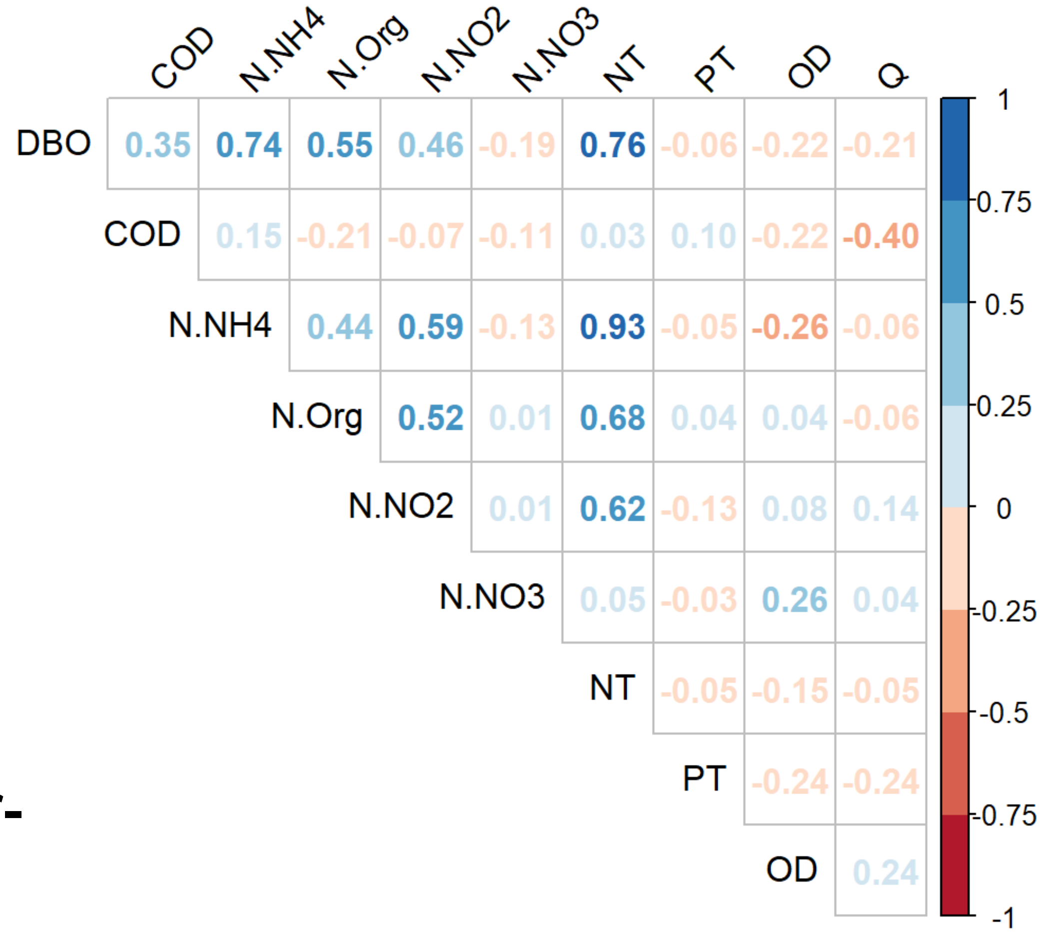
	COD	N.NH4	N.Org	N.NO2	N.NO3	NT	PT	OD	Q
DBO	0.35	0.74	0.55	0.46	-0.19	0.76	-0.06	-0.22	-0.21
COD		0.15	-0.21	-0.07	-0.11	0.03	0.10	-0.22	-0.40
N.NH4			0.44	0.59	-0.13	0.93	-0.05	-0.26	-0.06
N.Org				0.52	0.01	0.68	0.04	0.04	-0.06
N.NO2					0.01	0.62	-0.13	0.08	0.14
N.NO3						0.05	-0.03	0.26	0.04
NT							-0.05	-0.15	-0.05
PT								-0.24	-0.24
OD									0.24

Interpretações:

1. A DBO possui boa correlação com as variáveis NT, N.NH4 e N.Org.
2. N.NO2 e COD possuem correlações mais fracas, mas não desprezíveis

Regressão | regressão linear múltipla

Exemplo: (cont.)



3. N.NH4 e NT são altamente correlacionadas, indicando um possível problema de multicolinearidade
4. Em um modelo de RLM, a escolha deve recair em uma das variáveis
5. As demais correlações não sugerem multicor-linearidade ($r < 0{,}8$)

Regressão | regressão linear múltipla

Sobre a construção de um modelo de RLM

processo por partes: incorporar ao modelo uma variável por vez

inicia com a variável explicativa que tem a maior correlação com a variável independente

inclui-se outra variável explicativa e avalia-se o desempenho do modelo. Caso ele seja pior, descarta-se a variável incluída

No R:

Ajuste por meio da função "lm"

```
reg <- lm(amostra1 ~ amostra2 + amostra3 + … + amostrak, data = dados)
```

Exibir os resultados do ajuste

```
Summary(reg)
```

Regressão | regressão linear múltipla

Exemplo (proposto): ajustar o melhor modelo de RLM para explicar a variação de DBO, usando como critério de qualidade o coeficiente de determinação ajustado.

```
Call:
lm(formula = DBO ~ COD + NOrg + N.NH4 + Q, data = dadosReg)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.740558   4.044539  -0.183 0.855746
COD          1.016486   0.302610   3.359 0.001860 **
NOrg         1.126946   0.293114   3.845 0.000473 ***
N.NH4        0.638564   0.129060   4.948 1.76e-05 ***
Q           -0.005654   0.029813  -0.190 0.850652
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.646 on 36 degrees of freedom
Multiple R-squared:  0.7207,     Adjusted R-squared:  0.6896
F-statistic: 23.22 on 4 and 36 DF,  p-value: 1.498e-09
```

Revisão

Coeficientes de correlação quantificam a associação entre variáveis

contudo, não indicam relação causa-efeito

Coeficiente de Spearman não é afetado por transformações nas séries e é resistente a outliers

Modelos de regressão são utilizados para explicar a variação de uma variável em função de uma (RLS) ou outras variáveis (RLM)

Existem outras técnicas para avaliar modelos de regressão, mas que não foram mostradas em aula (ver Helsel et al., 2020)

Estatística Aplicada a Ciências Ambientais

Daniel Detzel
detzel@ufpr.br