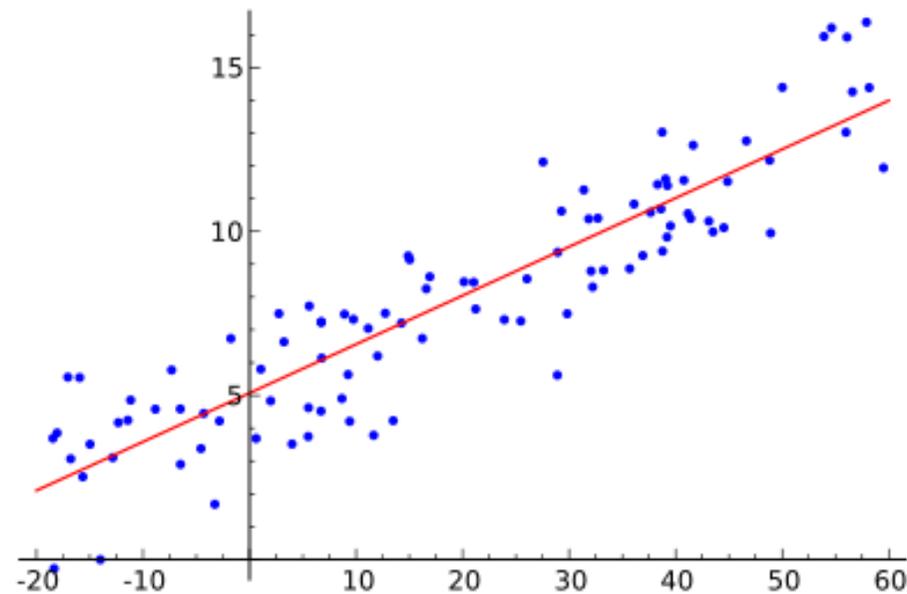


Regressão linear



Camila de Toledo Castanho

2022

Conteúdo da aula

1. Regressão linear simples: quando usar
2. A reta de regressão linear
3. Teste de significância da regressão
4. Coeficiente de determinação (r^2)
5. Pressupostos do teste
6. Procedimentos diagnósticos
7. Roteiro

1. Quando usar?

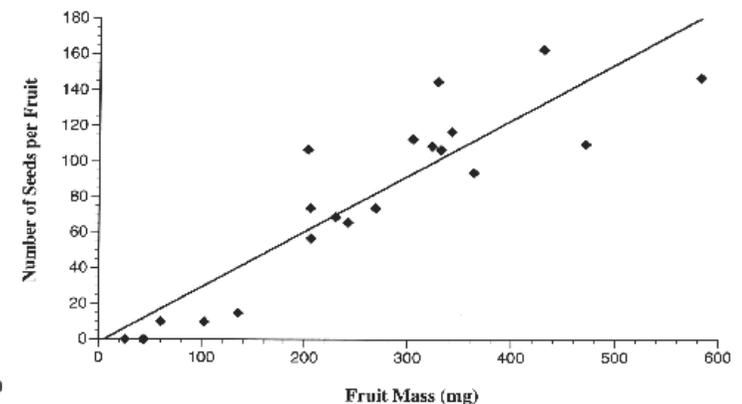
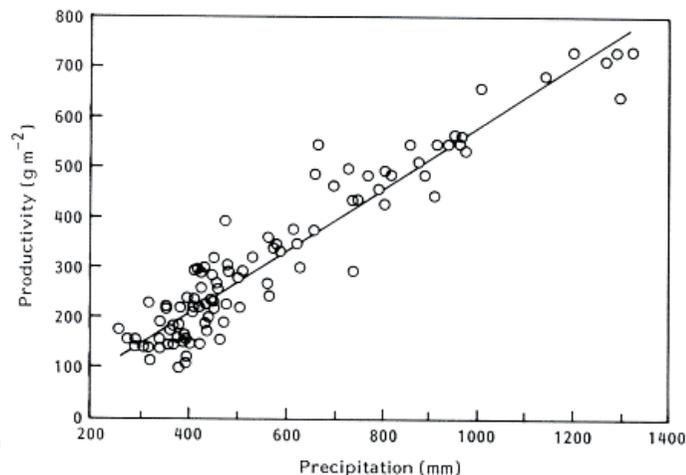
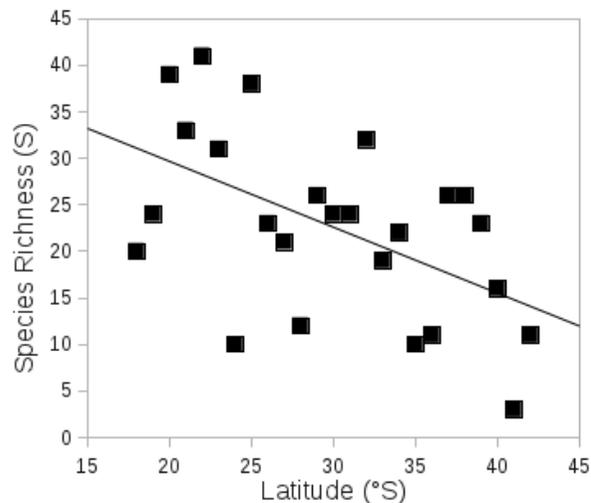
- Suposição de relação de causa-efeito entre duas variáveis contínuas

Eixo X= variável preditora; explicativa ou independente

Eixo Y= variável resposta ou dependente

- ✓ Para cada valor de x observa-se o valor correspondente de y
- ✓ Os valores de x são em geral selecionados no sentido de obter ampla variação desta variável

Objetivos: avaliar possível dependência de y em relação à x e expressar matematicamente essa relação



2. A reta de regressão linear

- Primeiro passo: visualização dos dados → gráfico de dispersão dos pontos

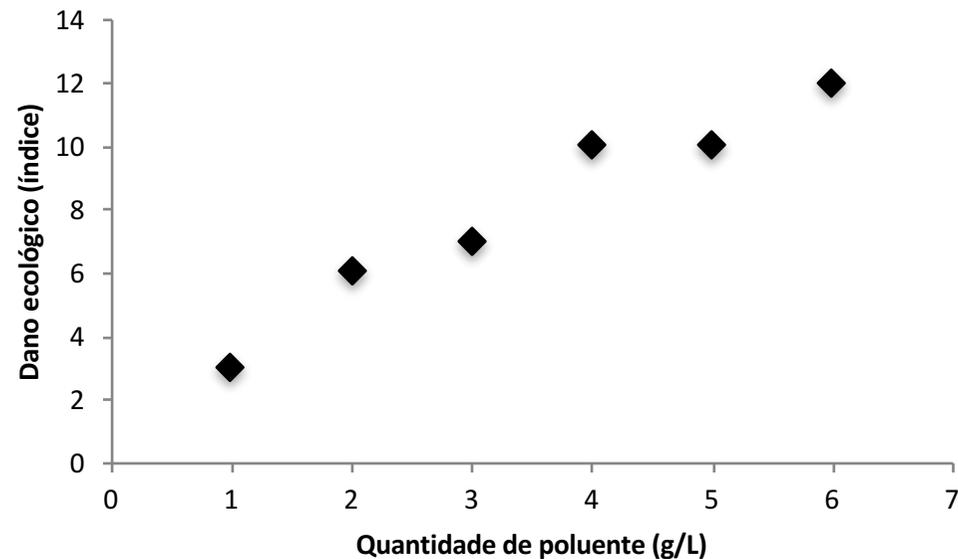


Fornece uma boa idéia da existência de dependência

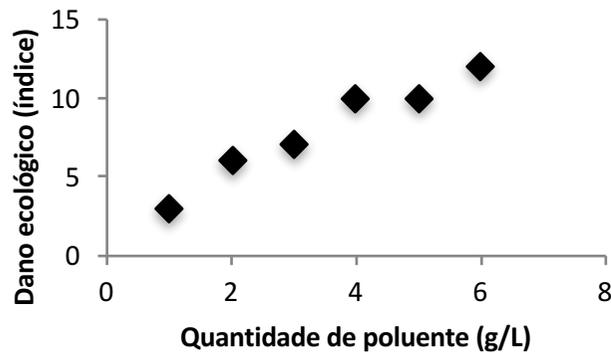
EXEMPLO

Relação entre certo poluente despejado por uma fábrica em um riacho e o dano ecológico na água, medido por um índice.

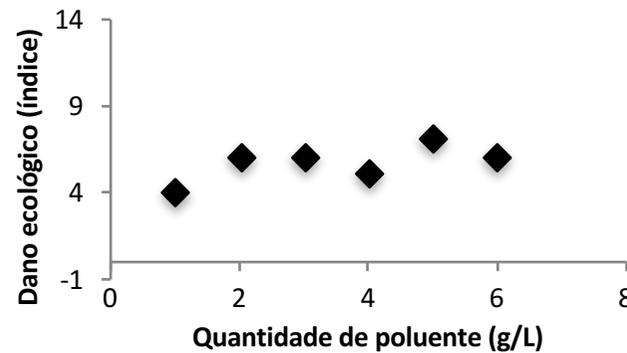
Aparentemente há uma **dependência positiva** de y em relação à x



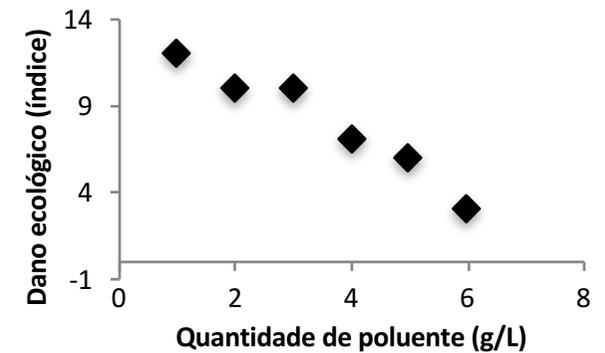
2. A reta de regressão linear



Dependência positiva



Ausência de dependência

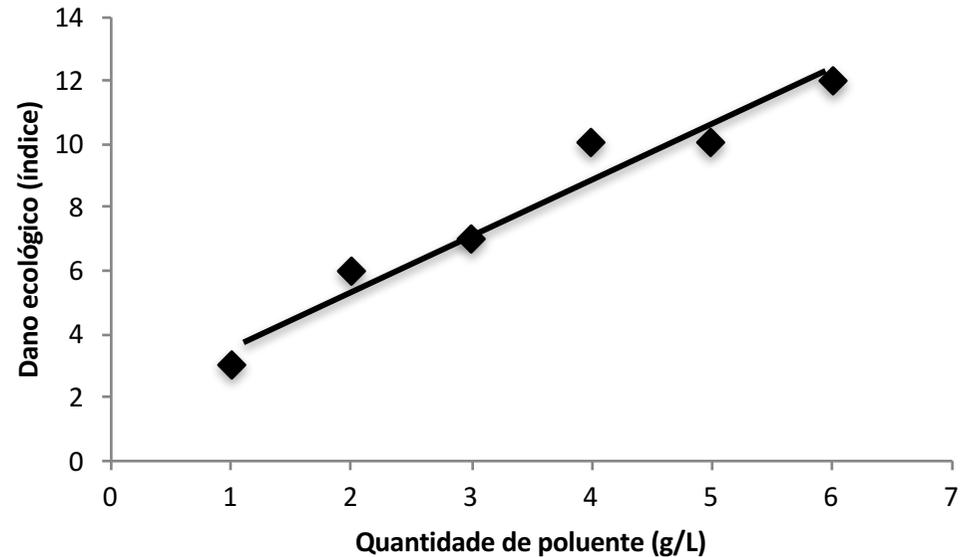


Dependência negativa

2. A reta de regressão linear

EXEMPLO

Relação entre certo poluente despejado por uma fábrica em um riacho e o dano ecológico na água, medido por um índice.



*Tal dependência poderia ser genericamente representada por uma **linha reta***

Análise de regressão linear simples

- procedimento que fornece equação de *linha reta* → **linear**
- *uma* variável preditora → **simples**

2. A reta de regressão linear

EQUAÇÃO DA RETA

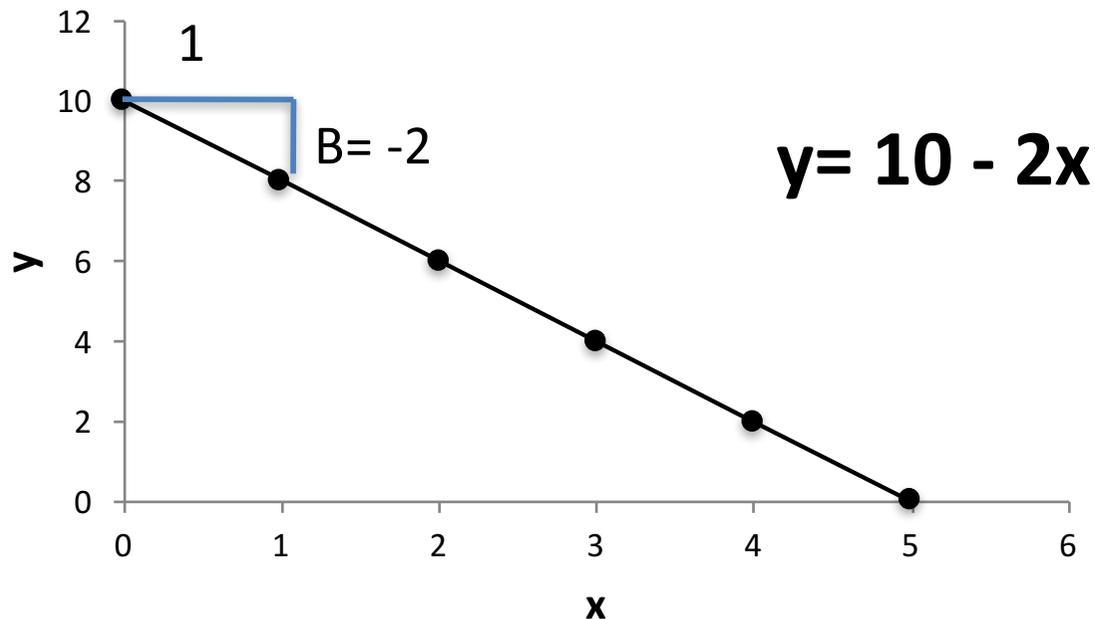
$$y = A + Bx$$

y = variável dependente

A = intercepto (valor de y qdo $x=0$)

B = coeficiente angular (inclinação da reta: acréscimo ou decréscimo em y para cada acréscimo de unidade em x)

x = variável independente



2. A reta de regressão linear

EQUAÇÃO DA RETA

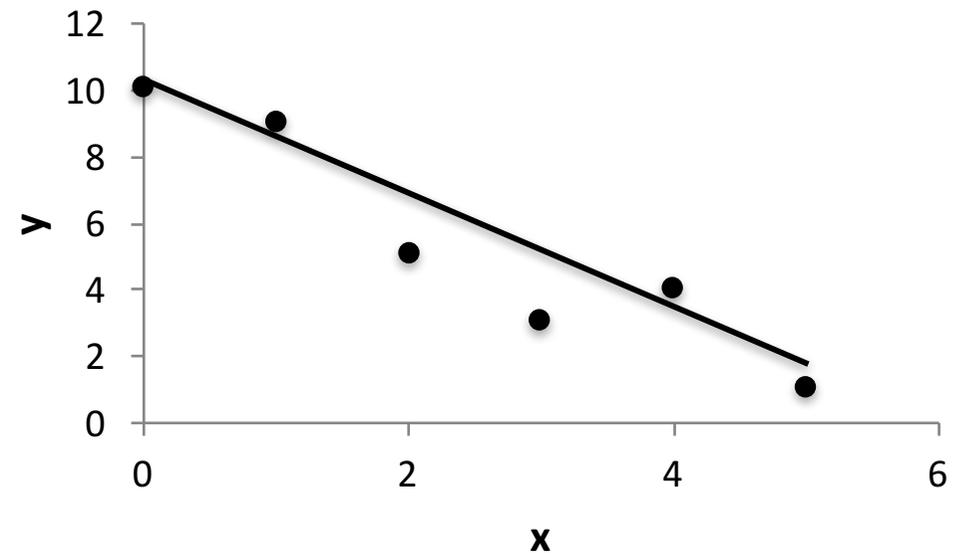
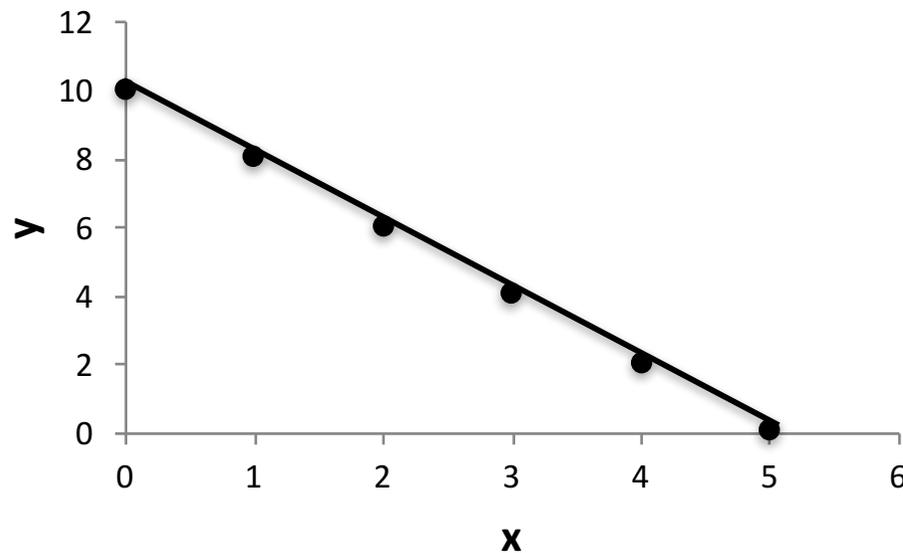
$$y = A + Bx$$

y é um valor que depende de x , mas uma vez que x assume um valor y é fixo

Dados
biológicos



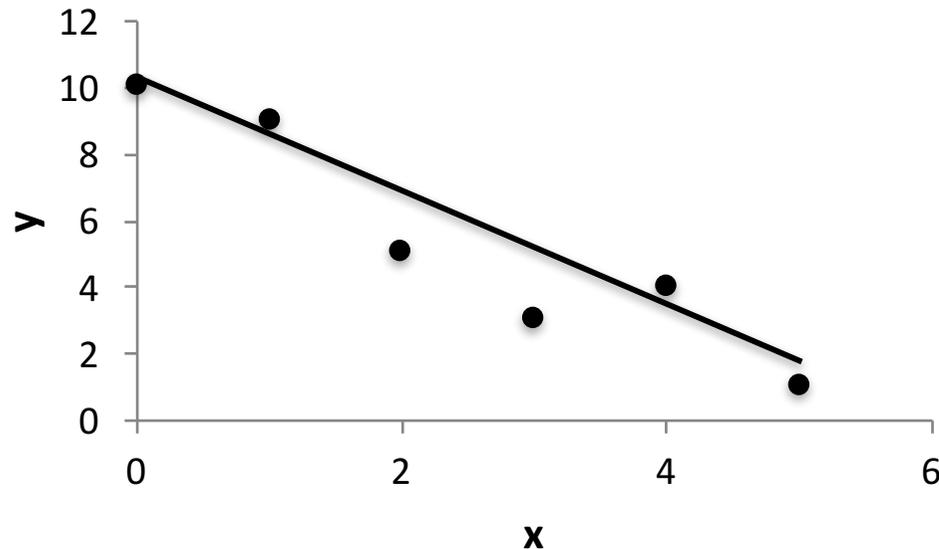
Variação



Desalinhamentos → interpretados como **desvios, ao acaso, do comportamento geral**

$$y = A + Bx + \epsilon$$

2. A reta de regressão linear

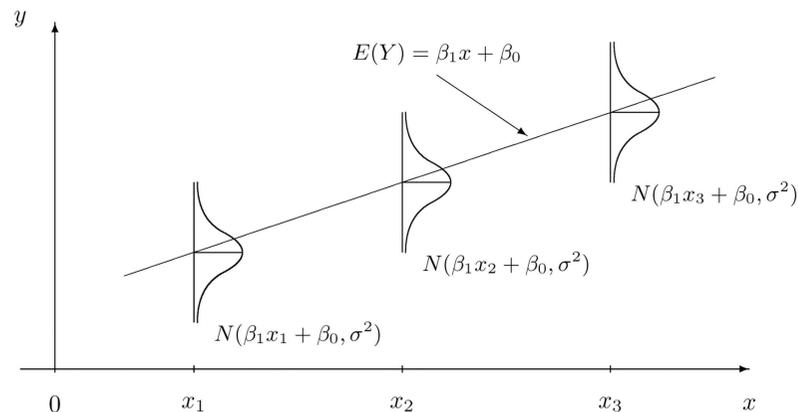


$$y = A + Bx + \varepsilon$$

Desalinhamentos →
interpretados como **desvios, ao acaso, do comportamento geral**

ε = erro ou resíduo

- A linha reta representa o comportamento de valores de y médios esperados para distintos valores de x

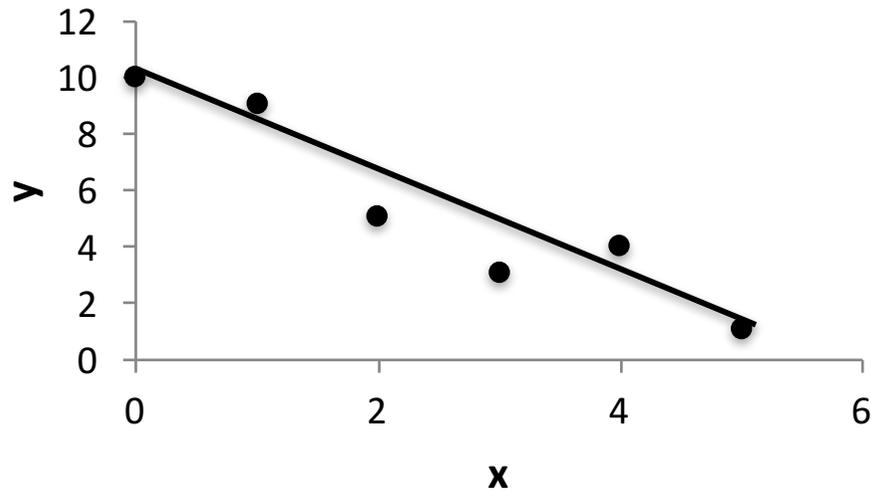


- Exemplo: para $x=2$ existe um conjunto de valores de y possíveis, sendo que a média destes valores está sobre a reta de regressão
- Pressuposto: a variação é sempre a mesma

2. A reta de regressão linear

OBTENÇÃO DA RETA DE REGRESSÃO

- A reta de regressão **verdadeira** seria obtida se fossem conhecidos os valores de x e y para **todos os indivíduos da população**



- No entanto, em geral temos apenas uma **amostra** da população



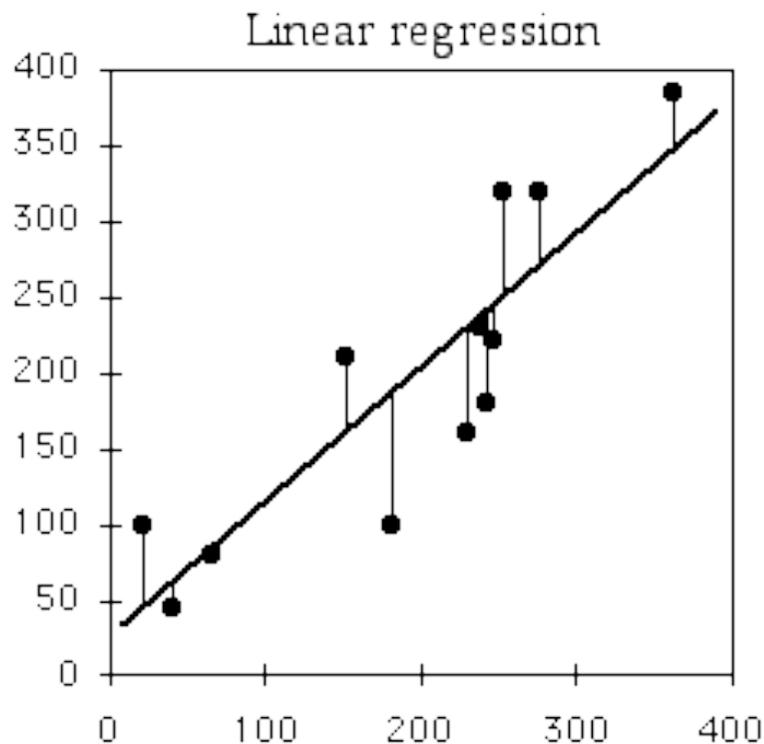
Estimativa dos parâmetros A e $B \rightarrow a$ e b

- *Método dos mínimos quadrados*: método usado para definir a reta e obter a e b

2. A reta de regressão linear

- *Método dos mínimos quadrados*: método usado para definir a reta e obter a e b

Garante que a reta obtida é aquela na qual se tem as menores distâncias (ao quadrado) entre os valores observados (y) e a própria reta (**soma dos quadrados dos resíduos – SQR**)



$$\hat{y} = a + bx$$

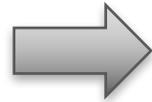
\hat{y} = valor esperado de y
para cada valor de x

$$a = \frac{\sum y - b \sum x}{n}$$

$$b = \frac{n \sum (xy) - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

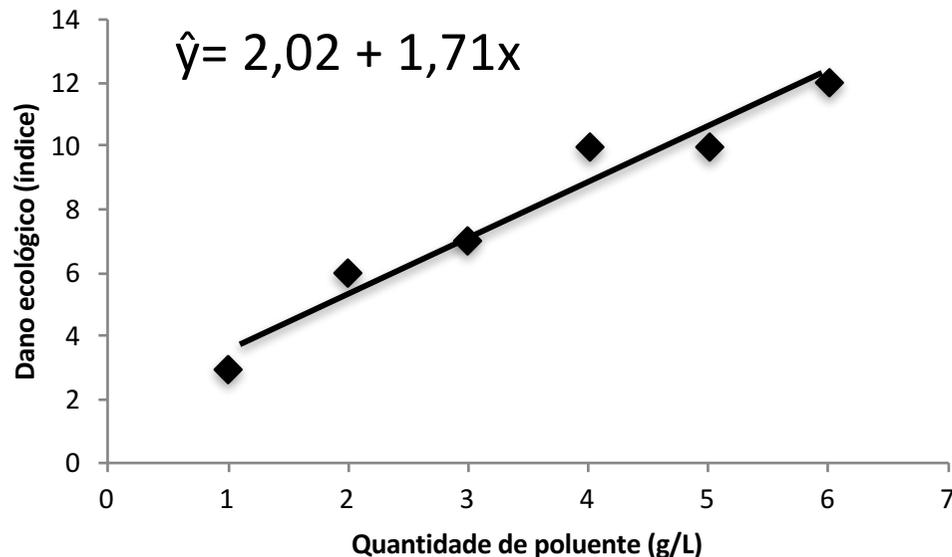
3. Teste de significância da regressão

Coeficiente angular (b)



Representa a dependência de y em relação a x

No entanto, trata-se de uma estimativa do B verdadeiro já que baseia-se em uma amostra



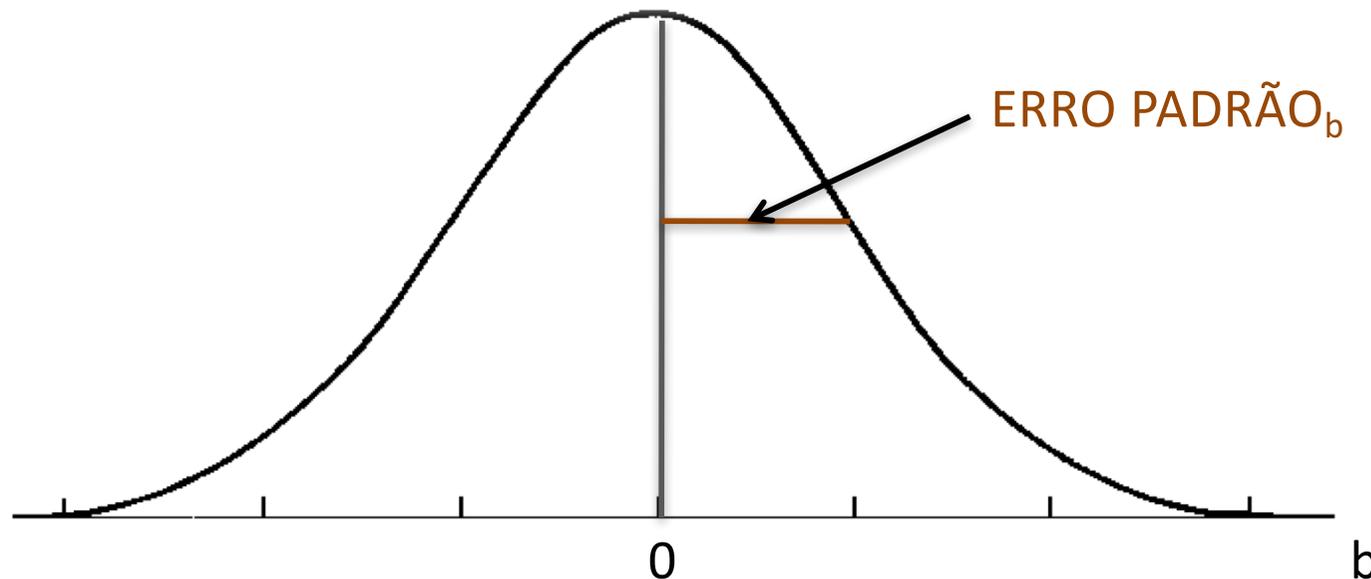
b = 1,71 representa uma dependência real de y em relação à x?

TESTE DE HIPÓTESE sobre a existência de dependência na população

3. Teste de significância da regressão

RACIOCÍNIO DO TESTE

- Testar a hipótese de que B é diferente de 0 $B=0 \rightarrow y$ não depende de x



- Para testar a hipótese de que B não é zero, determina-se o número crítico de erros padrão permitido para um afastamento não-significativo entre b e B , em unidades de erro padrão (t_{calc}).
- Se o valor calculado exceder o valor crítico, rejeita-se a hipótese de que b representa um desvio ao acaso de $B=0 \rightarrow y$ depende de x

3. Teste de significância da regressão

ETAPAS DO TESTE

Exemplo: poluente no riacho e dano ecológico

1) Hipóteses estatísticas

$$H_0: B = 0$$

$$H_1: B \neq 0$$

2) Nível de significância

$$\alpha = 0,05$$

3) Determinação do valor crítico do teste

$$gl = n - 2 \quad gl = 6 - 2 = 4$$

n = número de pontos

$$t_{\alpha; gl} = t_{0,01; 4} = 4,604$$

Atenção: bicaudal

3. Teste de significância da regressão

4) Determinação do valor calculado do teste

$$t_{calc} = \frac{b - B}{EP_b} = \frac{b}{EP_b}$$

B= 0 pois supõe-se que H_0 é verdadeira

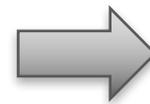
$$EP_b = \sqrt{\frac{\sum y^2 - a \sum y - b \sum xy}{(n-2) \left(\sum x^2 - \frac{(\sum x)^2}{n} \right)}}$$

$$t_{calc} = 1,71/0,187 = \mathbf{9,144}$$

5) Decisão

$$P \leq \mathbf{0,0001}$$

Como $|t_{calc}| = 9,144 > t_{0,01;4} = 4,604$



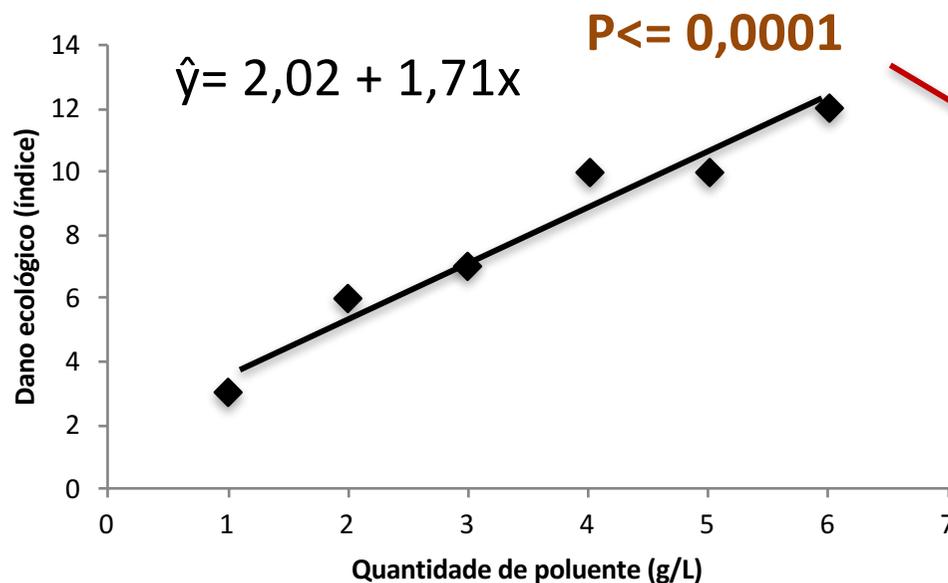
REJEITA-SE H_0

3. Teste de significância da regressão

6) Conclusão

Dado que o coeficiente angular populacional (B) não deve ser zero; logo admitimos que existe regressão de y sobre x ($\alpha= 0,05$)

O dano ecológico depende da concentração do poluente, de forma que para cada acréscimo de um g/L de poluente na água, espera-se que o índice de dano ecológico aumente **1,71 unidades**.



Tamanho de efeito
(relevância biológica)

Significância

- incompatibilidade dos dados com H_0
- clareza quanto ao efeito do poluente

4. Coeficiente de determinação – r^2

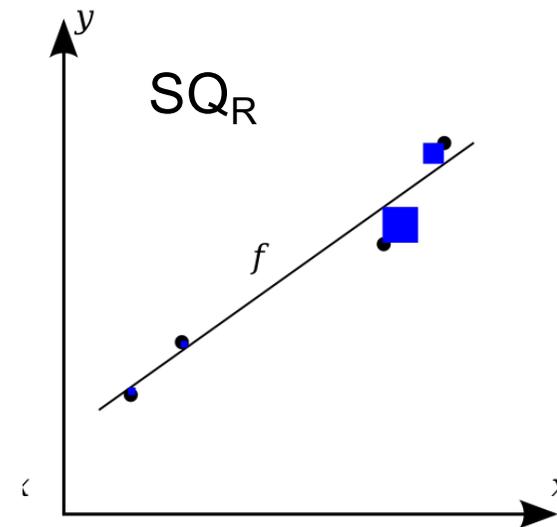
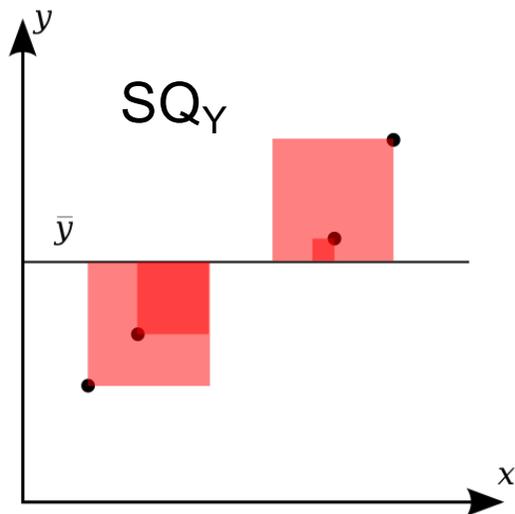
Componentes da variação

$$SQ_Y = SQ_{reg} + SQR$$

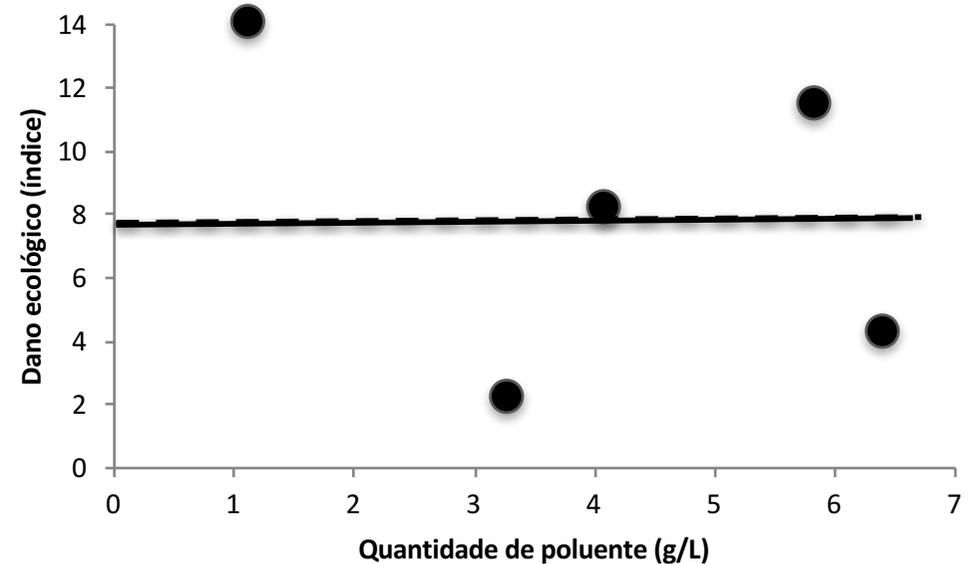
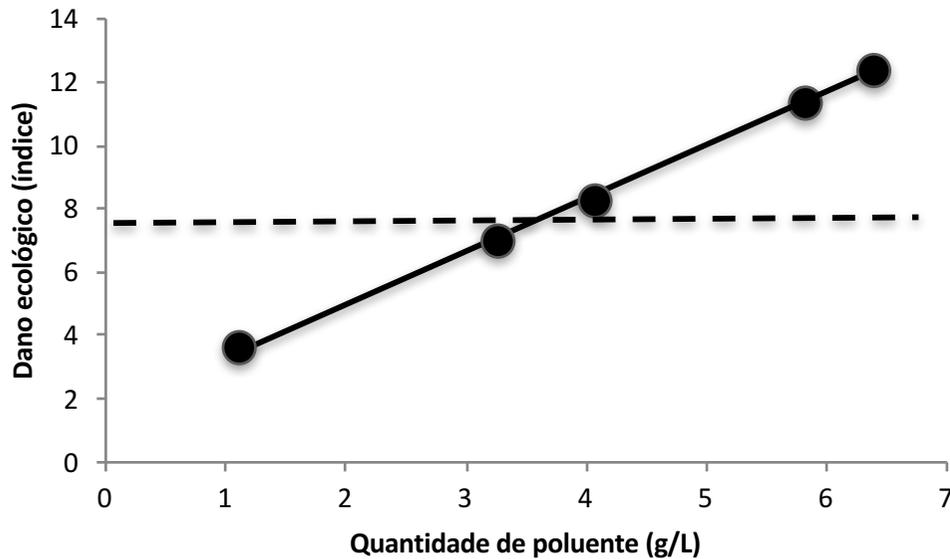
SQ_Y = soma dos quadrados da variável Y (variação total)

SQR = soma dos quadrados dos resíduos (erro aleatório)

SQ_{reg} = componente da variação atribuído ao modelo de regressão (sistemática)



4. Coeficiente de determinação – r^2



$$SQ_Y = SQ_{reg} + SQR$$

\searrow

$$SQ_Y = SQ_{reg} \quad 0$$

$$SQ_Y = SQ_{reg} + SQR$$

\searrow

$$SQ_Y = SQR \quad 0$$

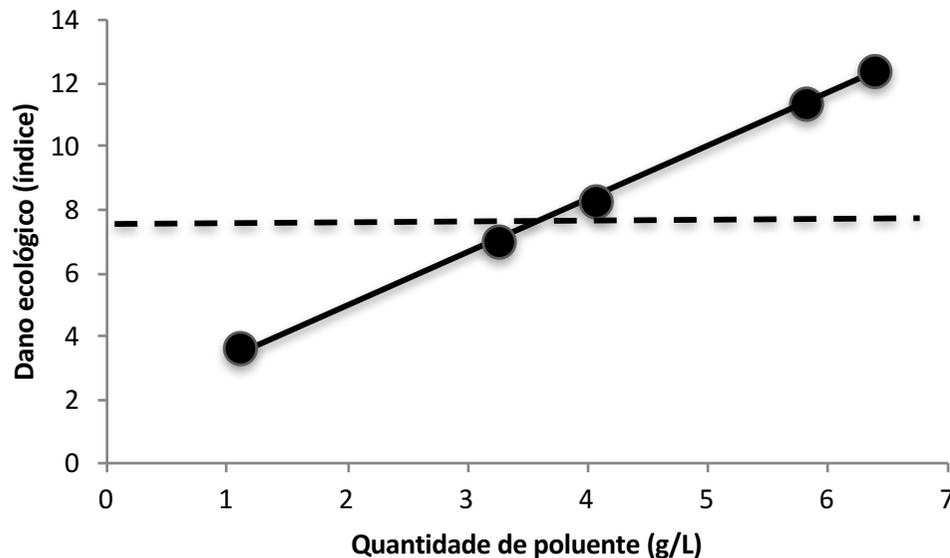
Entre estes dois extremos estão a maior parte dos dados ecológicos
(variação aleatória + variação sistemática)

4. Coeficiente de determinação – r^2

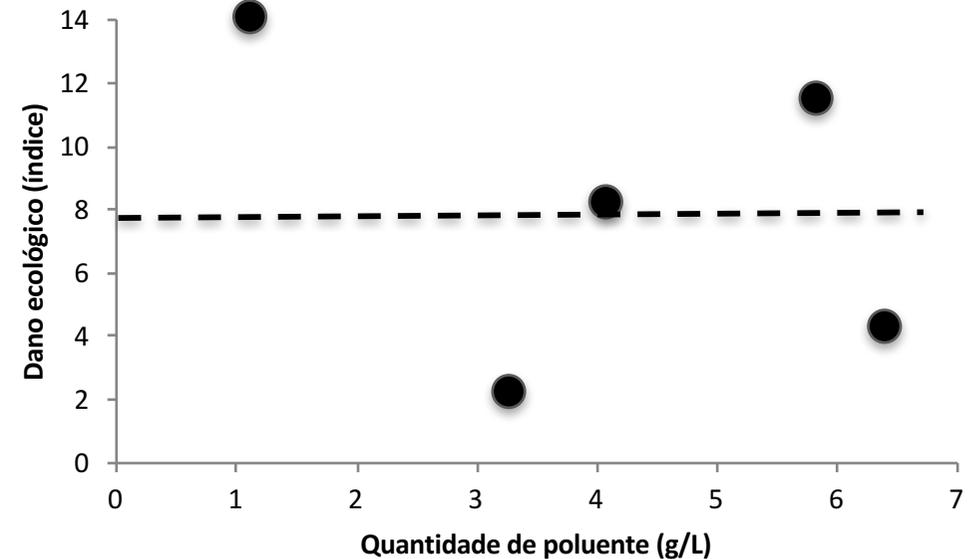
$$\text{Coeficiente de determinação } (r^2) = \frac{SQ_{\text{reg}}}{SQ_Y}$$

Descreve a proporção da variação em Y explicada pela regressão com X

Importância relativa da variação sistemática versus a aleatória



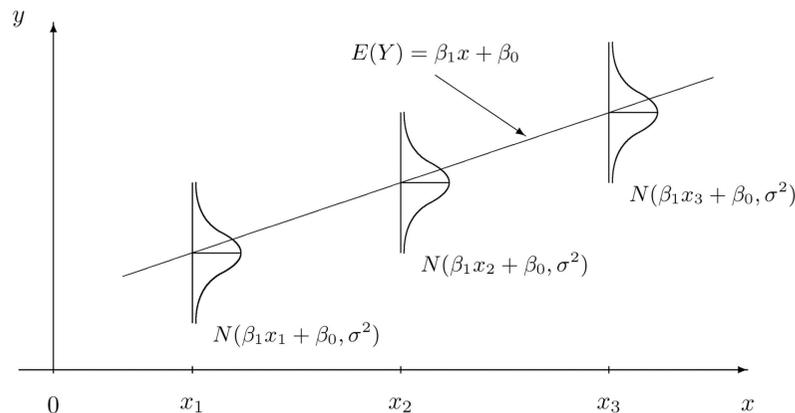
$$r^2 = 1$$



$$r^2 = 0$$

5. Pressupostos do teste

1. Pontos no gráfico devem apresentar tendência linear, caso contrário, a equação que melhor descreverá o fenômeno não será uma reta
2. A variável X é medida sem erros (alternativa: Regressão Modelo II) **PRESSUPOSTO FREQUENTEMENTE IGNORADO**
Risco: subestimar o B
3. Para cada valor de X , os valores de Y são independentes e com erros com distribuição normal -> resíduos
4. A variação é constante ao longo da linha de regressão (homogeneidade das variâncias)

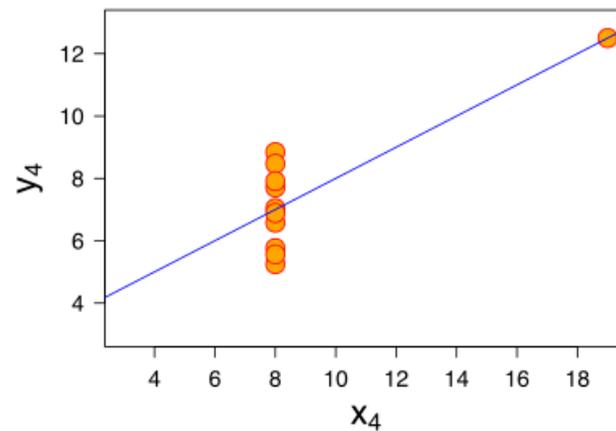
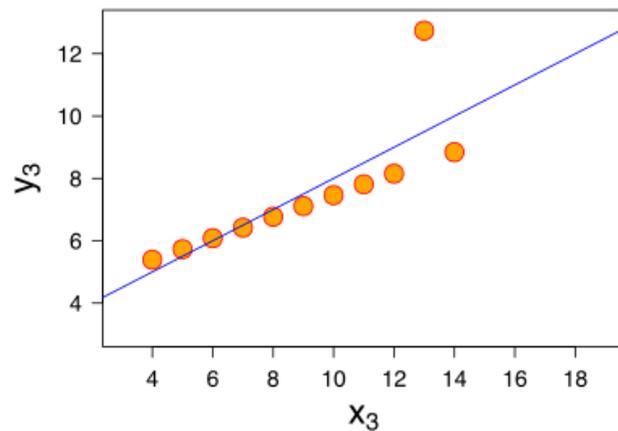
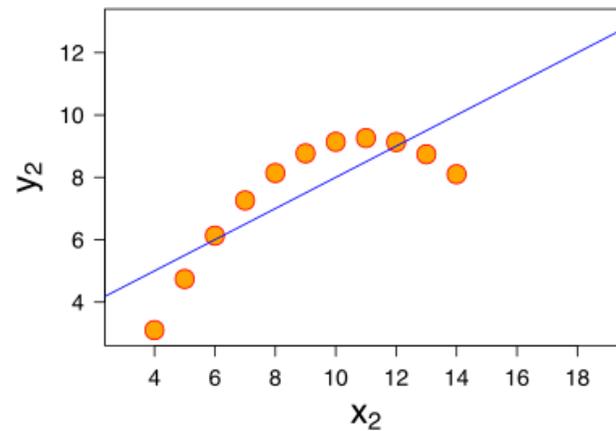
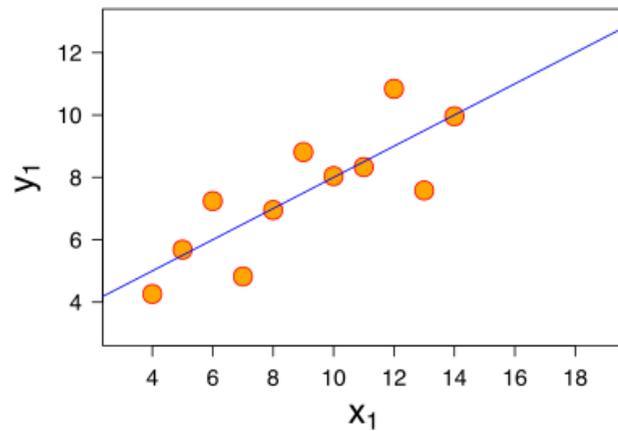


6. Procedimientos diagnósticos

Quarteto de Anscombe

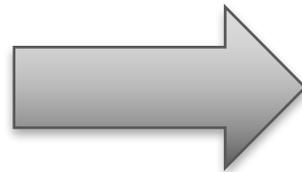
$$\hat{y} = 3,00 + 0,500x$$

$$R^2 = 0,67$$



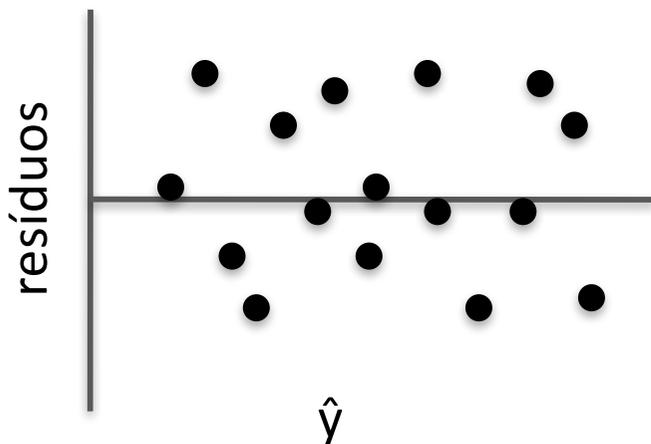
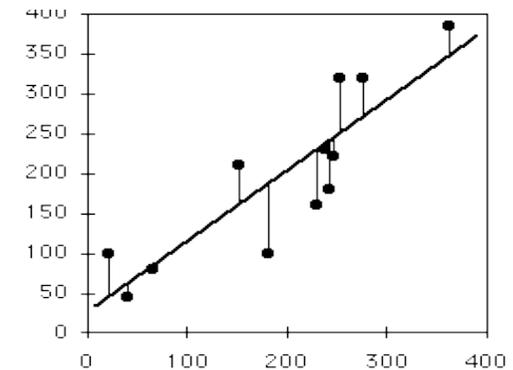
6. Análise dos resíduos

Gráfico diagnóstico para checar os pressupostos da regressão

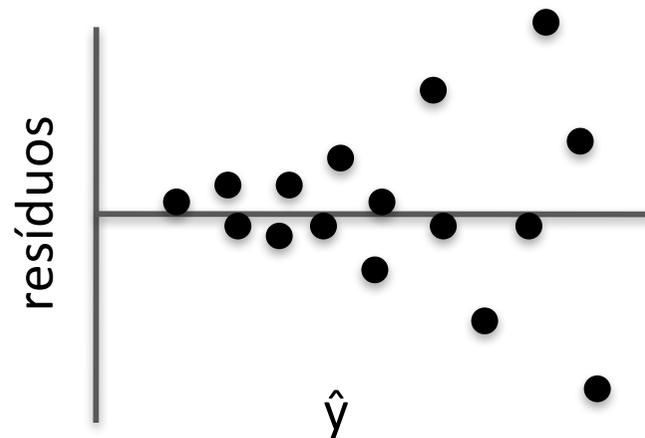


Resíduos no eixo vertical e os valores esperado de y (\hat{y}) no eixo horizontal

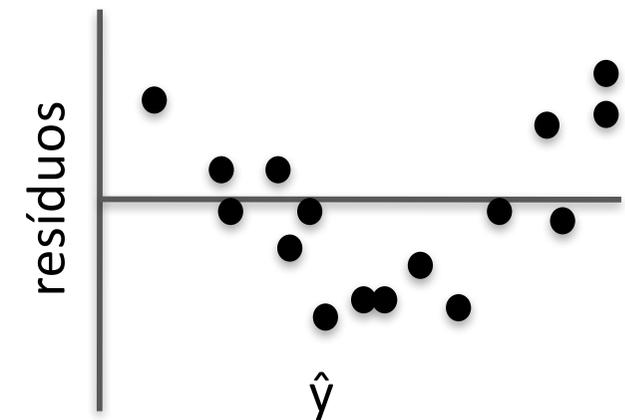
$$\text{Resíduo} = \varepsilon = y - \hat{y}$$



APROVADO!



Variâncias não homogêneas



Não linearidade

POSSÍVEL SOLUÇÃO:
transformação de dados

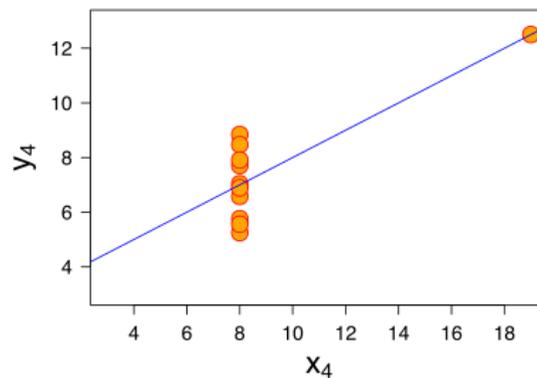
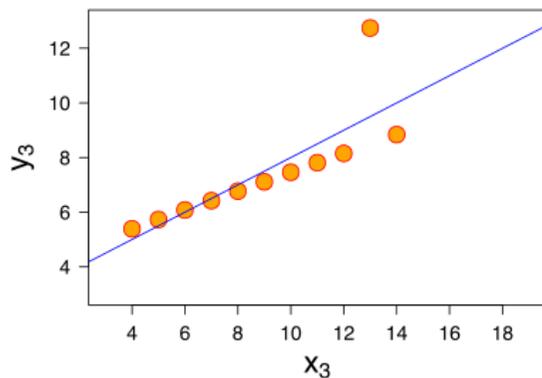
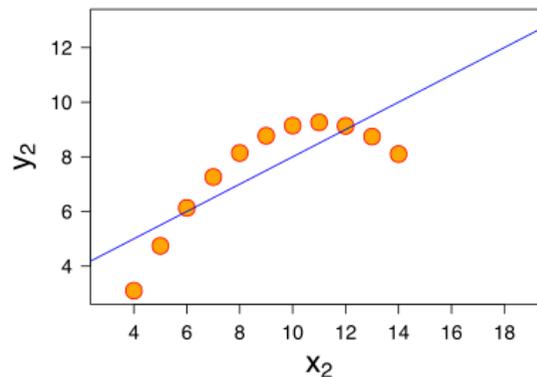
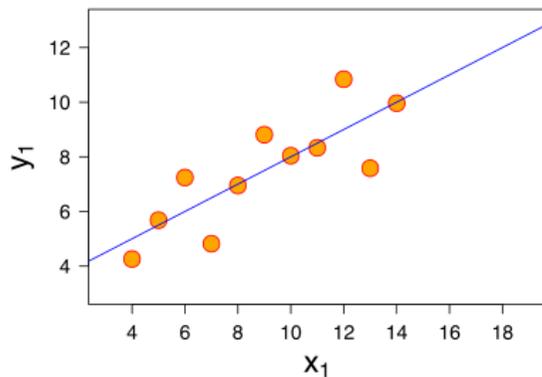
6. Outros diagnósticos

- sensibilidade ou função de influência



Forma de avaliar a estabilidade e validade geral das conclusões

Quarteto de Anscombe

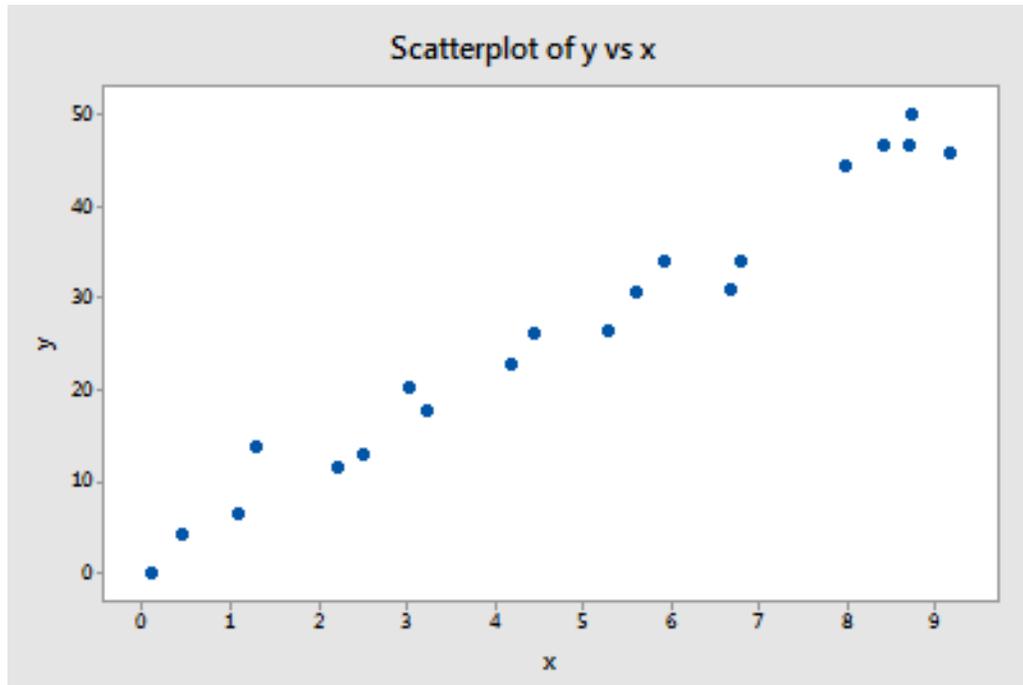


Distância de Cook

Mede a influência de cada dado considerando seu resíduo e sua “alavancagem”

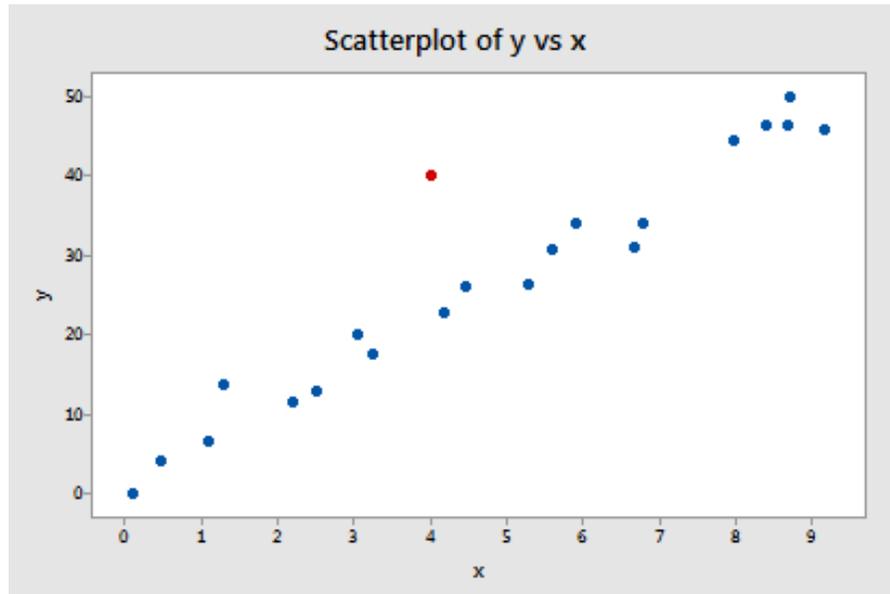
Alta alavancagem = valor extremo de X

6. Outros diagnósticos



- ✓ Sem outliers
- ✓ Sem observações com alta alavancagem

6. Outros diagnósticos



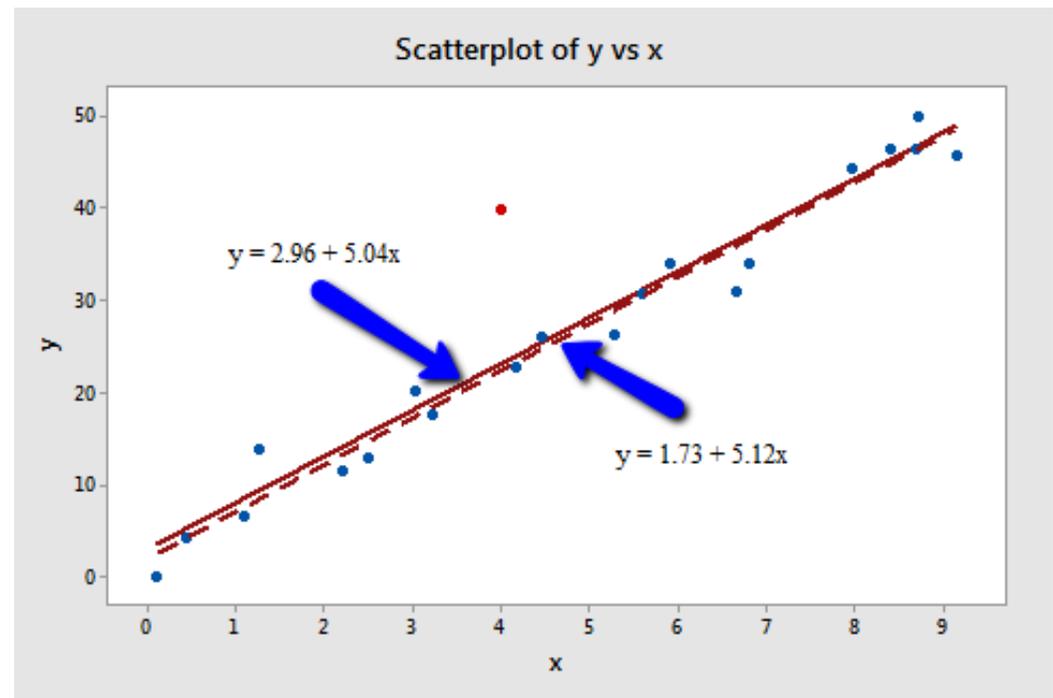
Resultados com pouca alteração nos coeficientes e r^2

Valor de P para $H_0: B=0$ é $<0,001$ no dois casos

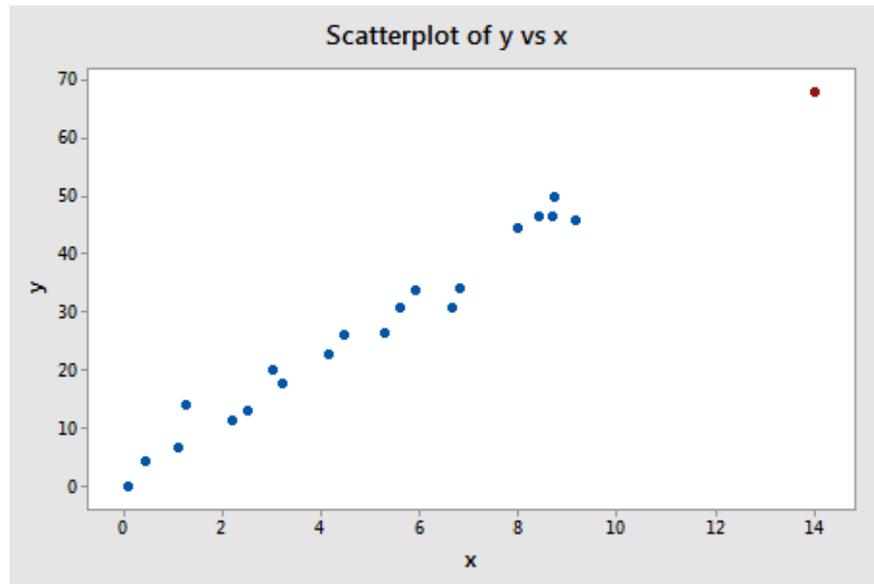
Não há dados influentes!

✓ Um outlier

✓ Sem observações com alta alavancagem



6. Outros diagnósticos

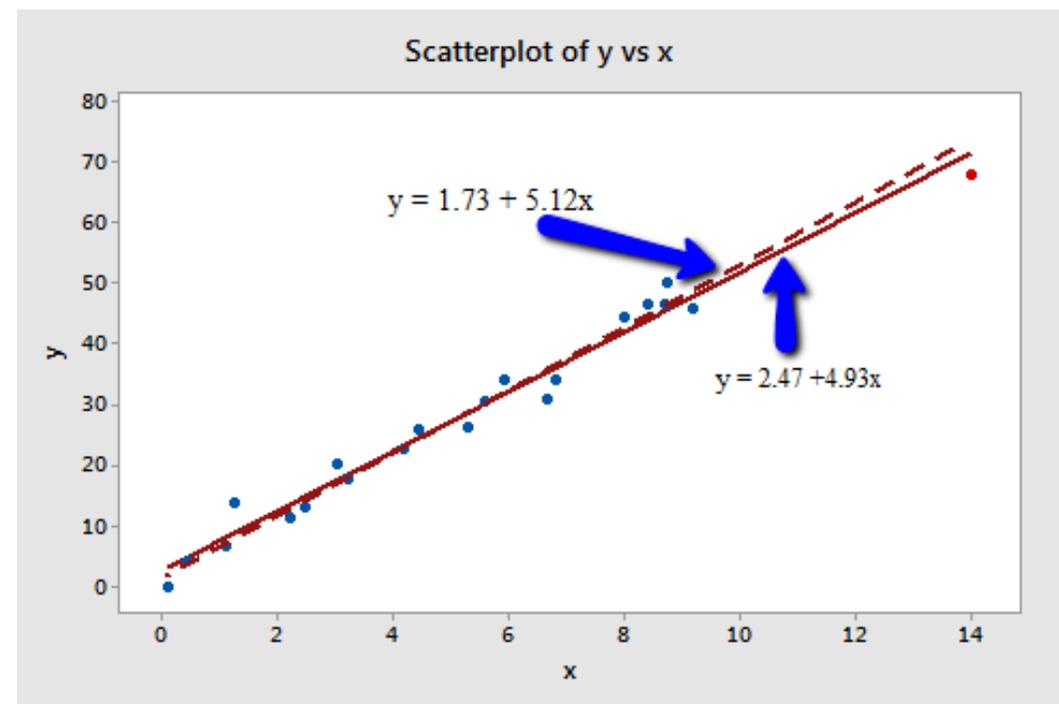


Resultados com pouca alteração

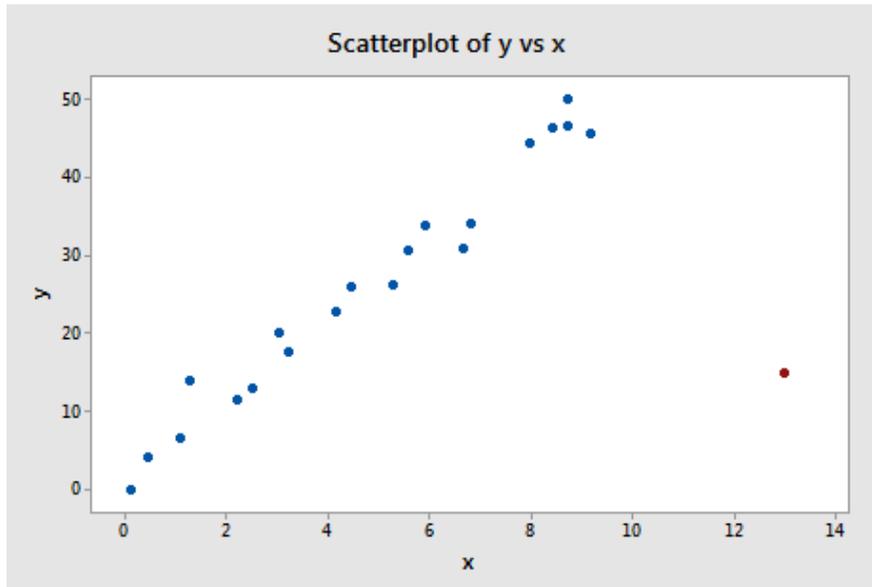
Não há dados influentes!

✓ Não há outliers

✓ Um dado com alta alavancagem



6. Outros diagnósticos



✓ **Ponto vermelho é um outlier e apresenta alta alavancagem**

Alteração substancial das estimativas dos coeficientes do modelo (a e b) e r^2 (97% \rightarrow 55%)

Valor de P para $H_0: B=0$ é $<0,001$ no dois casos

Dado influente!

