Introdução aos Modelos Lineares em Ecologia

Prof. Adriano Sanches Melo - Dep. Ecologia – UFG
asm.adrimelo no gmail.com

Página do curso: www.ecologia.ufrgs.br/~adrimelo/lm/

Livro-texto: Crawley, M.J. 2005. Statistics: An Introduction using R. John Wiley & Sons.

Página do livro na internet:
http://www3.imperial.ac.uk/naturalsciences/research/statisticsusingr

AULA 4

1. Resíduos

Propriedades dos Resíduos:

Média = 0

$$s^2 = \frac{\sum \left(e_i - \bar{e}\right)^2}{n-2} = \frac{\sum e_i^2}{n-2} = \frac{SSE}{n-2} = MSE$$

2. Problemas que podem ser avaliados por análise de resíduos

1. A regressão não é linear

2. Os erros não possuem constância de variância

3. Os erros não são independentes

4. O modelo se ajusta bem à maioria das observações, mas existem *outlyers*

5. Os erros não são distribuídos de acordo com a distribuição normal

6. Uma ou mais variáveis importantes não foram incluídas no modelo

3. Diagnóstico dos resíduos

1. Diagrama de dispersão dos resíduos contra a variável independente

2. Diagrama de dispersão dos valores absolutos ou de seu quadrado contra a variável independente

3. Diagrama de dispersão dos resíduos contra os valores ajustados

4. Diagrama de dispersão dos resíduos contra o tempo ou ordem de coleta etc. (existe autocorrelação?)

5. Diagrama de dispersão dos resíduos contra uma variável preditora potencial mas que não foi incluída no modelo

6. Box-plot dos resíduos

7. "*Normal probability plot*" dos resíduos (ver exemplo abaixo)

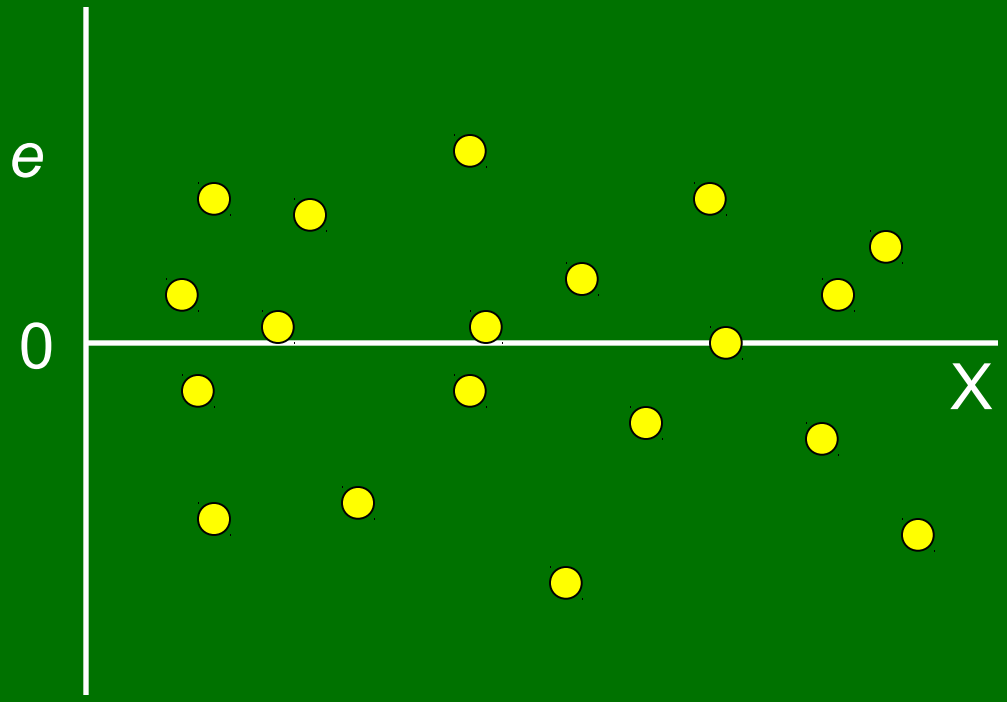
Como os resíduos DEVEM ser

-- Sem padrão aparente

-- Constância de variância (homogeneidade de variâncias)

-- Maioria dos valores próximos a zero.

-- Ausência de *outlyers*



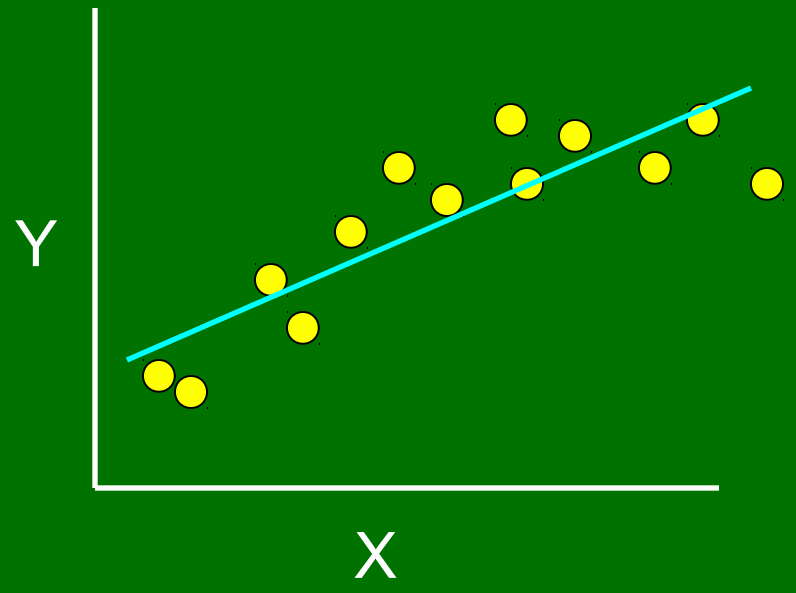
Como os resíduos NÃO DEVEM ser:

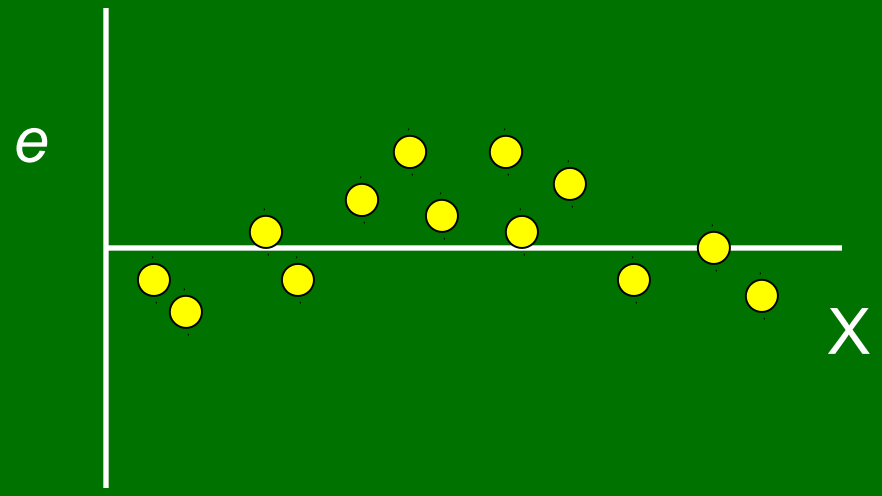
a) Não-linearidade da função de regressão

Pode ser avaliado a partir de:

1. Diagrama dispersão dos resíduos contra a variável independente
2. Diagrama de dispersão dos resíduos contra os valores ajustados

Exemplo:

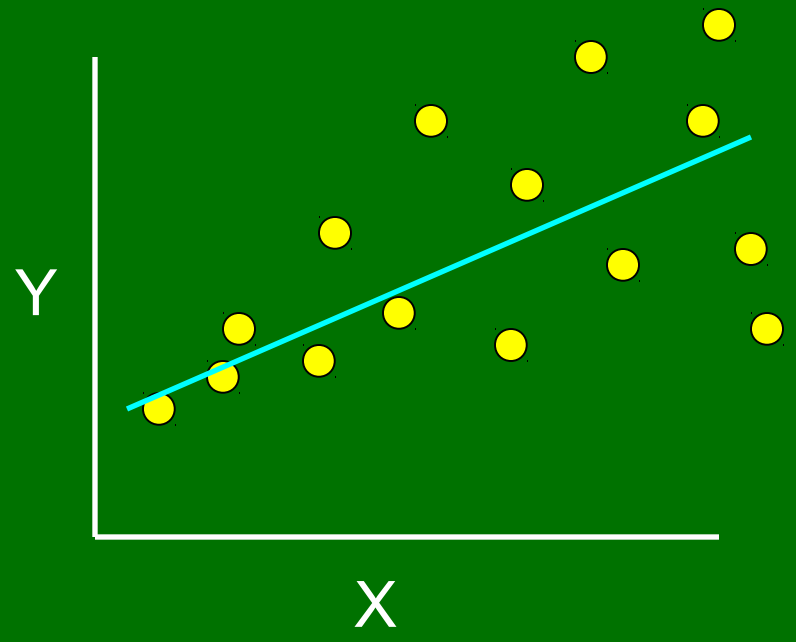


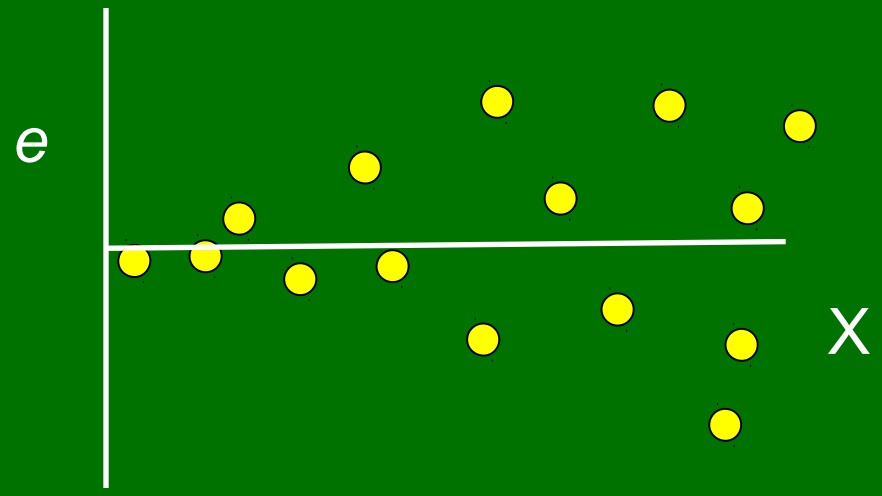


Como os resíduos NÃO DEVEM ser:

b) Não-homogeneidade de variâncias (MUITO IMPORTANTE!!)

Um gráfico de resíduos contra X pode revelar padrão de megafone (aumento de variância conforme aumenta X). Quando existem poucas observações, pode-se usar o valor absoluto do resíduo ou o seu quadrado.

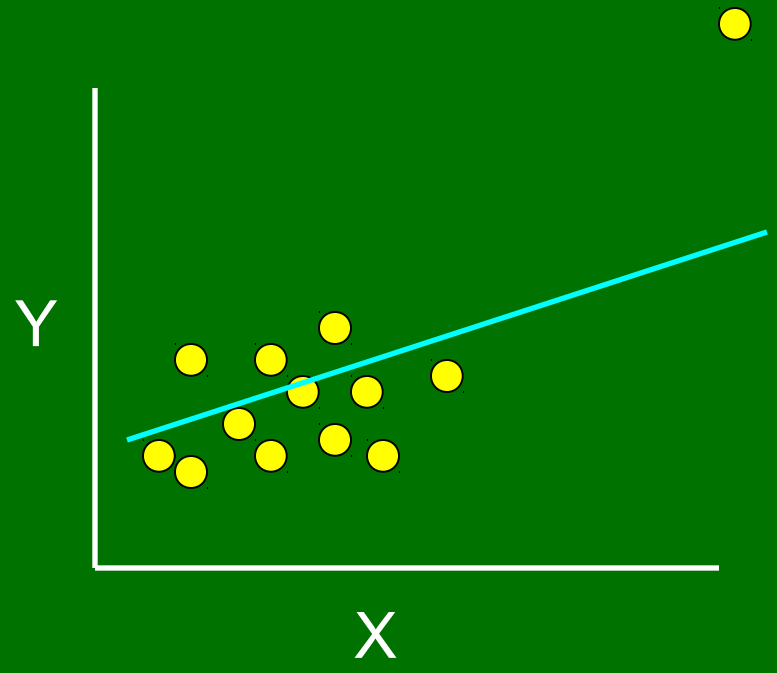


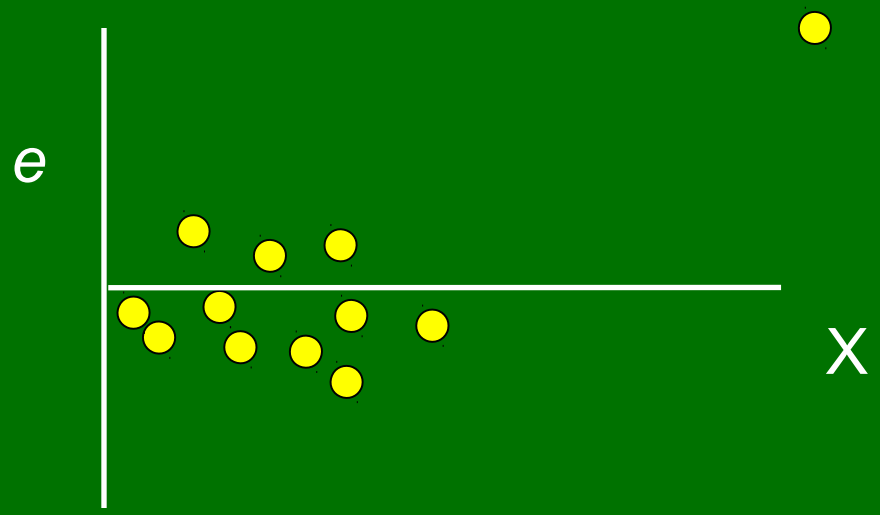


Como os resíduos NÃO DEVEM ser:

c) Presença de *outlyers*

Quando se têm poucas observações, uma simples observação pode alterar muito a relação.

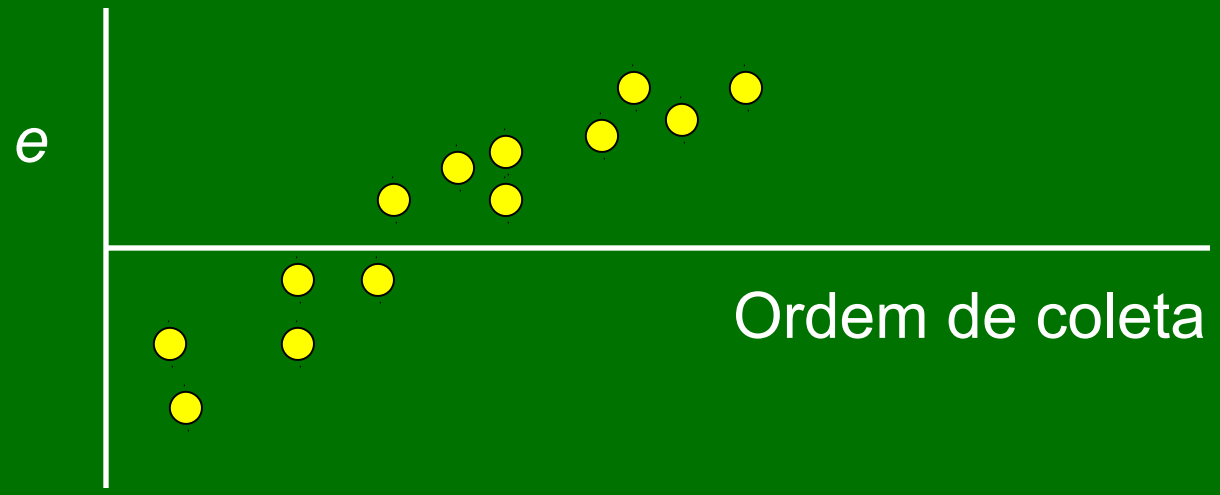




Como os resíduos NÃO DEVEM ser:

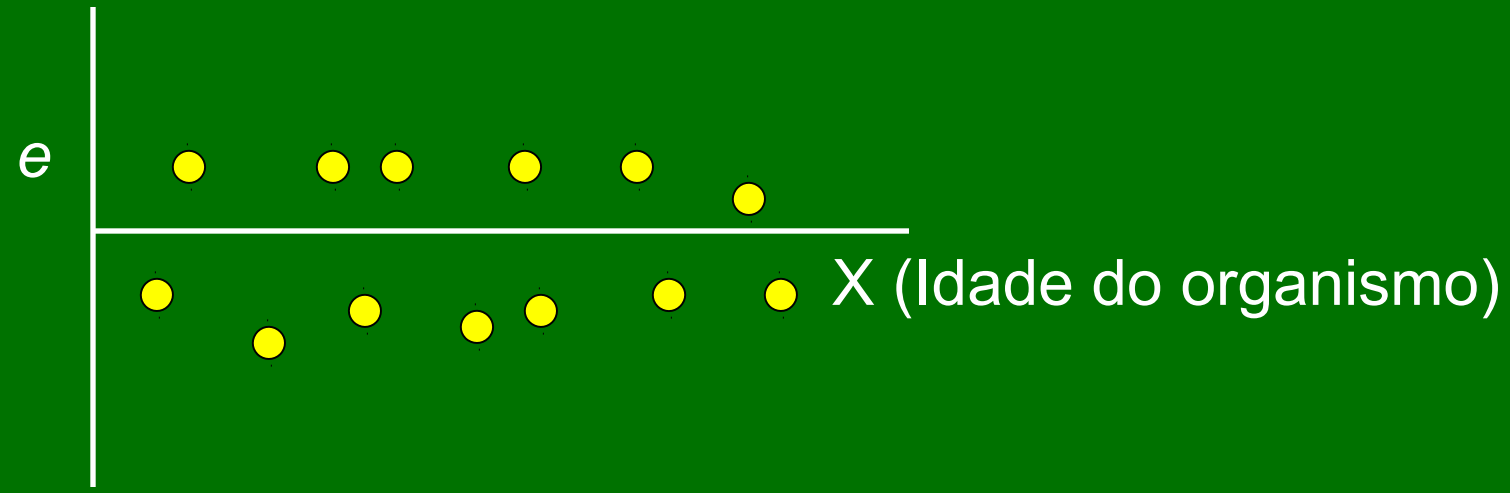
d) Não-independência dos erros

Sempre que dados forem coletados numa sequência temporal ou espacial, deve-se fazer um diagrama dos resíduos contra a referida sequência.

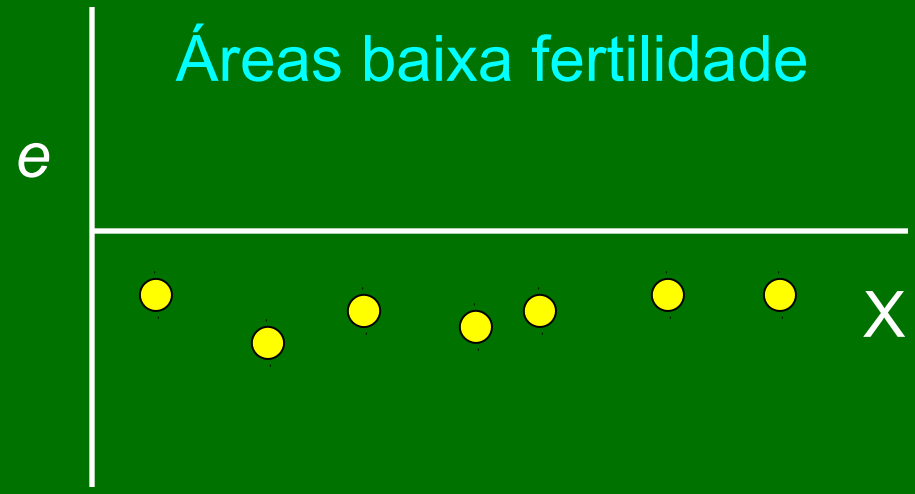


Como os resíduos NÃO DEVEM ser:

e) Omissão de variável importante.



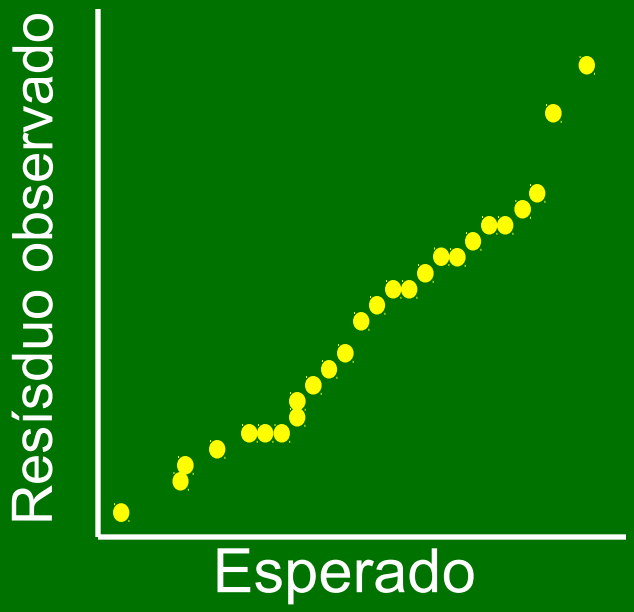
Áreas baixa fertilidade



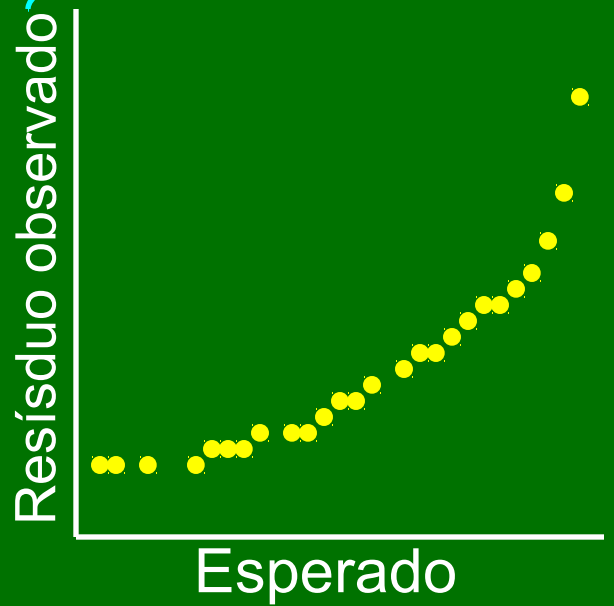
Áreas alta fertilidade

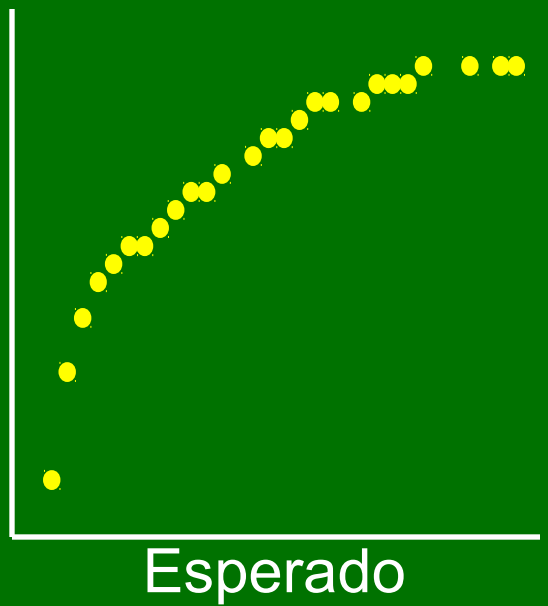
Normal Probability Plot

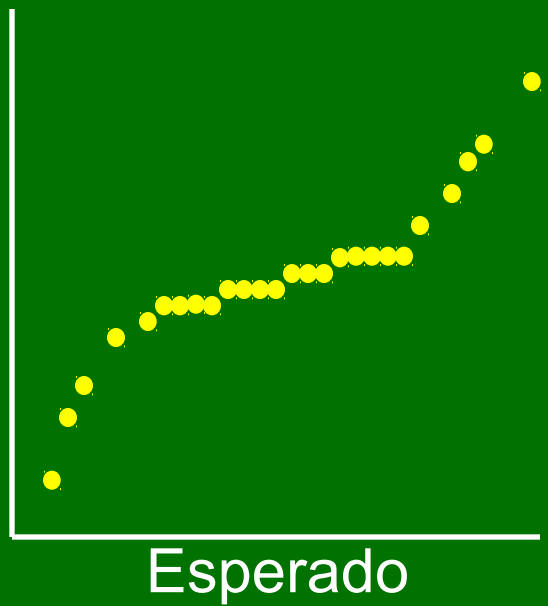
Como os resíduos DEVEM ser



Como os resíduos NÃO DEVEM ser

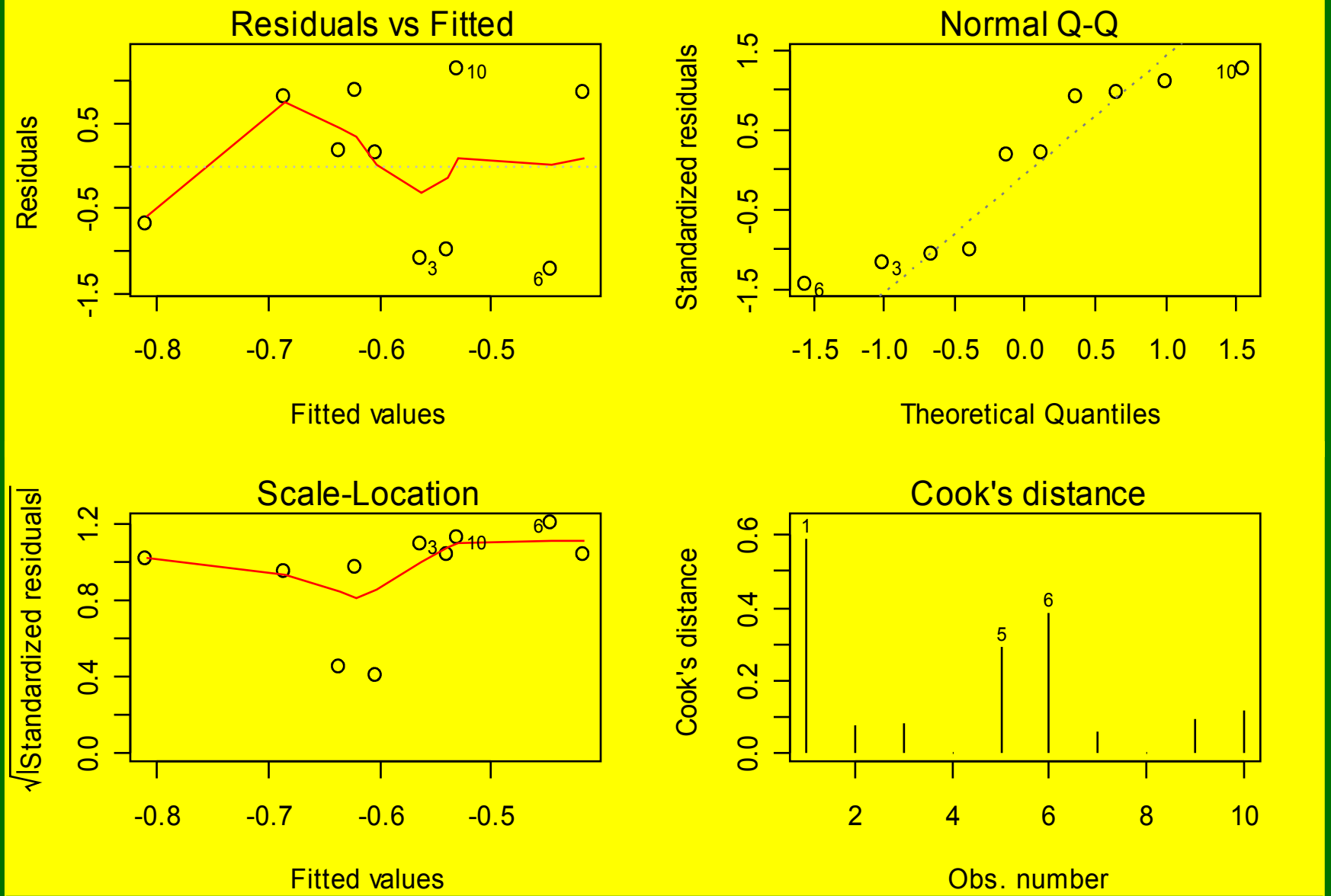






No R:

```
resu<-lm(y~x)
par(mfrow=c(2,2))
plot(resu, which=c(1,2,3,4))
```



Distância de Cooks

Influência da observação *i* sobre todos os valores ajustados.

Em outras palavras, os valores ajustados mudariam muito se a observação *i* fosse excluída?

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)} \right)^2}{p * MSE}$$

$\hat{Y}_j$ = Valores ajustados

$\hat{Y}_{j(i)}$ = Valores ajustados sem a observação *i*

p = número de parâmetros

Quando a distância de Cook é grande?

Segundo Kutner et al. (2004):

Obtenha percentil da distribuição F com (p, n-p) graus de liberdade

No R: `pf(cook, df1, df2)`

onde cook = distância

df1 = número de parâmetros estimados

df2 = nº observações – nº parâmetros estimados

Se valor obtido for:

< 0.2 = baixa influência

> 0.2 e < 0.5 = influência moderada

> 0.5 = grande influência

4. Visão geral de medidas remediadoras

Transformações mais comuns

- Log (y) ou Log (y + 0,5)
- Raíz quadrada de y
- Inverso → 1 / y

5. Transformação Box-Cox $Y' = Y^{\lambda}$

Objetivo é achar λ mais adequado

$\lambda = 2 \longrightarrow Y' = Y^2$

$\lambda = 0.5 \longrightarrow Y' = \sqrt{Y}$

$\lambda = 0 \longrightarrow Y' = \log_e Y$ (por definição)

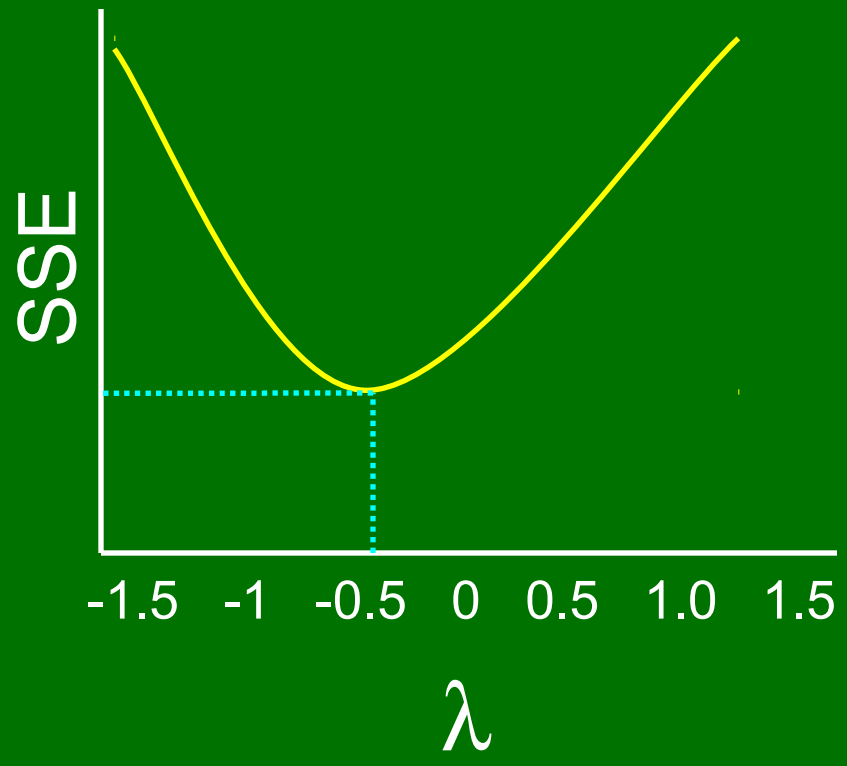
$\lambda = -0.5 \longrightarrow Y' = 1/\sqrt{Y}$

$\lambda = -1 \longrightarrow Y' = 1/Y$

Como descobrir o melhor λ: Para cada valor de λ, as observações Y^{λ}_i são inicialmente padronizadas de forma que a magnitude da Soma de Quadrados do Erro (SSE) não dependa do valor λ. Repete-se o procedimento com vários valores de λ até achar aquele valor que minimiza SSE. Usa-se então este valor de λ para transformar a variável de estudo.

Neste exemplo, o menor valor de SSE é encontrado com $\lambda = -0.5$.

Usa-se portanto a variável transformada ($Y' = Y_i^{-05}$) na análise.



Para cada valor de λ, as observações Y^{λ}_i são inicialmente padronizadas de forma que a magnitude da Soma de Quadrados do Erro (SSE) não dependa do valor λ:

$$W_i \begin{cases} K_1(Y^{\lambda}_i - 1) & \text{para } \lambda \neq 0 \\ K_2(\log_e Y_i) & \text{para } \lambda = 0 \end{cases}$$

onde:

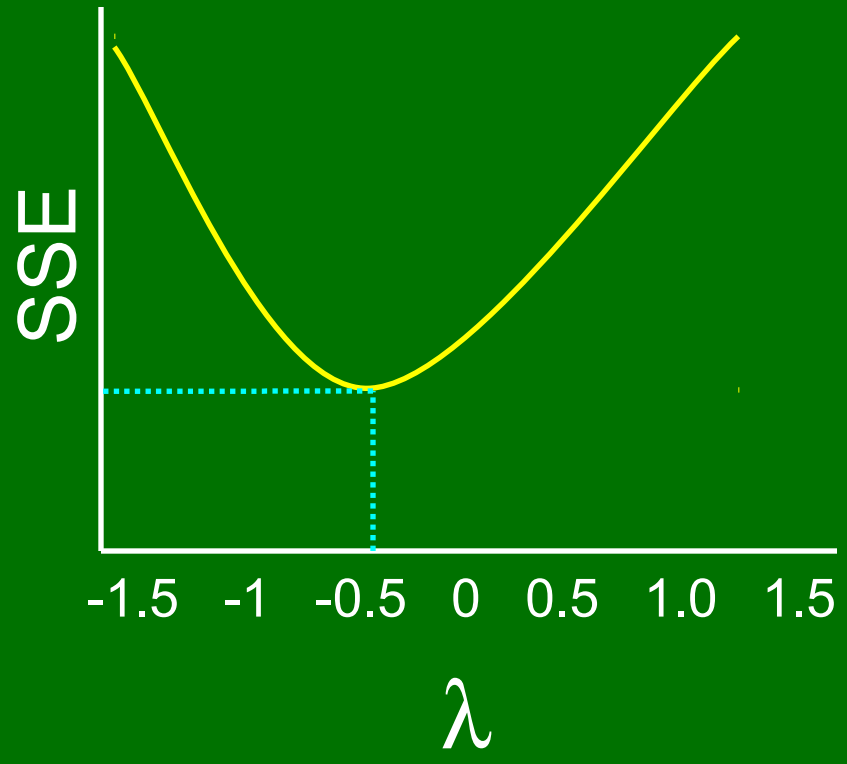
$$K_2 = \left(\prod_{i=1}^{n} Y_i \right)^{1/n}$$

(média geométrica observações Y_i)

$$K_1 = \frac{1}{\lambda K_2^{\lambda - 1}}$$

Após obtenção de W_i para cada observação, calcula-se o modelo de regressão e anota-se a SSE.

Repete-se o procedimento com vários λ e usa-se aquele que minimize SSE.



Box-Cox no R:

```
K2<-prod(Yplasma)^(1/25) #para variável Yplasma com 25 observações
K1<-1/(0.3*(K2^-0.7))        # para variável Y com lambda = 0.3
W<-K1*((Yplasma^0.5)-1)      # para obter variável padronizada
summary(lm(W~X)) # para examinar o SSE. Agora repete-se o procedimento
```

com vários λ para descobrir com qual deles minimiza-se SSE.

O valor que minimizar será usado para transformar Yplasma e então fazer a análise dos dados.

Para descobrir λ automaticamente; dentro do pacote MASS

```
library(MASS)
boxcox(Y~X)
locator()     ##vá até figura e clique com esquerdo no pico. Depois clique com direito e escolha 'parar'.
print(boxcox(Y~X))
```

Exercícios e estudo individual:

-- Lista em sala de aula

-- Crawley: Cap. 8 (pp. 143-145)

-- Gotelli & Ellison: Cap. 9 (pp. 259-264)