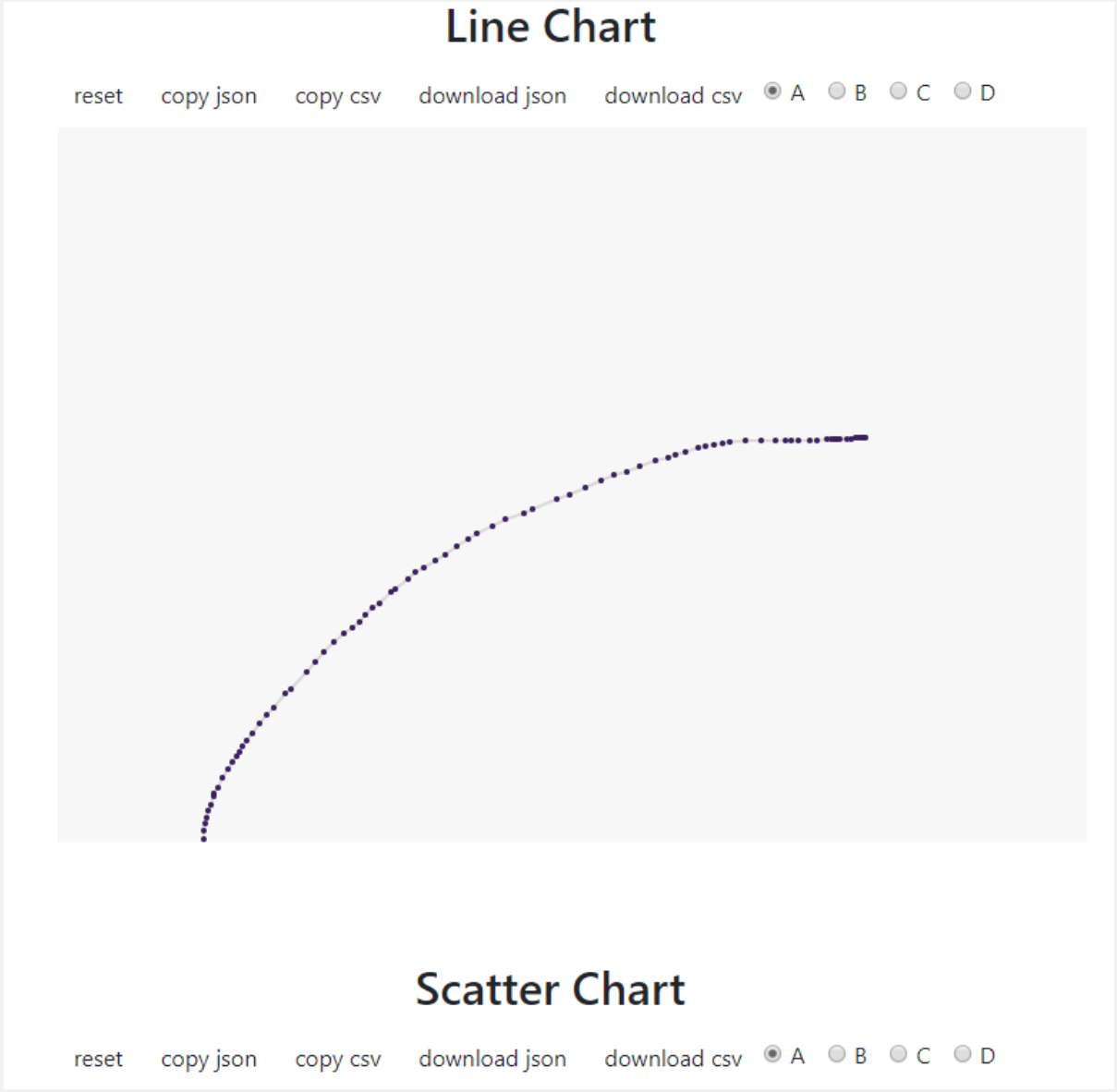
Modelação Ecológica

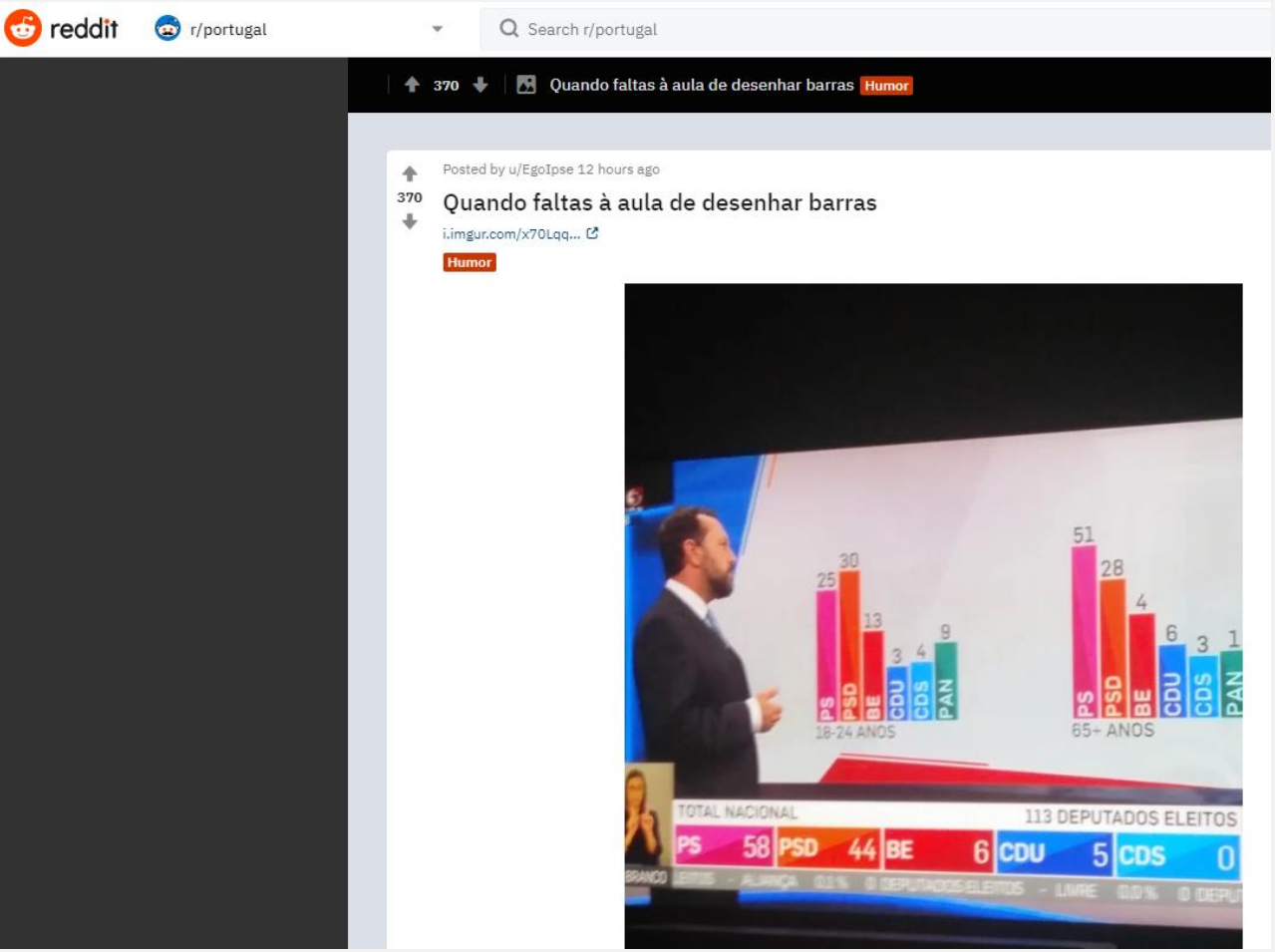
CLASS 7

08 October 2019 – 13:30-16:00 – room 2.4.37

Tiago A. Marques



http://drawdata.xyz/



SIC Noticias:
Provavelmente, o pior conjunto de gráficos da noite.

Se alguém algum dia me entregar uns gráficos destes pode desitir da cadeira de Ecologia Numérica – dedique-se antes à pesca.

Sem escala, sem eixos, sem explicação do que é o quê, e com informação claramente contraditória… alguêm consegue sequer adivinhar o que isto era?

Um bom exemplo de péssimo jornalismo, absolutamente deplorável.

Tweet

Richard Feynman
@ProfFeynman

Teach your students
- to doubt,
- to think,
- to communicate,
- to question,
- to make mistakes,
- to learn from their mistakes, and most importantly
- have fun in their learning.

3:34 AM · Oct 6, 2019 · Twitter for Android

3.8K Retweets **10.6K** Likes

A FCT e a Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) informam que se encontra aberto concurso **até dia 13 de novembro de 2019**, para o apoio a 10 projetos conjuntos a decorrerem no biénio 2020/2021.

O apoio destina-se ao intercâmbio de investigadores no âmbito de projetos comuns de investigação nas seguintes áreas

- Alterações Climáticas
- Ciências do Espaço
- Ciências do Mar
- Computação Avançada
- Inteligência Artificial
- Medicina Oncológica

O financiamento por parte da FCT destina-se, estritamente, à mobilidade de investigadores participantes nos projetos: despesas de viagem e estadia da equipa portuguesa no Brasil (o montante previsto para cada projeto será no valor de 4500€/ano).

A **não inclusão de jovens investigadores na equipa portuguesa** tem **caráter eliminatório** (ver definições nas FAQ's).

Será dada **prioridade** a **novos projetos** e a **equipas que não obtiveram financiamento nos últimos concursos.**

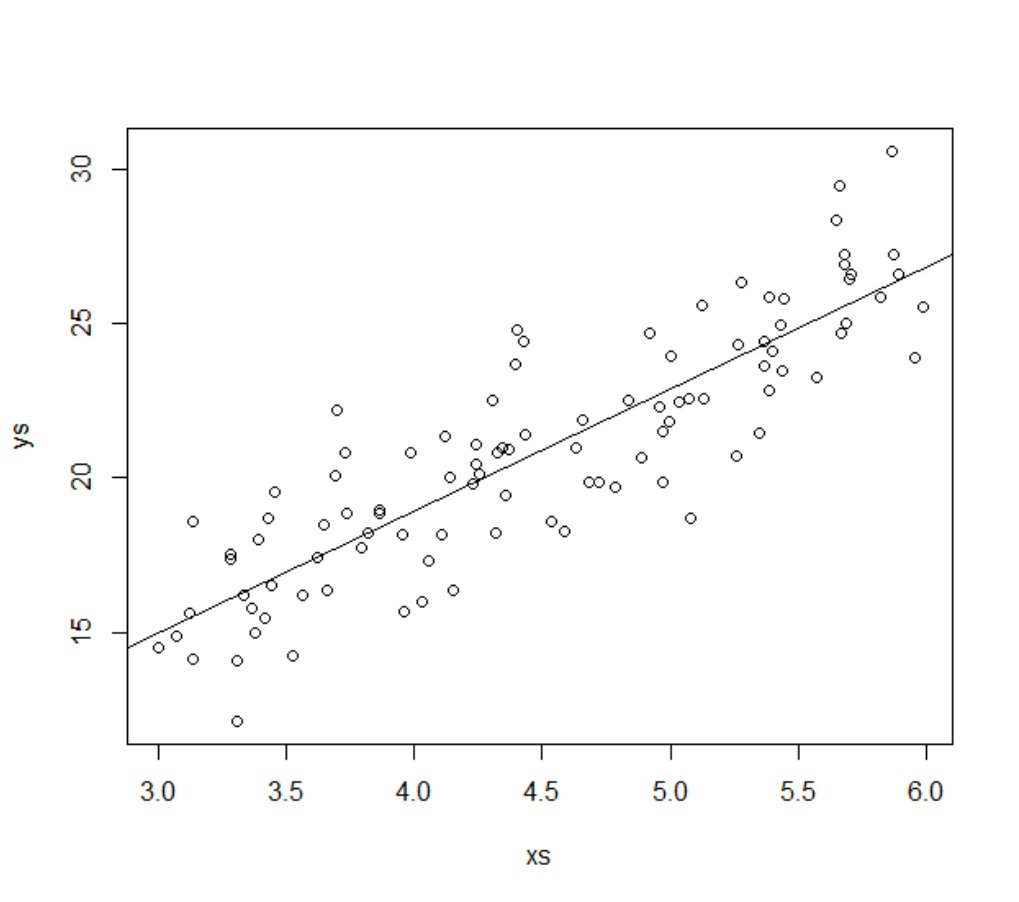
Os organismos executores deste Acordo procederão à respetiva avaliação e seleção. Não serão consideradas as candidaturas que não forem apresentadas aos organismos executores dos dois países.
A atribuição do subsídio para o 2º ano do projeto fica dependente do envio e aprovação de um relatório técnico/financeiro (modelo AQUI), a enviar ao DRI – Departamento das Relações Internacionais, 12 meses após o início do projeto.

Remember last week's hands-on on-the-fly exercise about:

1. simulating data from a regression model, and
2. fitting regression models to data.

My attempt at it – explore it – really important concepts here!!!

Regression Analysis



Regression and GLMs

- What is the analysis one should implement when we have a dependente variable and one, or more, independente variables?
- What is the potential, and the limitations, of regression analysis?
- How to interpret results?
- How can we generalize regression when the usual assumptions do not hold?

Regression analysis

Scope:

Evaluate the relationship between two (or more) variables

Objectives:

- Modeling/Explaining (an ecological event)
- Testing hypothesis
- Prediction (via a predictive model)

Regression and GLMs

The (simplest) linear model (just the equation of a line!)

$$Y = \beta_0 + \beta_1 X$$

Each observation is given by:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

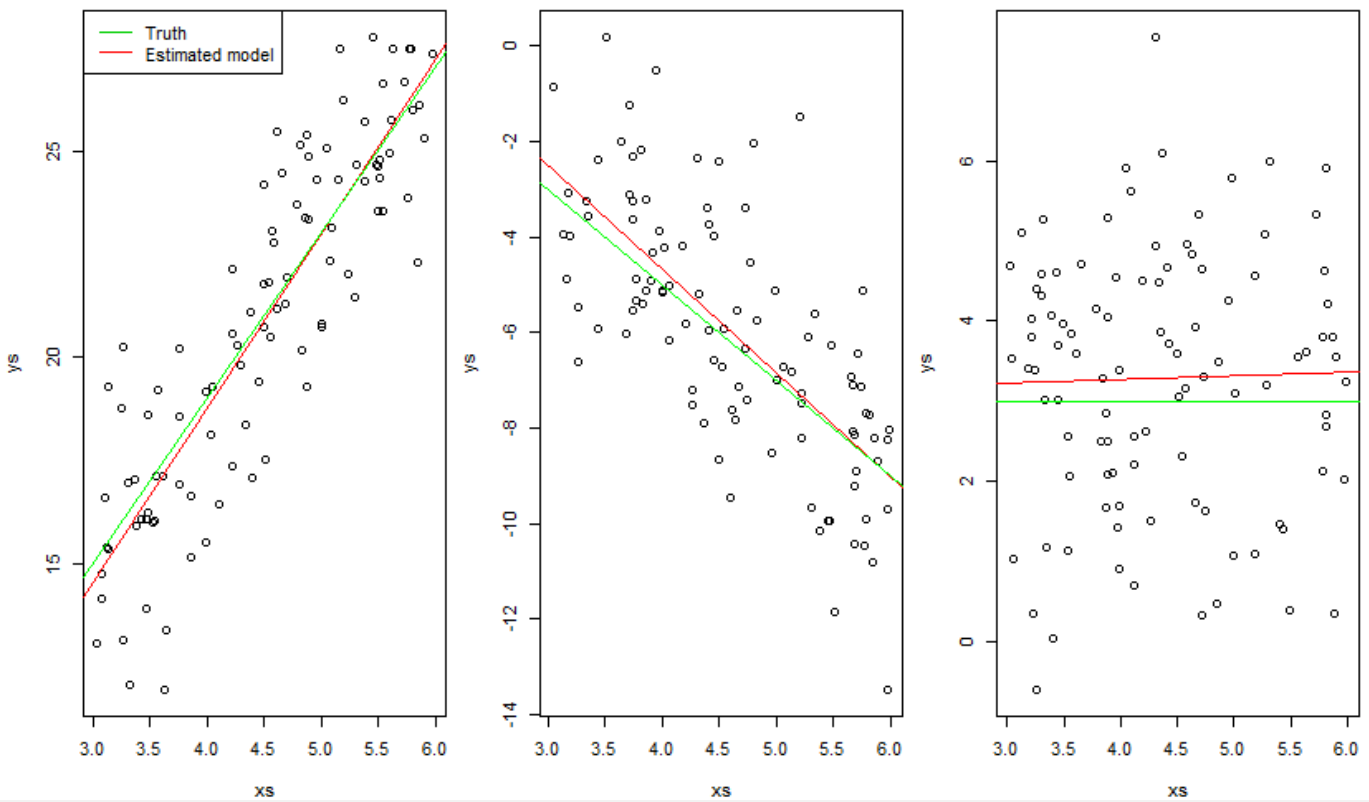
Estimators of the regression coeficients: (minimum square estimators)

$$\beta_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

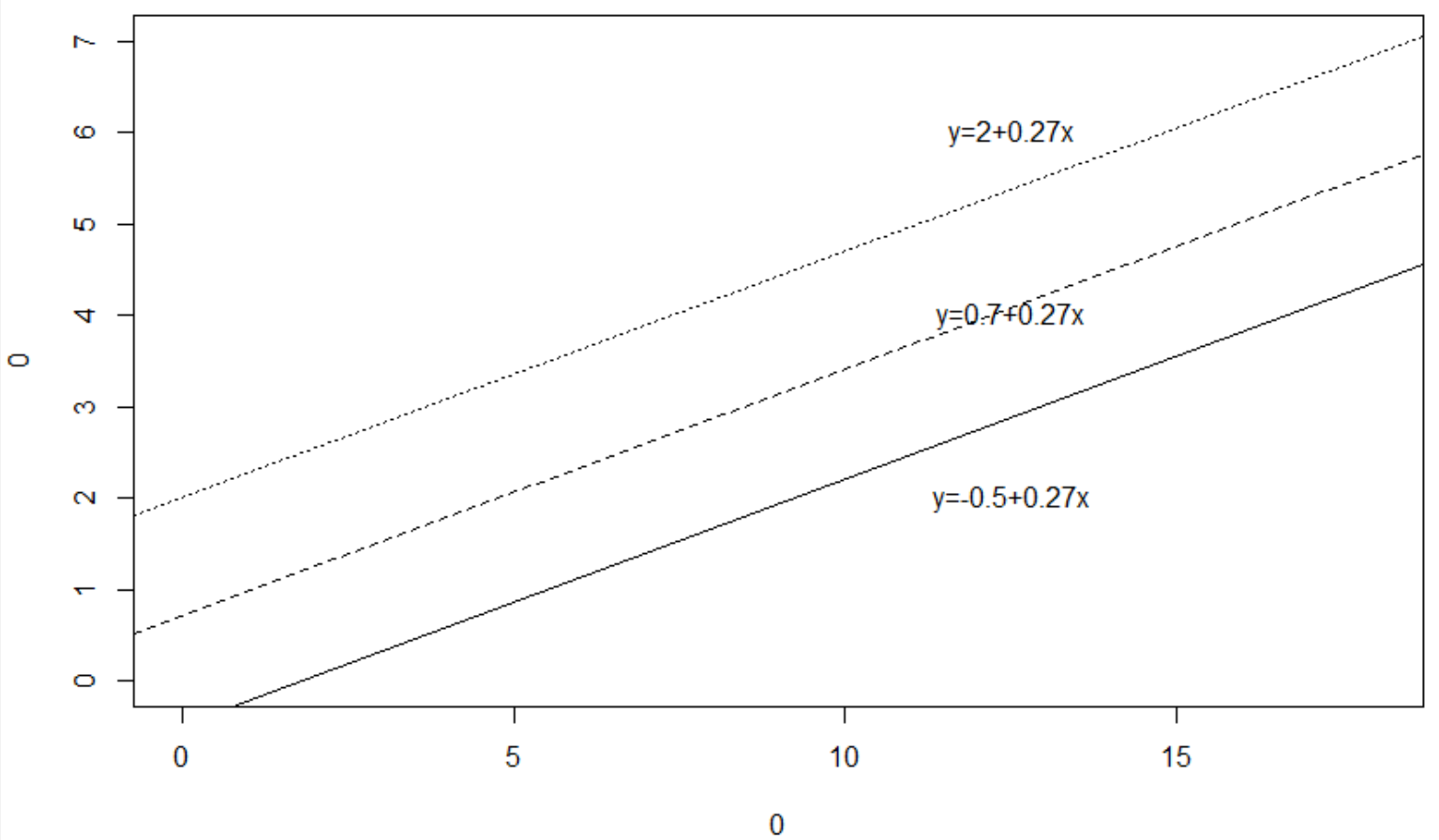
In practice in fact, we (typically) use maximum likelihood estimators!
(in this simple case, they are one and the same!)

Regression analysis



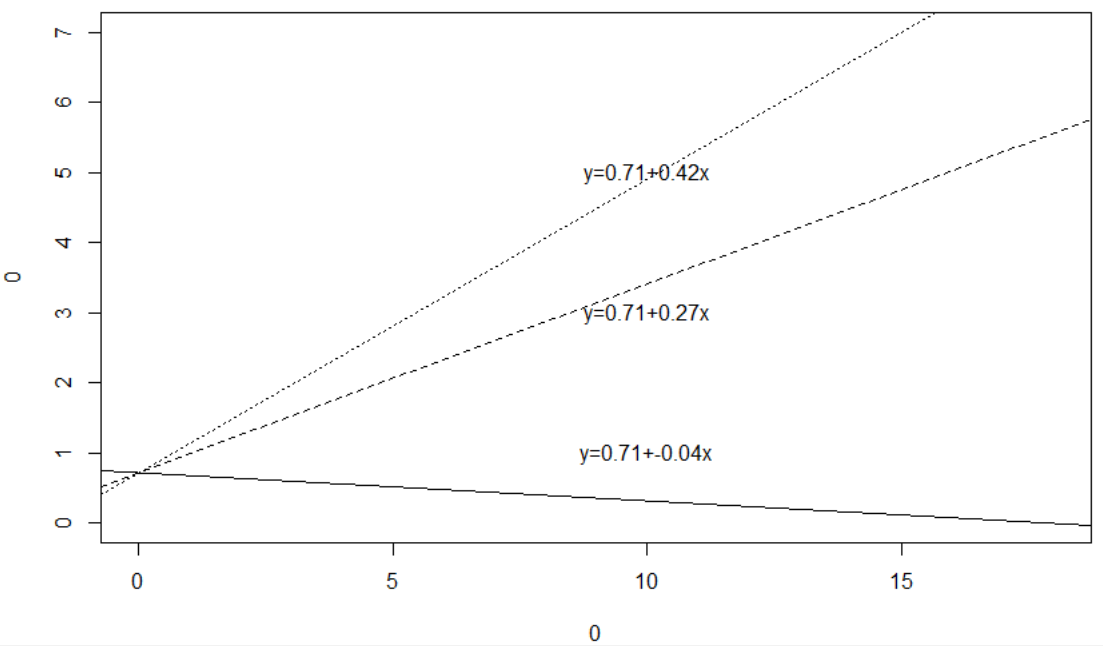
The slope (positive, negative ou null i.e. = 0) determines the type of relationship !

Análise da regressão



Same slope, different intercepts

Análise da regressão

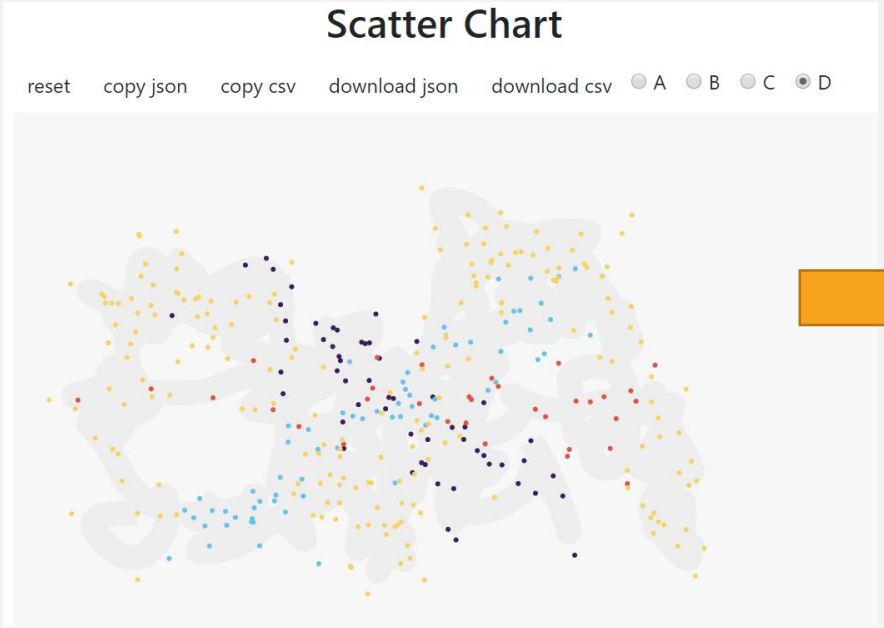


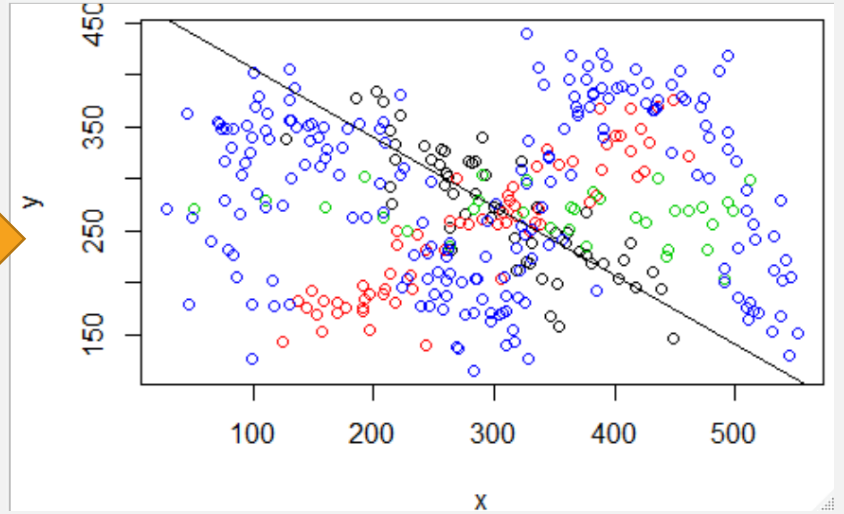
Same intercept, different slopes

http://drawdata.xyz/

Do it yourself: regression

```
data <- read.csv("C:/Users/tam2/Downloads/data.csv")
with(data,plot(y~x,col=as.numeric(z)))
dataA=data[data$z=="a",]
with(dataA,abline(lm(y~x)))
```





Hypothesis tests in regression analysis

F tests

Correct!

Hypothesis :

H_0: $\beta_0 = \beta_1 = 0$
H_1: At least one of the β_i is different from 0

Wrong!

H_0: $\beta_1 = 0$
H_1: β_1 is different from 0

$$SQ_{TOTAL} = SQ_{REGRESSION} + SQ_{residual\ (error)}$$

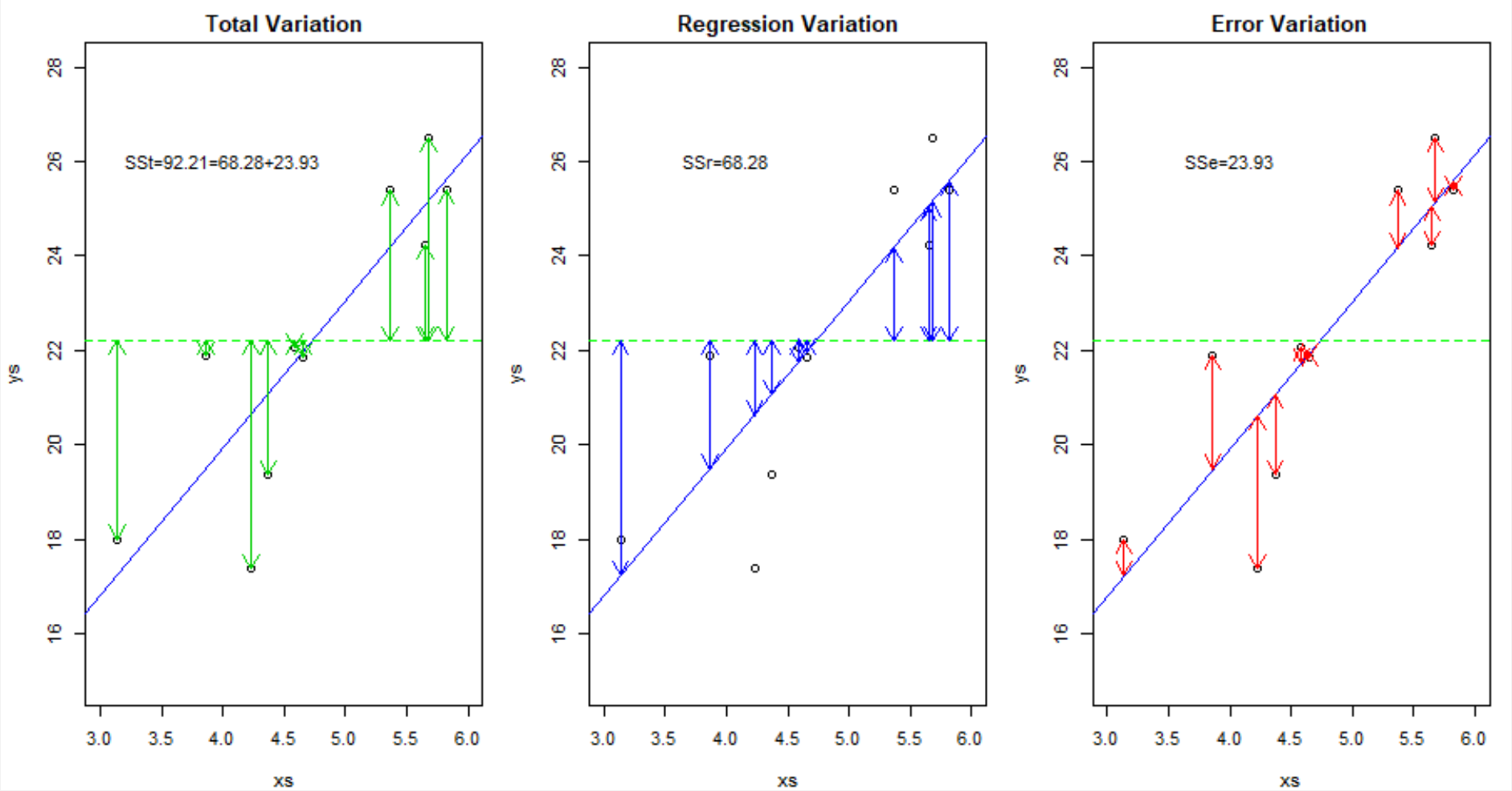
Hypothesis tests in regression analysis

$$SQ_{TOTAL} = SQ_{REGRESSION} + SQ_{residual\ (error)}$$

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

$$SQ_{TOTAL} = SQ_{REGRESSÃO} + SQ_{residual\ (erro)}$$

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$



Testes de hipóteses na análise da regressão

Estatística de teste:

$$F = \frac{\frac{SQ_{REGRESSÃO}}{\upsilon_{REGRESSÃO}}}{\frac{SQ_{ERRO}}{\upsilon_{ERRO}}} = \frac{QM_{REGRESSÃO}}{QM_{ERRO}}$$

sendo $gl_{regressão}$ = p -1 (p=nº de coeficientes) e
gl_{erro} = n - 2 (n=nº observações)

Valor crítico:

$$F_{\alpha,\upsilon_{reg},\upsilon_{erro}}$$

Critério de decisão:

Rejeitar H_0 se: $F > F_{\alpha,\upsilon_{reg},\upsilon_{erro}}$

Não rejeitar H_0 caso contrário

Implementing in R

```
> summary(lm(y~x,data=dataA))

Call:
lm(formula = y ~ x, data = dataA)

Residuals:
    Min      1Q  Median      3Q     Max
-78.797 -18.089   3.943  24.103  59.315

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 472.28619   19.20313   24.59  < 2e-16 ***
x            -0.66473    0.06147  -10.81 3.13e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.98 on 55 degrees of freedom
Multiple R-squared:  0.6801,    Adjusted R-squared:  0.6743
F-statistic:   117 on 1 and 55 DF,  p-value: 3.135e-15
```

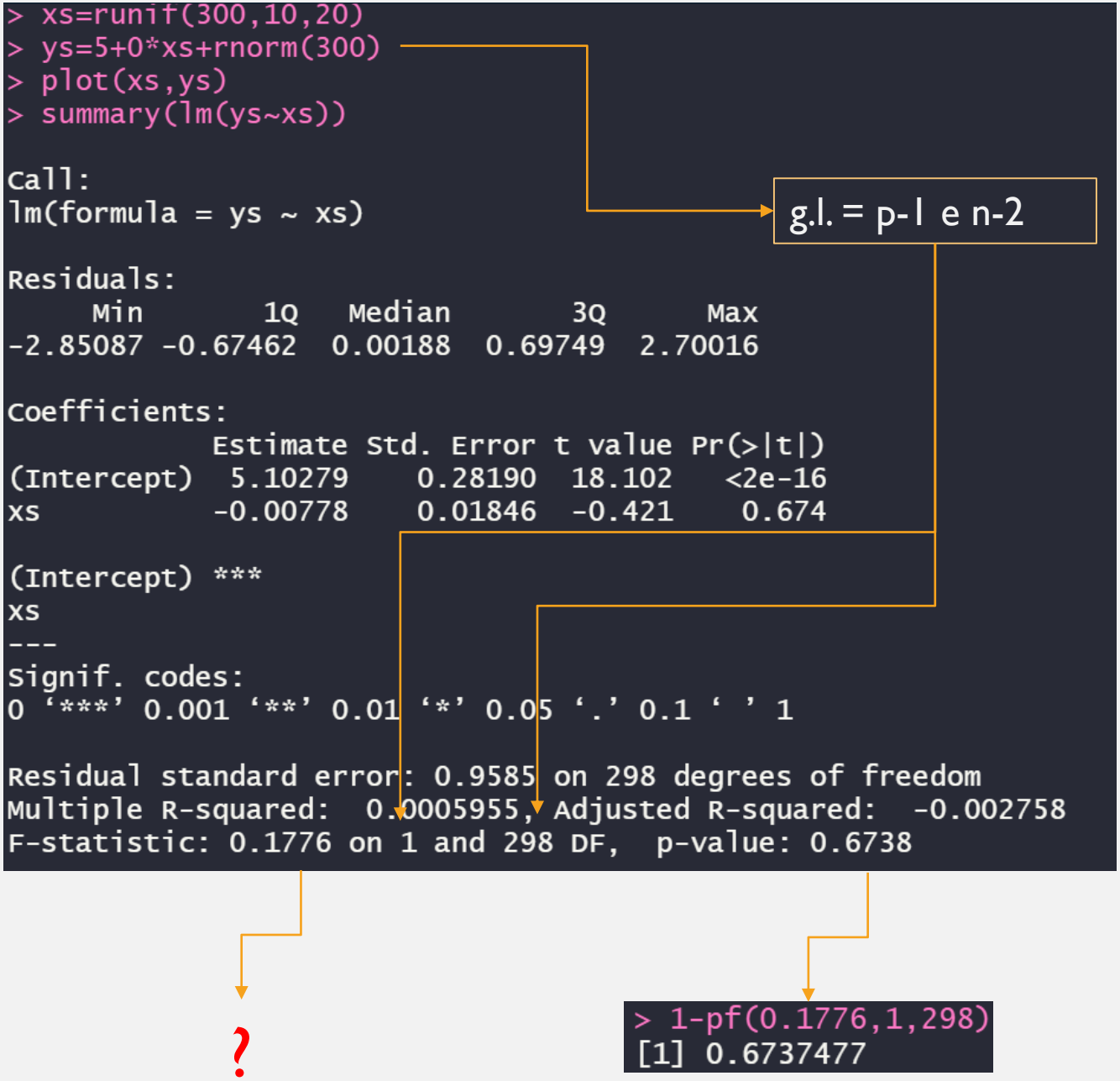
http://analyticspro.org/2016/03/15/r-tutorial-how-to-interpret-f-statistic-in-regression-models/

Challenge:

Do it yourself, based on the formulas!

Calculate:

- Degrees of freedom
- Test statistic
- P-value!



Variance explained by the regression model

The determination coeficient R^2

$$R^2 = \frac{SQ_{REGRESSÃO}}{SQ_{TOTAL}}$$

Is the portion of total variability explained by the regression model, and it is a measure of the adequacy of the linear relationship, i.e. of the fit of the model.

$$0 \leq R^2 \leq 1$$

R is the (linear) correlation coefficient between Y and X

Multiple Regression

Multiple regression analysis

Scope:

Evaluate the relationship between a dependent variable and multiple independent variables

Objectives (just as before!!):

- Modeling/Explaining (an ecological event)
- Testing hypothesis
- Prediction (via a predictive model)

Análise da regressão múltipla

When we have more than a single independente variable

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Each observation is now given by:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2_i} + \ldots + \beta_p X_{p_i} + \varepsilon_i$$

Multiple Regression Analysis

Variability explained by the model

O coeficiente de determinação R^2 depende de *p* e *n* e, por isso, uma estimativa mais adequada é o R^2 ajustado, dado por:

$$R_a^2 = 1 - \frac{MS_{RESIDUAL}}{MS_{TOTAL}} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

Testes de hipóteses na análise da regressão

Testes F

Hipóteses:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$

H_1: Pelo menos um dos β_i é diferente de 0

Testes de hipóteses na análise da regressão

Estatística de teste:

$$F = \frac{\frac{SQ_{REGRESSÃO}}{\upsilon_{REGRESSÃO}}}{\frac{SQ_{ERRO}}{\upsilon_{ERRO}}} = \frac{QM_{REGRESSÃO}}{QM_{ERRO}}$$

sendo $gl_{regressão}$ = p -1 (p=nº de coefientes) e
gl_{erro} = n - 2 (n=nº observações)

Valor crítico:

$$F_{\alpha,\upsilon_{reg},\upsilon_{erro}}$$

Critério de decisão:

Rejeitar H_0 se:

$$F > F_{\alpha,\upsilon_{reg},\upsilon_{erro}}$$

Não rejeitar H_0 caso contrário

gl=p-1=2-1=1

gl=n-2=10-2=8

```
> summary(aov(ys~xs))
            Df Sum Sq Mean Sq F value Pr(>F)
xs           1  68.28   68.28   22.82 0.0014 **
Residuals    8  23.93    2.99
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(lm(ys~xs))

Call:
lm(formula = ys ~ xs)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2431 -0.6694  0.1048  1.1045  2.3948

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.454      3.135   2.377   0.0447 *
xs             3.115      0.652   4.778   0.0014 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.73 on 8 degrees of freedom
Multiple R-squared:  0.7405,    Adjusted R-squared:  0.708
F-statistic: 22.82 on 1 and 8 DF,  p-value: 0.001395
```

H0: xs e ys são independentes, ou xs não influencia ys

```
> 1-pf(22.82,1,8)
[1] 0.001395977
> 2*(1-pt(4.778,8))
[1] 0.00139424
```

H0: $\beta_0=0$
H0: $\beta_1=0$

No caso de uma regressão multipla com p parâmetros n-2= n-p

Testes de hipóteses na análise da regressão

Hipótese:

H_0: $\beta_i=0$
H_1: $\beta_i \neq 0$

Estatística de teste:

$$t = \frac{\hat{\beta} - 0}{s_{\hat{\beta}}}$$

onde

$$s_{\hat{\beta}} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n-2}} \Big/ \sum x^2$$

Valor crítico:

$$t_{\alpha(2),n-2}$$

Critério de decisão:

Rejeitar H_0 se: $t > t_{\alpha(2),n-2}$

Não rejeitar H_0 caso contrário

Assumptions of linear regression

- The distribution of the errors is Gaussian with mean 0 and constant variance σ^2;

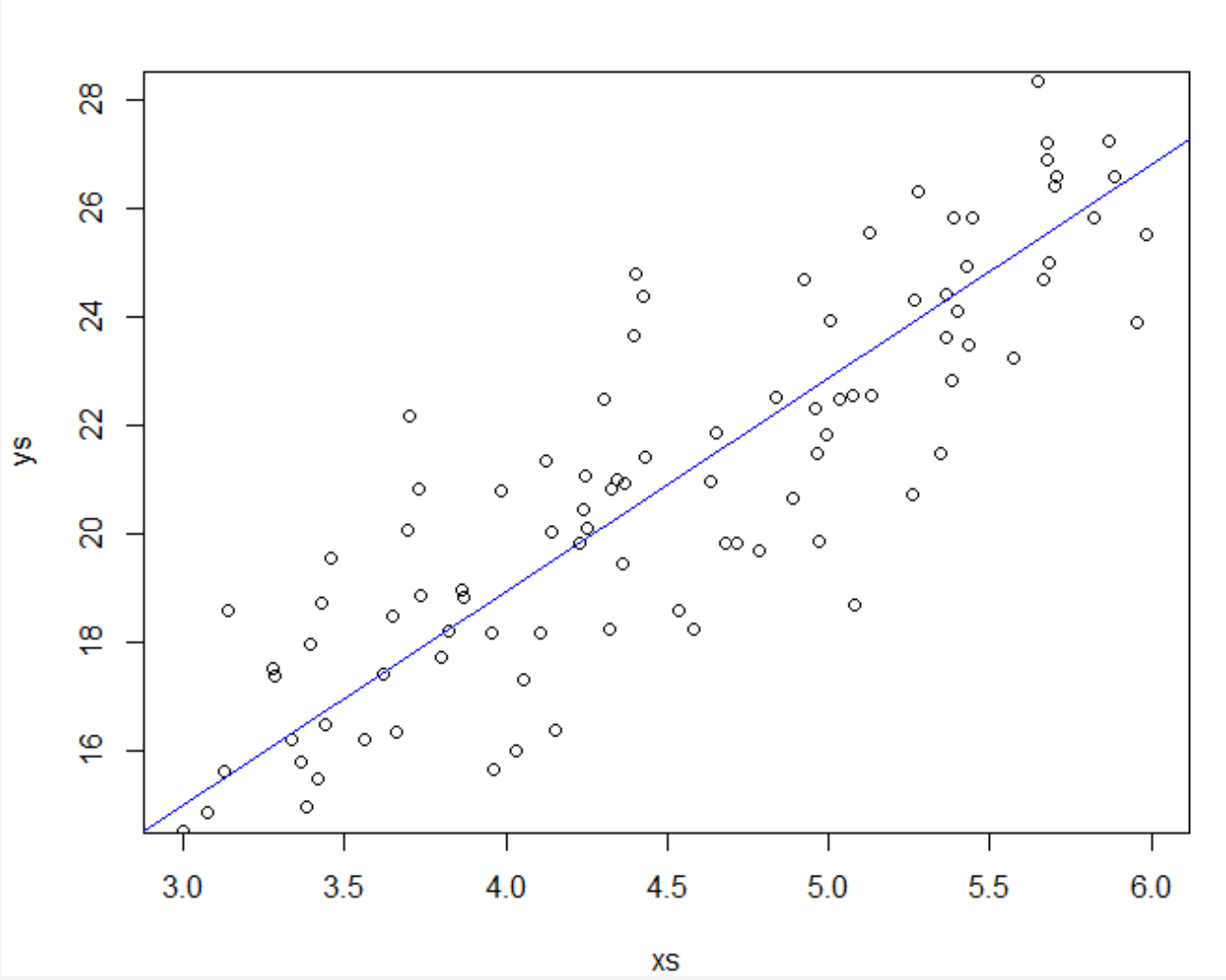
- The residuals/errors are independent, i.e. uncorrelated;

- There are no errors in X (in practice, measurements in X are made with negligible error compared to those of Y)

How to evaluate assumptions: residual analysis

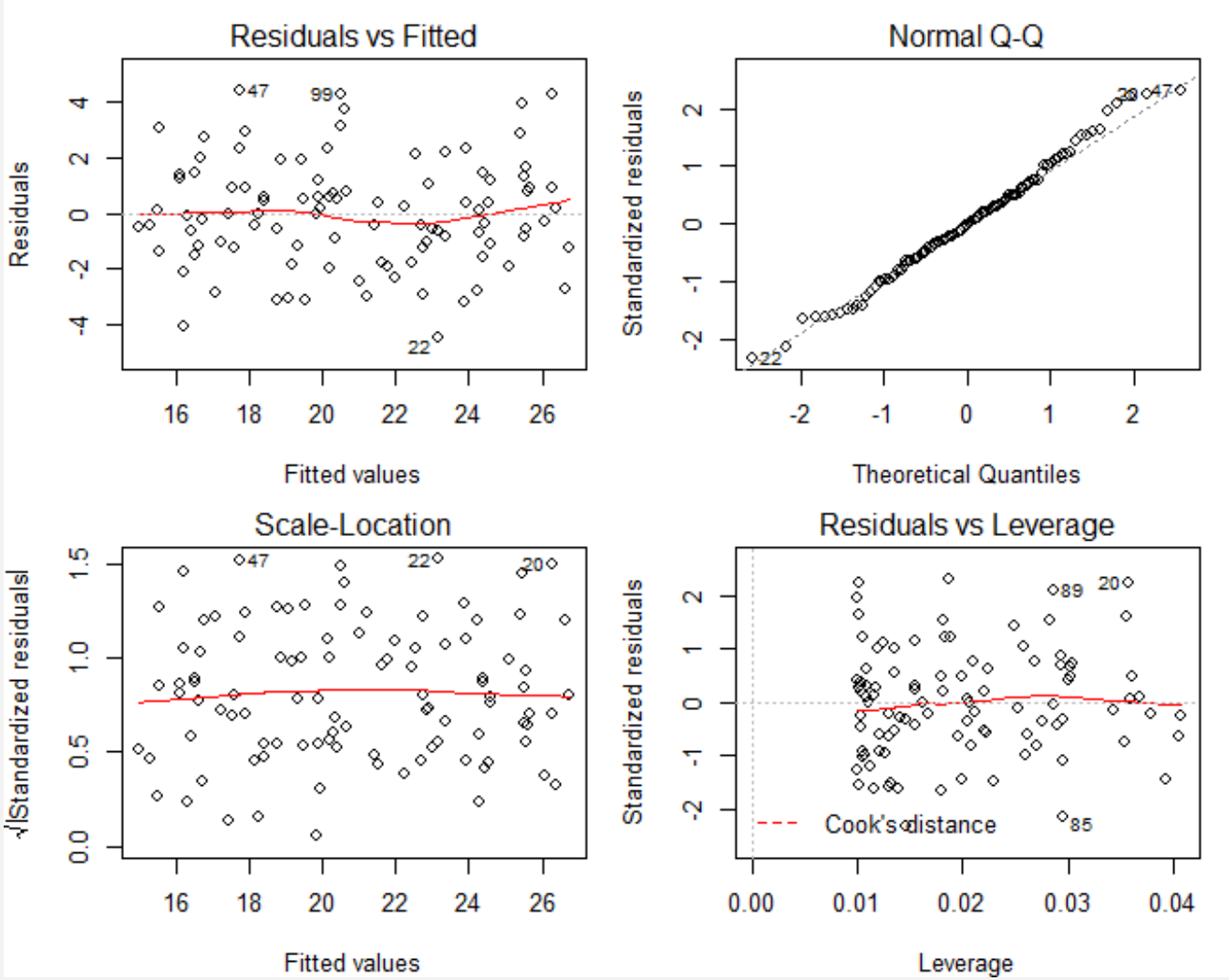
- Gaussian errors
- Homogeneity of variances
- Independent errors (absence of autocorrelation of residuals)

```
par(mfrow=c(1,1),mar=c(4,4,2,1))
set.seed(123)
n=100;slope=4;intercept=3
xs=runif(n,3,6)
ys=intercept+slope*xs+rnorm(n,mean=0,sd=2)
plot(xs,ys,ylim=c(15,28),xlim=c(3,6),main="")
mylm1=lm(ys~xs)
abline(mylm1,col=4)
```

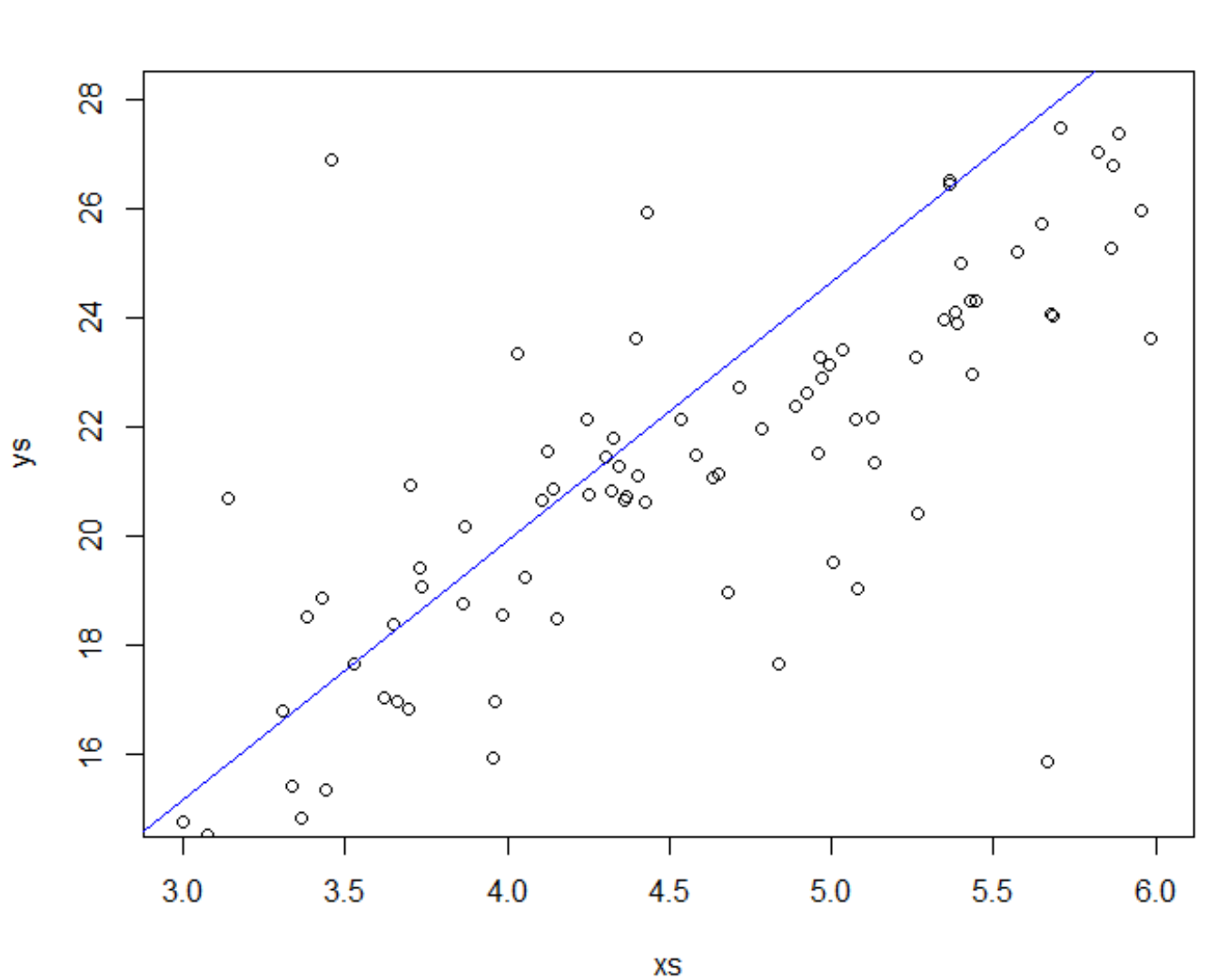


```
par(mfrow=c(2,2),mar=c(4,4,2,1))
plot(mylm1)
```

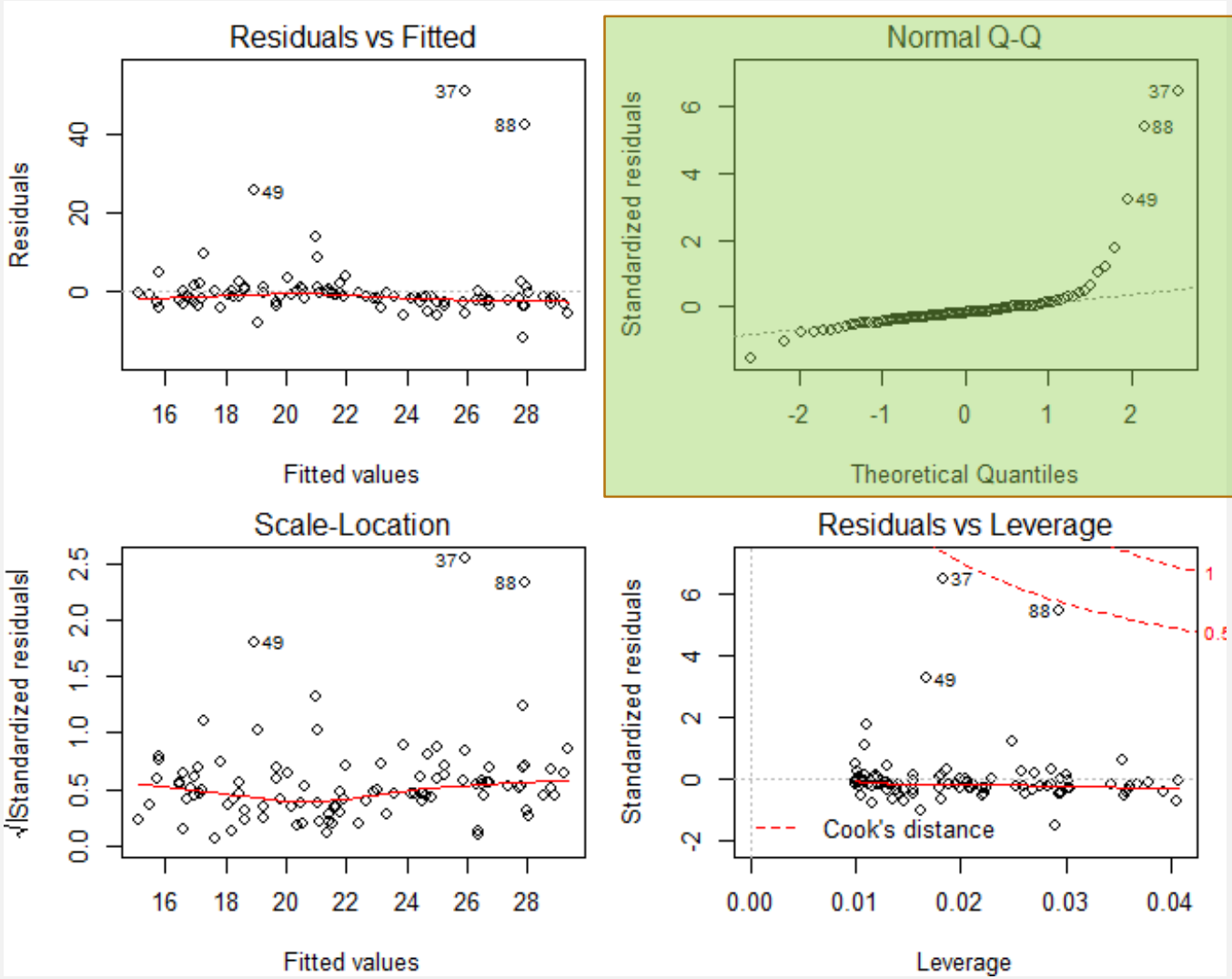
Análise de resíduos: normalidade, homocedasticidade e independência



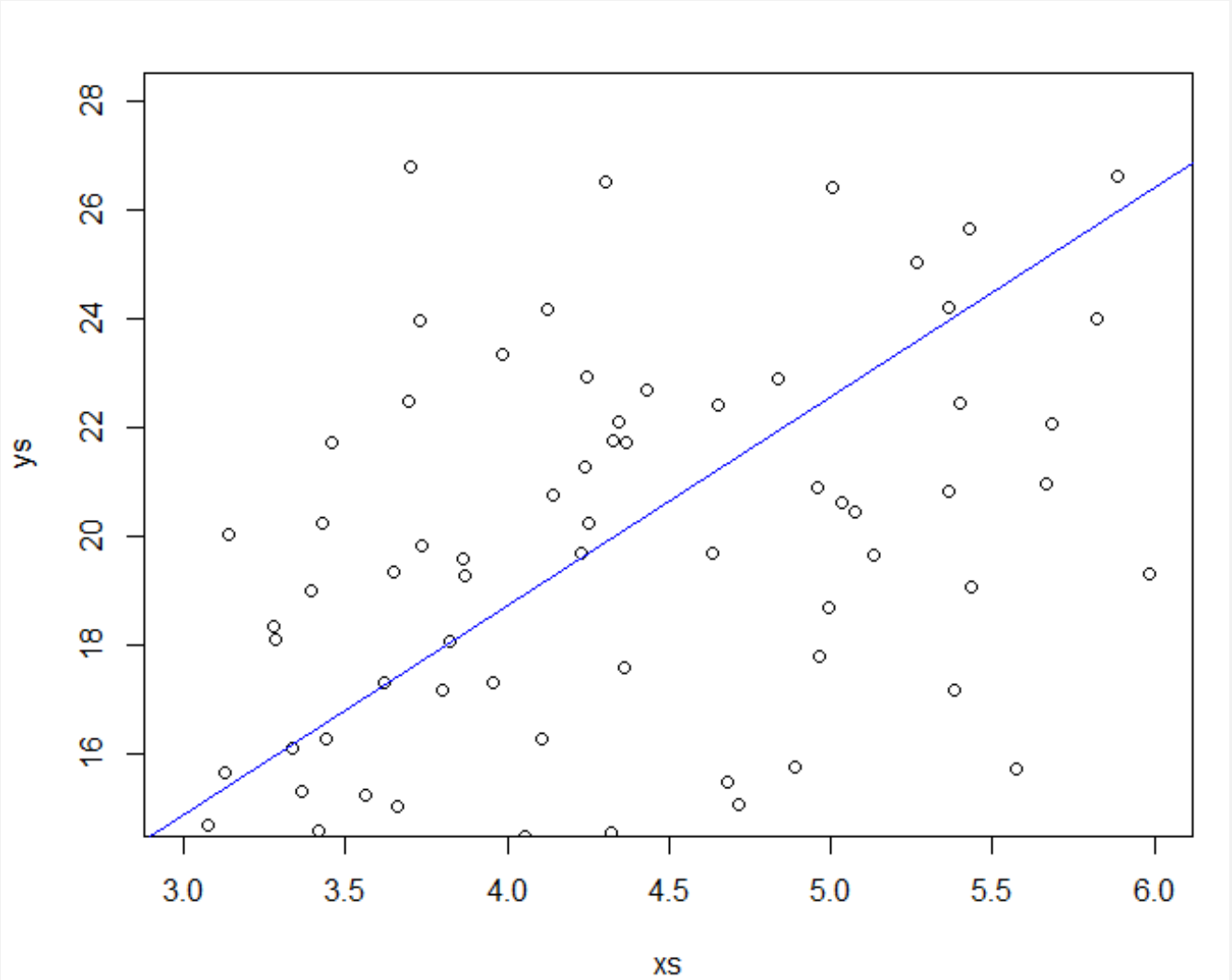
Non-Gaussian Errors



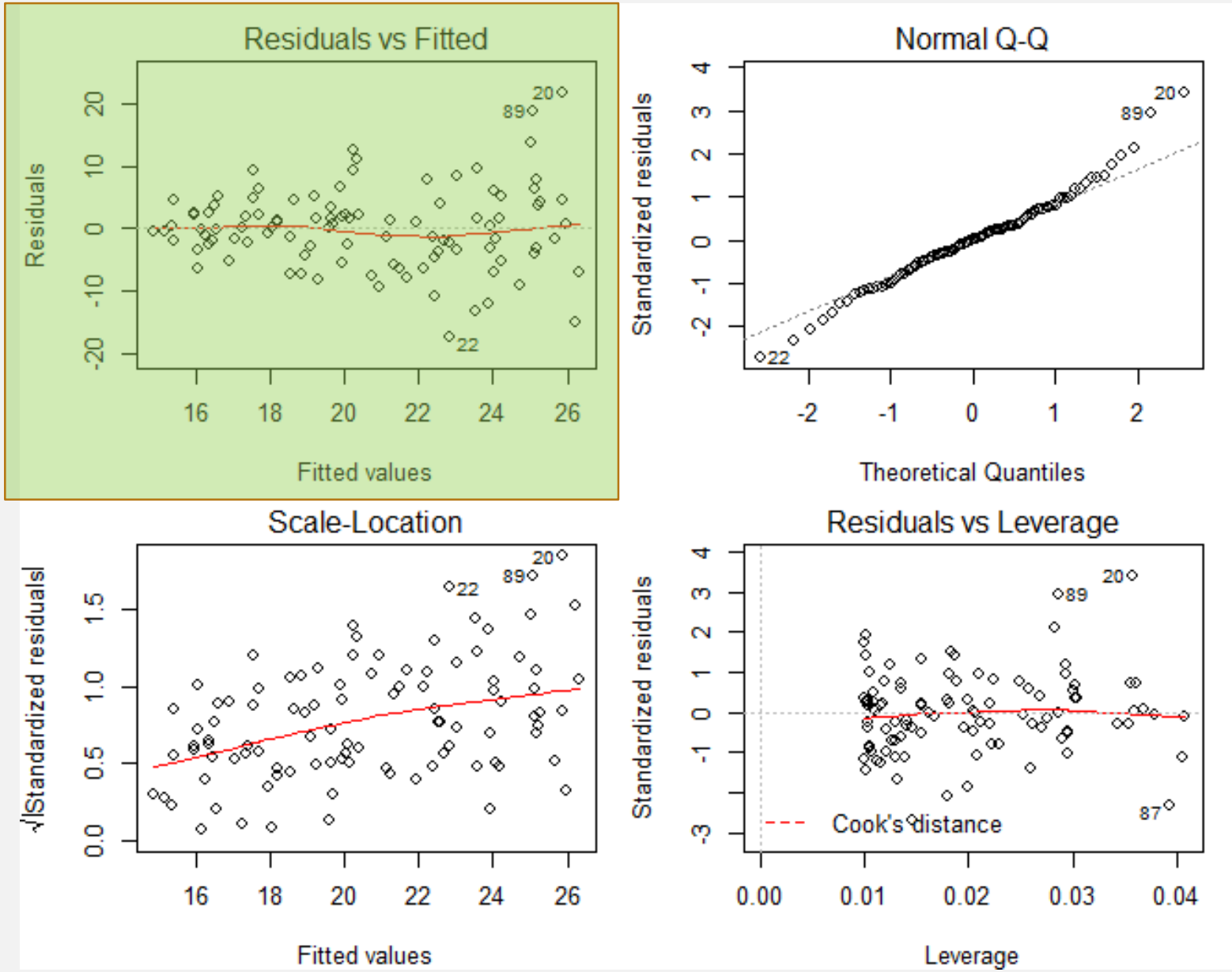
Non-Gaussian Errors



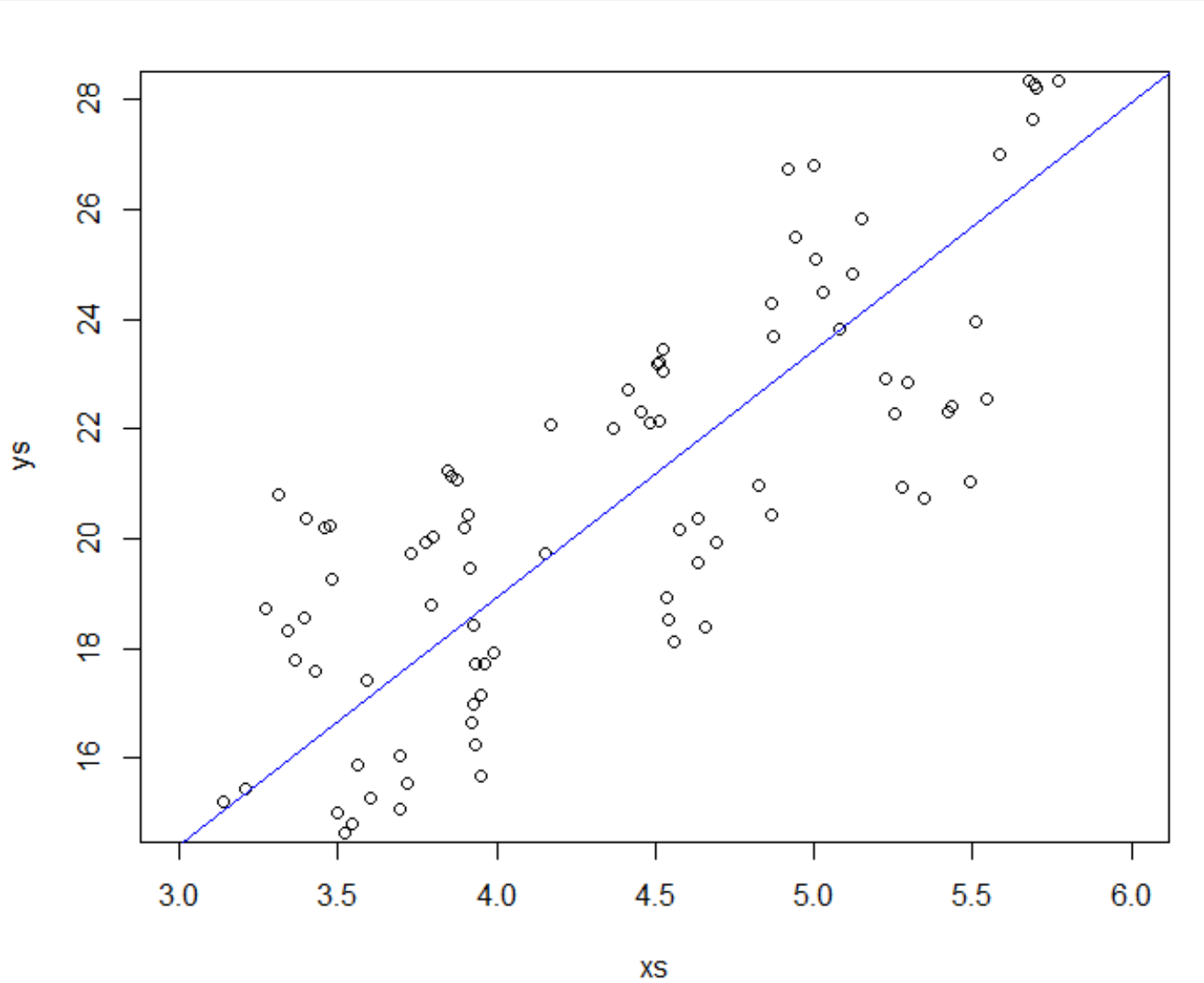
Heterocedastic errors



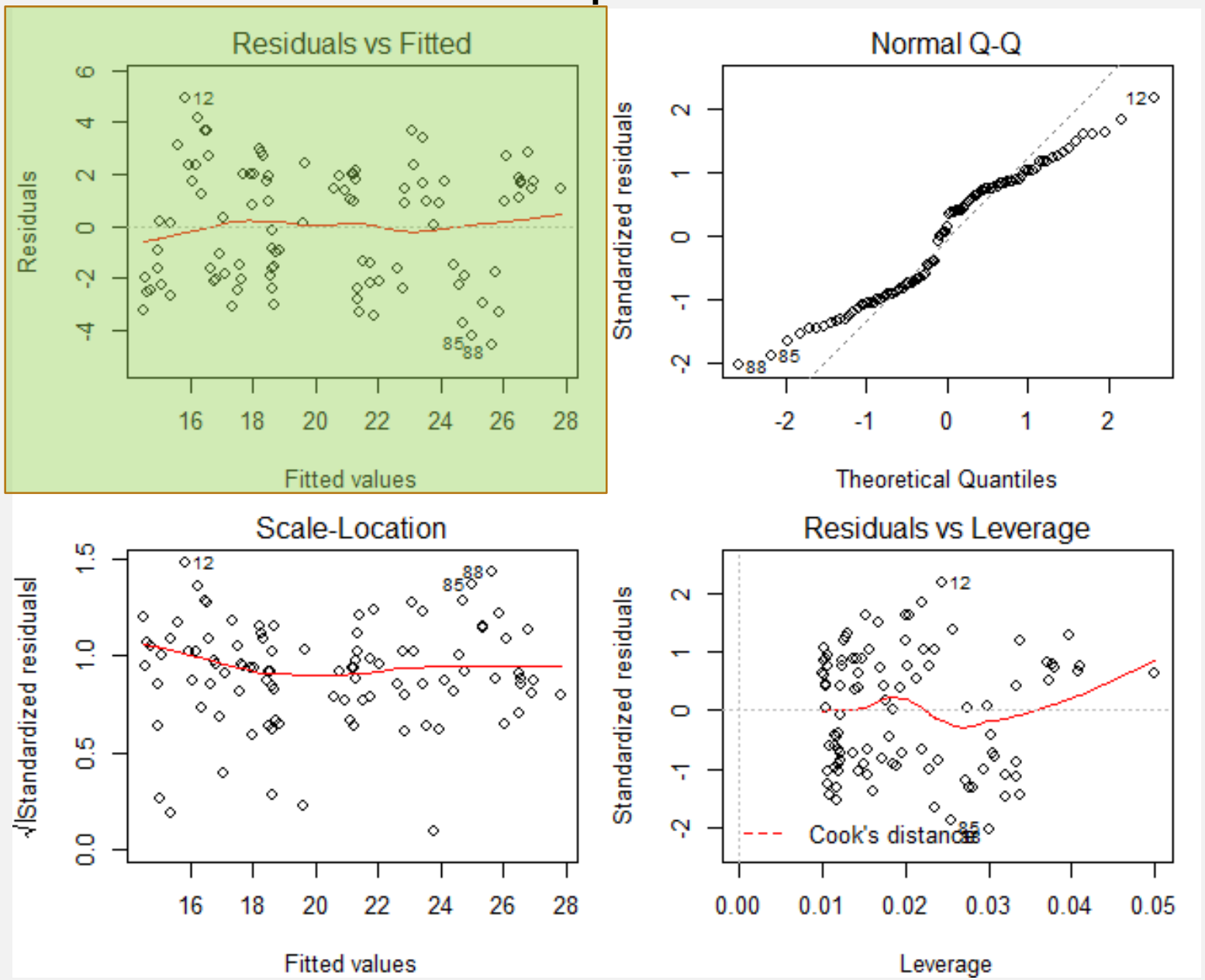
Heterocedastic errors



Non-independent errors



Non-independent errors



Análise de resíduos

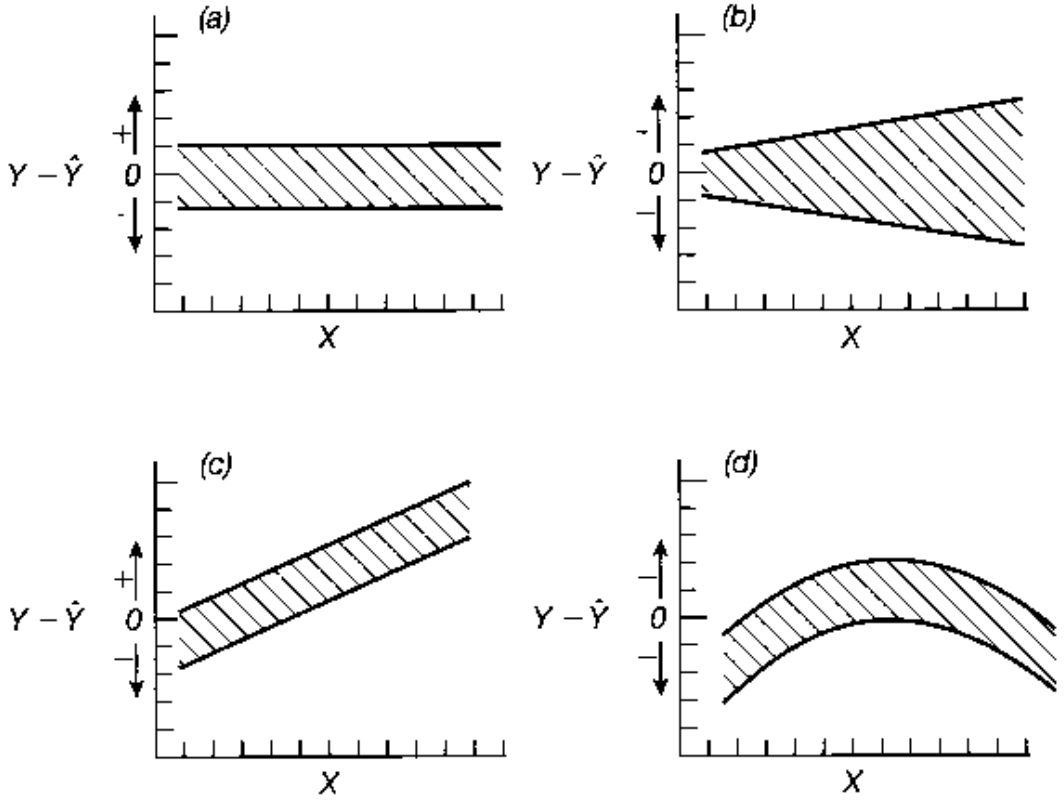
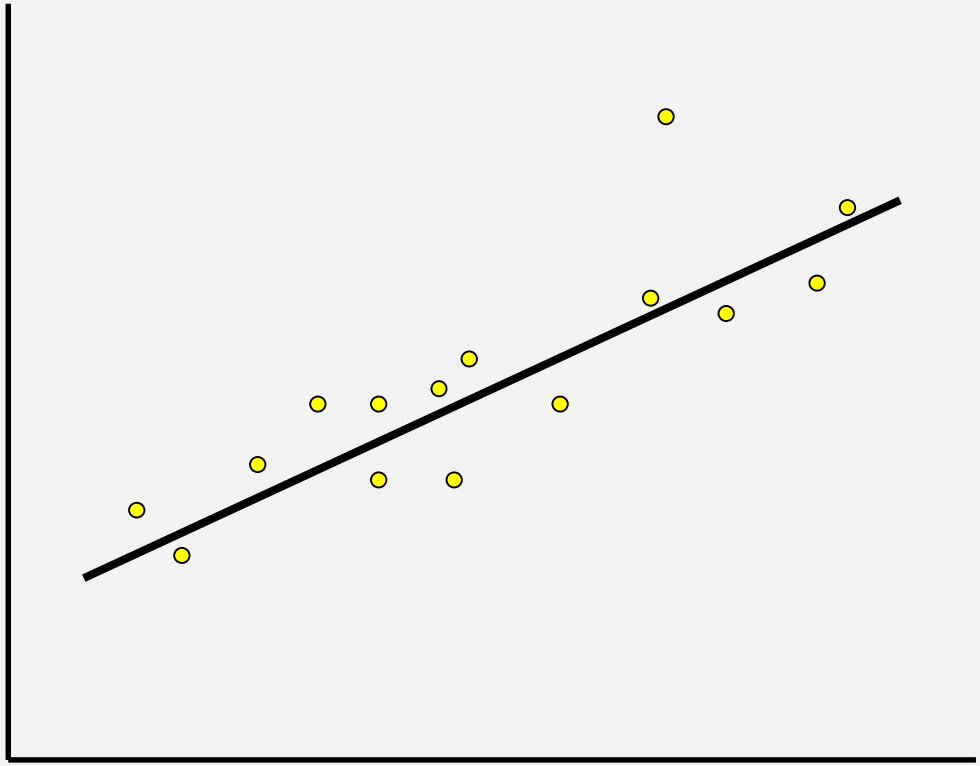


Figure 16.12 The plotting of residuals. (a) Data exhibiting homoscedasticity. (b) Data with heteroscedasticity of the sort in Example 16.9. (c) Data for which there was likely an error in the regression calculations, or an additional variable is needed in the regression model. (d) Data for which a linear regression does not accurately describe the relationship between Y and X, and a curvilinear relationship should be considered.

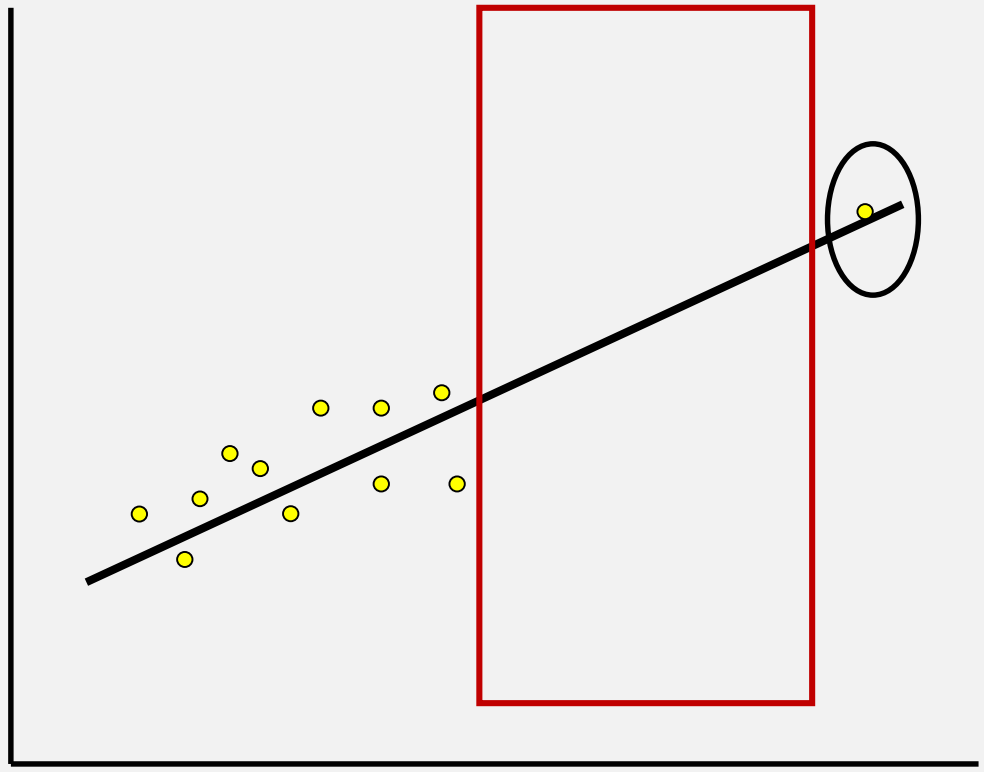
Residual Analysis : outliers e influential observations

- Outliers (exploratory analysis and residual analysis)
- Influential observations (e.g. Cook distance)

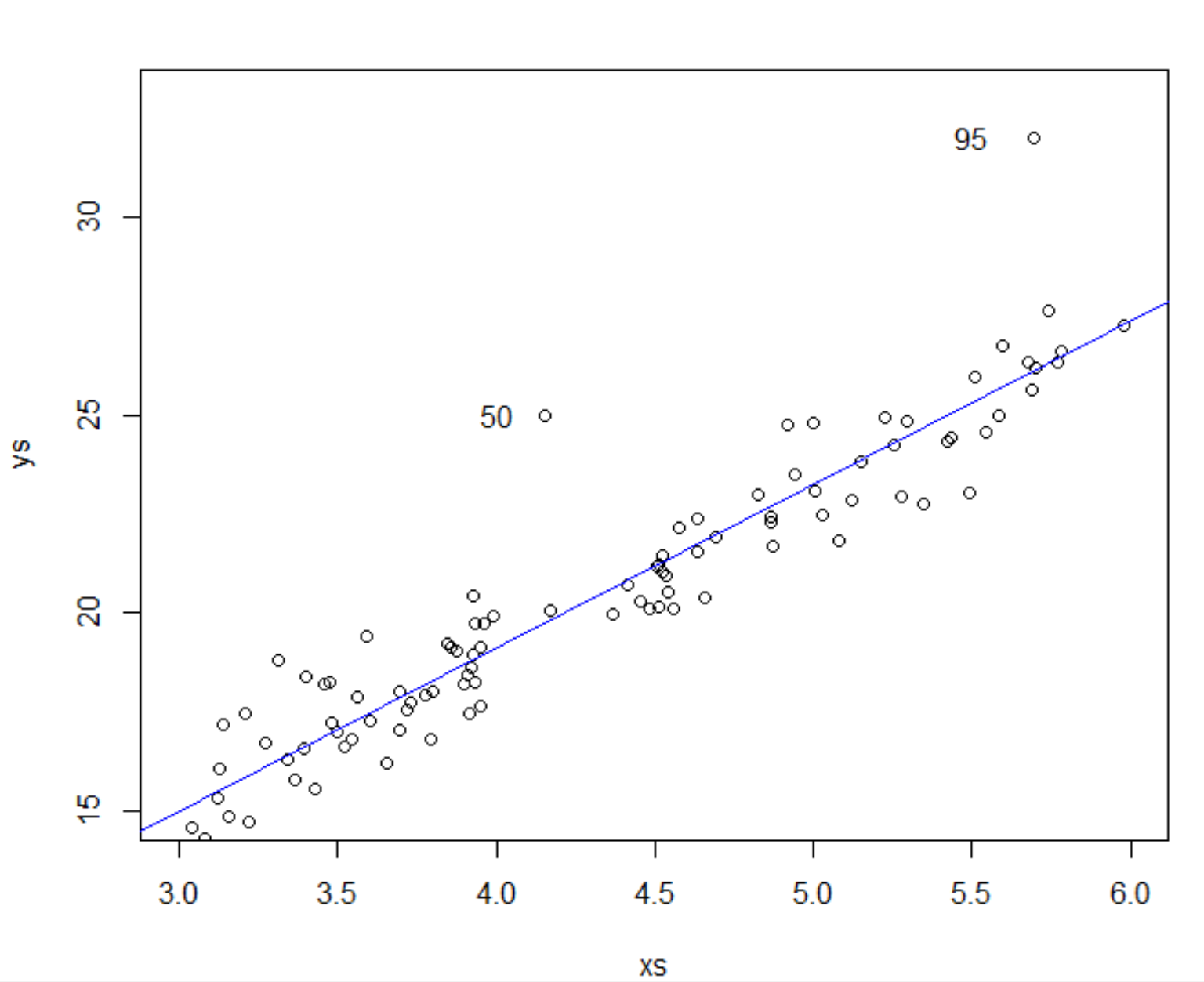
Análise de resíduos: outliers e observações influentes



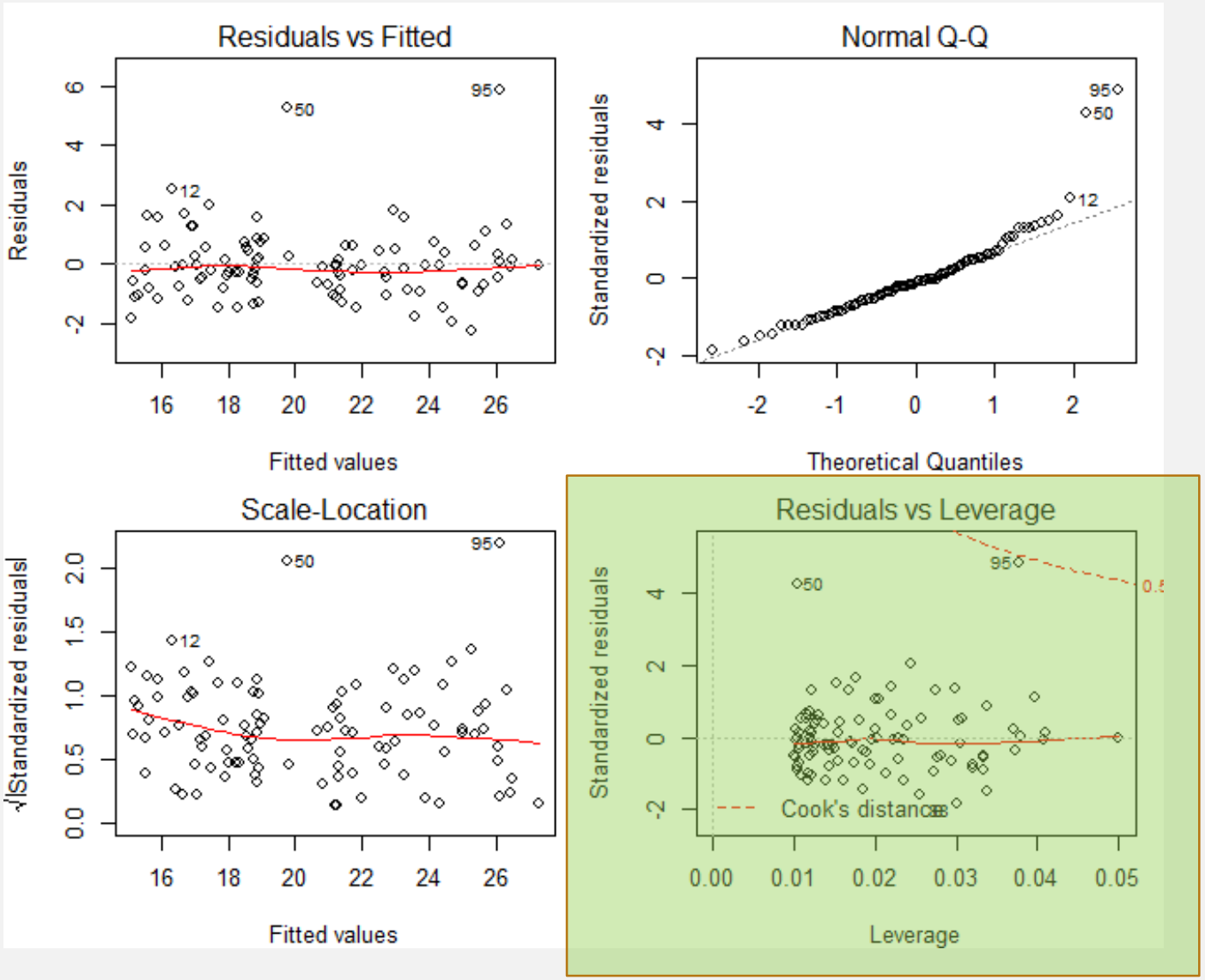
Análise de resíduos: outliers e observações influentes



Influential observations



Influential observations



Model selection in regression

We often want to select a subset of available covariates to use in a regression model (those more important/relevant to explain the response) and hence obtain a simplified model.

The regression coefficients of the simplified models are different from those of the original model (saturated model, if the model with all variables)

Framework:

- statistical tests
- Information criteria

Method:

- Forward selection
- Backward elimination
- Stepwise selection
- Test all combinations

Some aditional topics about regression

- "Dummy" variables (used to code factos, R does it for you! – but is key to understand it – interpreting model coefficients depends on it)

- Interaction between independent variables (multicollinearity – one should only consider variable ecologically relevant and remove those highly correlated, especially those less ecologically relevant)

- Other fitting approaches not MSM (Maximum likelihood, REML)

- Non-linear models