Aula 18 Goodies*

* Goodies related to animals, plants and numbers…

EMMA 2019

ENCONTRO MAMÍFEROS MARINHOS 2019

DATA E INSCRIÇÃO ORGANIZAÇÃO PROGRAMA SOBRE O EMMA 2019

Sobre o EMMA 2019

Há 10 anos foi realizada em Peniche uma iniciativa inédita, o 1º Simpósio Biologia e Conservação de Mamíferos Aquáticos. Foi um momento importante em que um número elevado de investigadores Portugueses tiveram a oportunidade de interagir e discutir sobre o trabalho desenvolvido em Portugal até ao momento.

O Encontro de Mamíferos Marinhos 2019 assinala uma década sobre esse primeiro esforço. E pretende juntar num mesmo local a comunidade científica e empresarial que trabalha, trabalhou e pretende trabalhar em mamíferos marinhos em Portugal. Irão ser discutidos os desafios que se apresentam à investigação nesta área, o que está a ser desenvolvido, o que falta fazer e como se poderá colmatar estas lacunas no futuro.

EMMA

http://emma2019.campus.ciencias.ulisboa.pt/

Spurious correlations

Now a ridiculous book!

- Spurious charts
- Fascinating factoids
- Commentary in the footnotes

Amazon | Barnes & Noble | Indie Bound

http://tylervigen.com/spurious-correlations

We all know the truism "Correlation doesn't imply causation," but when we see lines sloping together, bars rising together, or points on a scatterplot clustering, the data practically begs us to assign a reason. We want to believe one exists.

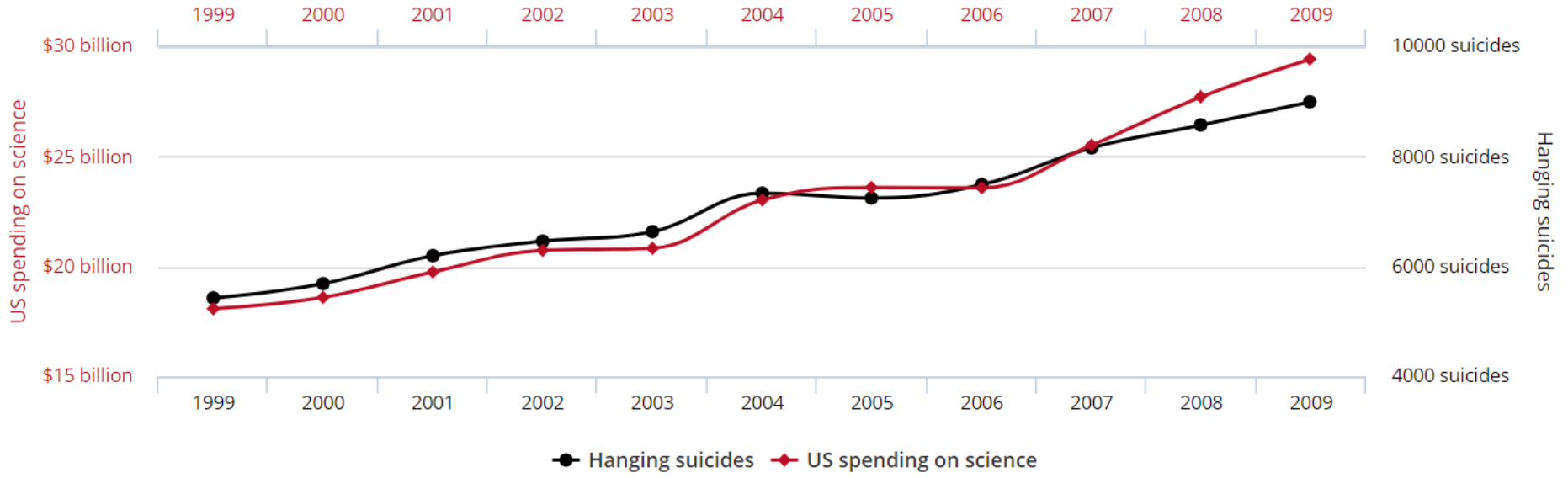
Statistically we can't make that leap, however. Charts that show a close correlation are often relying on a visual parlor trick to imply a relationship. Tyler Vigen, a JD student at Harvard Law School and the author of *Spurious Correlations,* has made sport of this on his website, which charts farcical correlations—for example, between U.S. per capita margarine consumption and the divorce rate in Maine.

Vigen has programmed his site so that anyone can find and chart absurd correlations in large data sets.

https://hbr.org/2015/06/beware-spurious-correlations

US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation

Correlation: 99.79% (r=0.99789126)

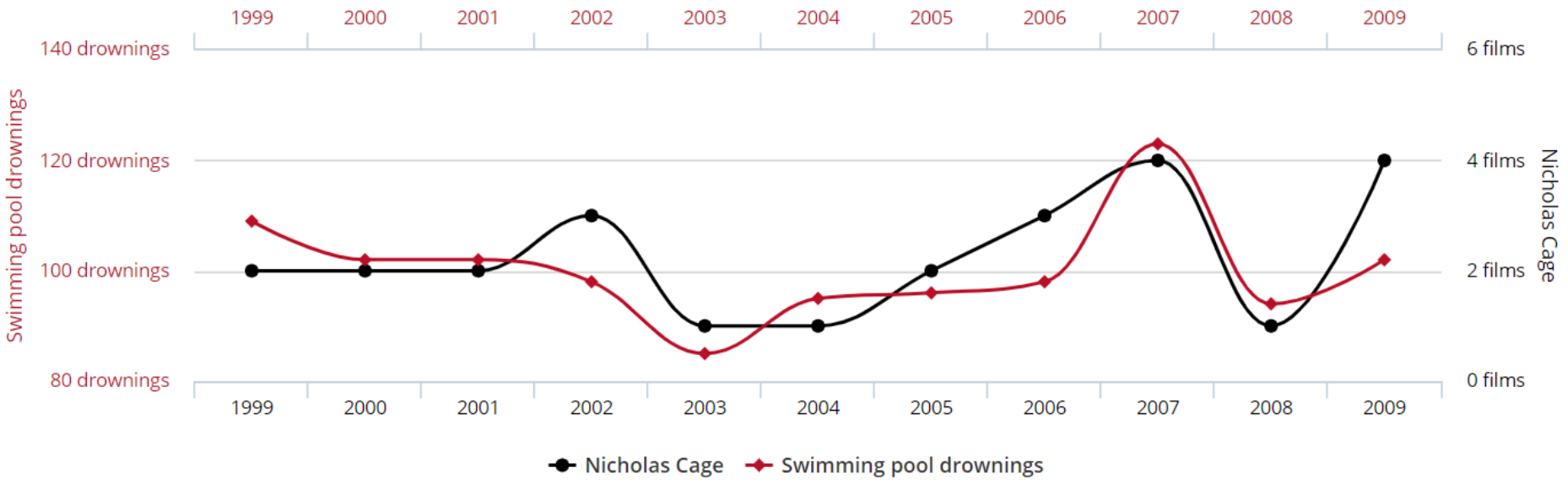


tylervigen.com

Data sources: U.S. Office of Management and Budget and Centers for Disease Control & Prevention

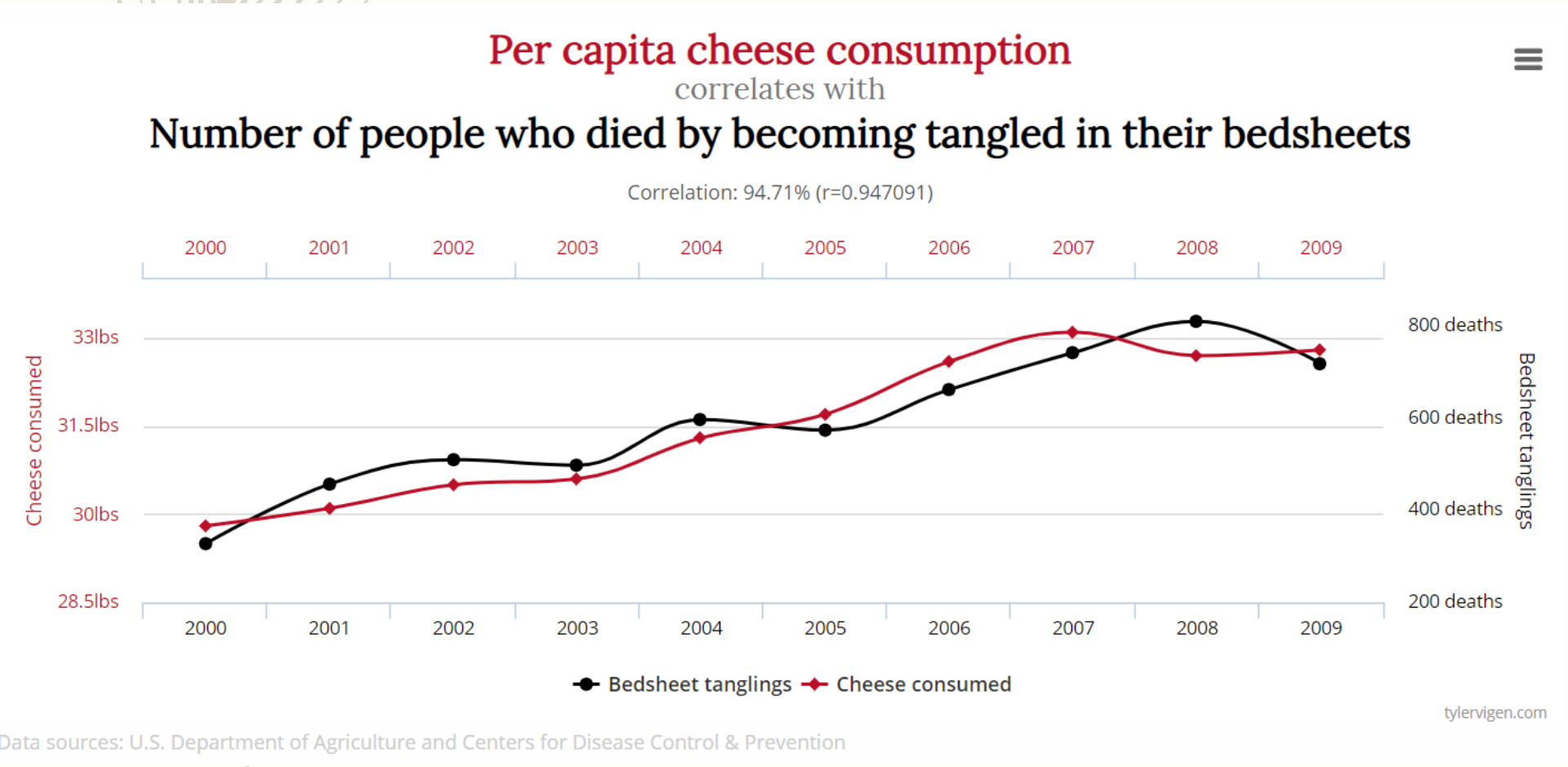
Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in

Correlation: 66.6% (r=0.666004)



tylervigen.com

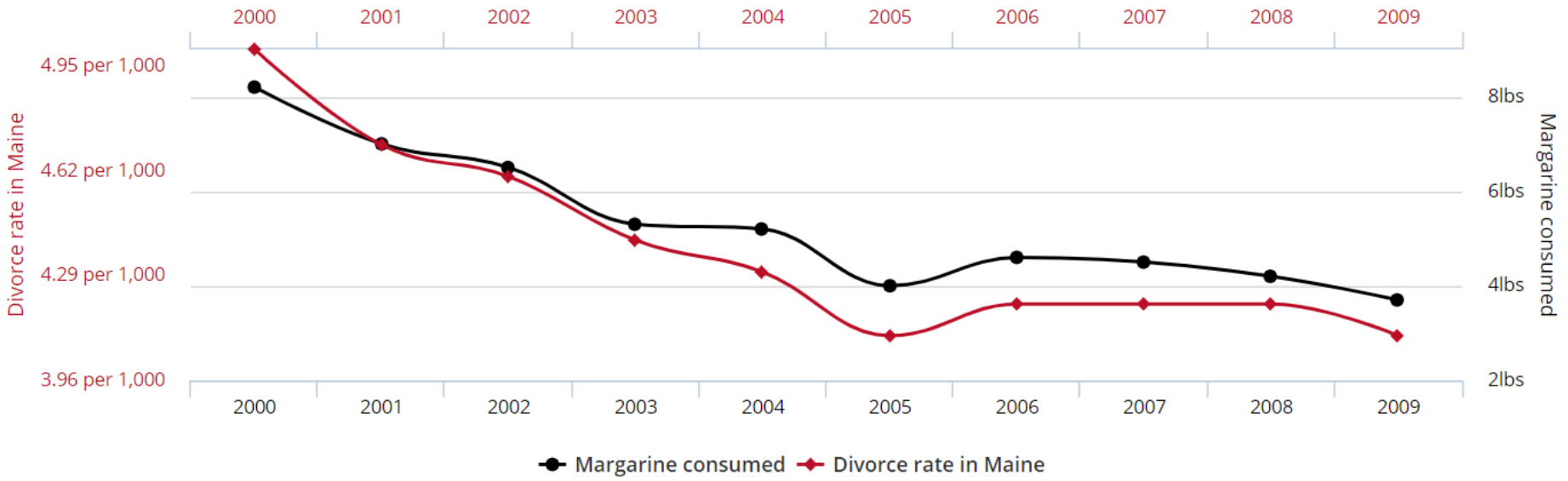
Data sources: Centers for Disease Control & Prevention and Internet Movie Database

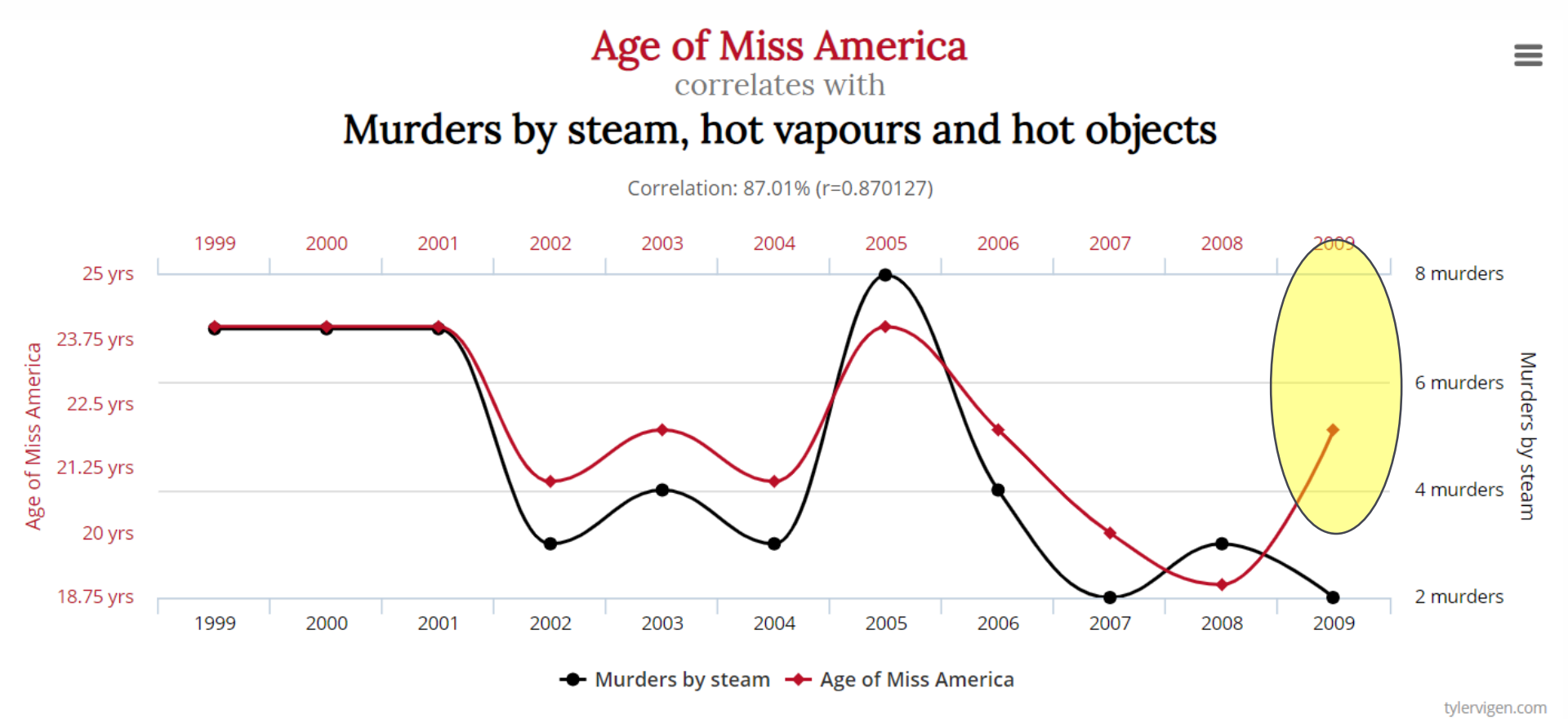


Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

Divorce rate in Maine
correlates with
Per capita consumption of margarine

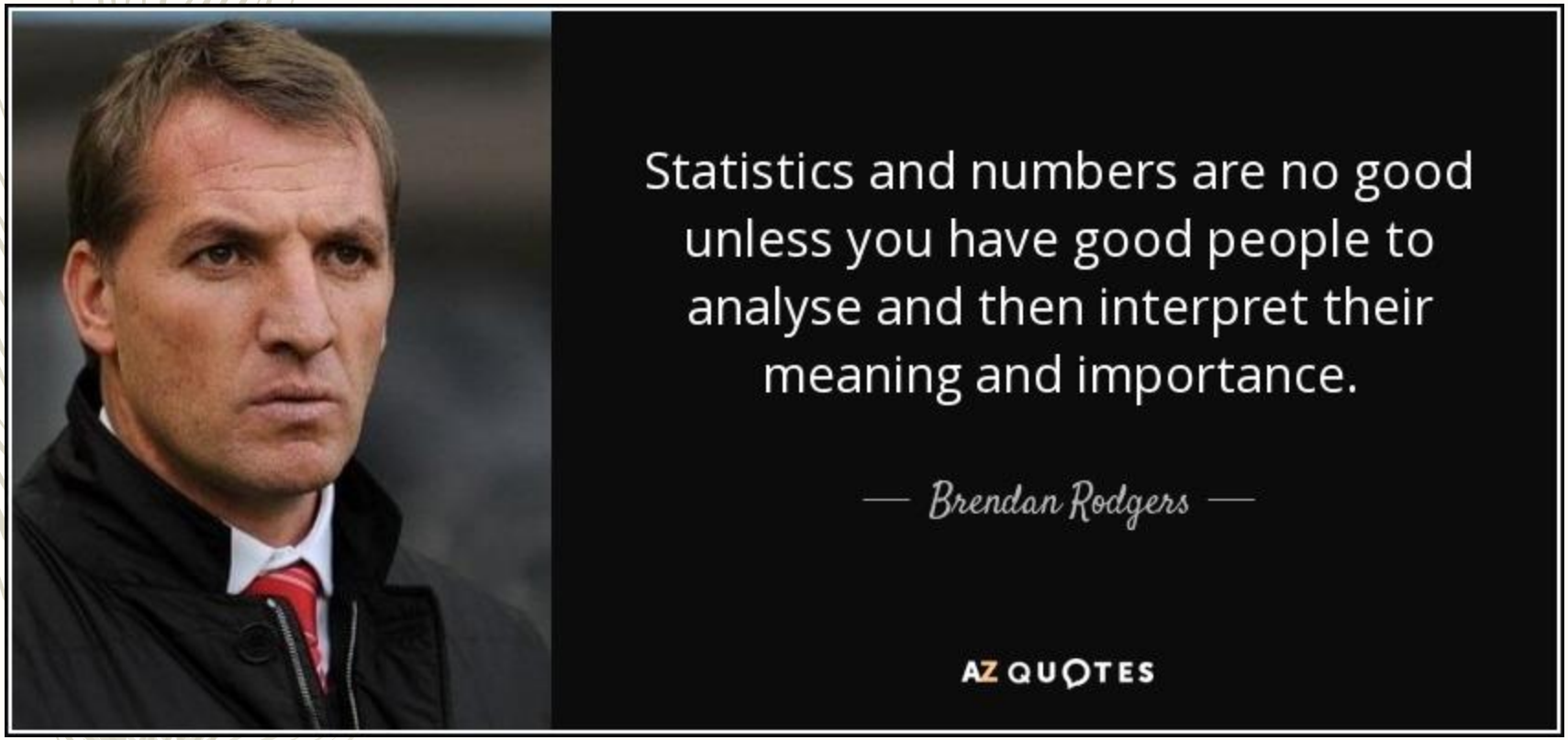
Correlation: 99.26% (r=0.992558)





Remember Slide 6, aula 2 ? – correlações espurias!

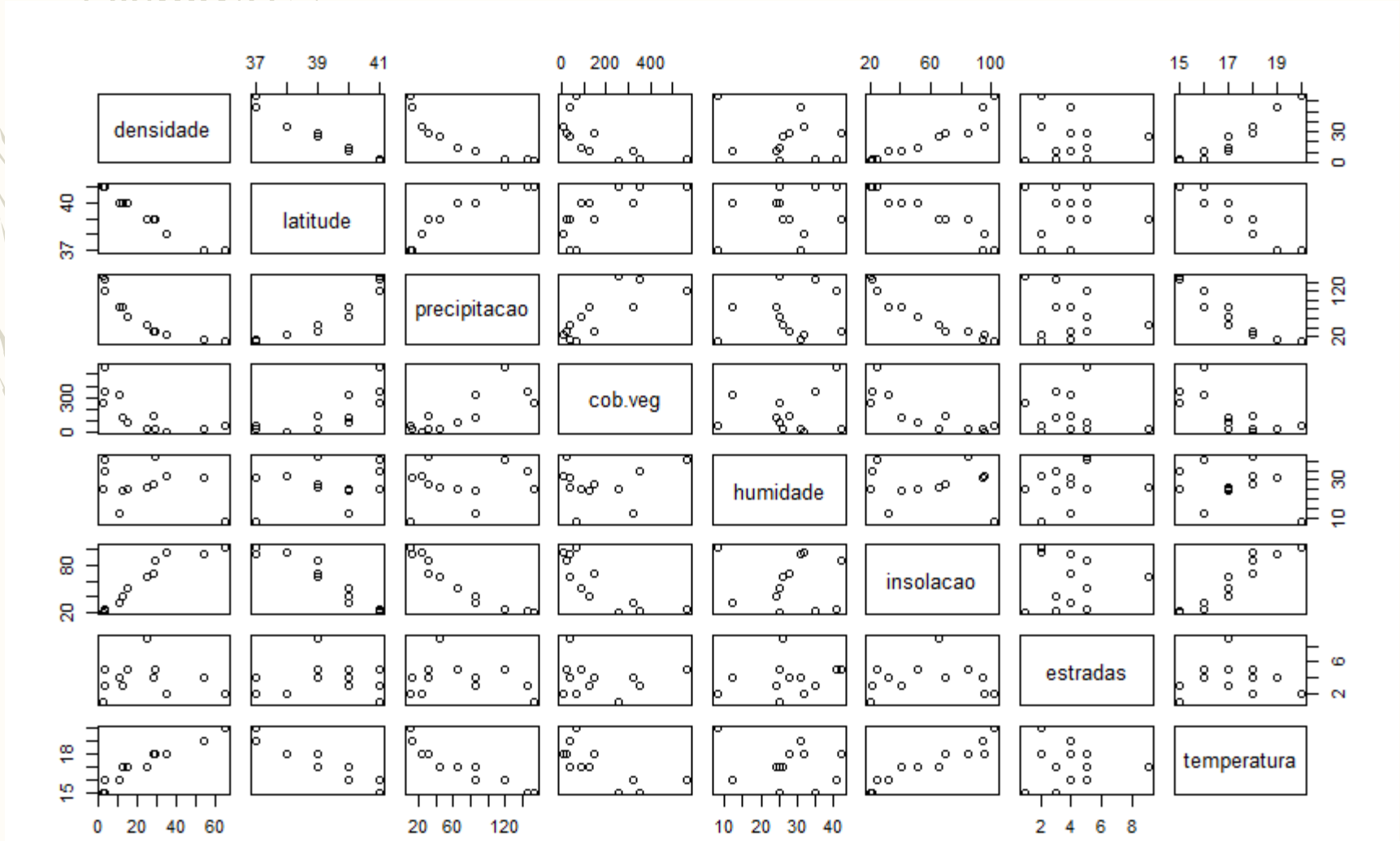
Ecologia Numérica - Aula Teórica 18 – 18-11-2018



https://www.azquotes.com/quote/1304958

regressão e MLG

Antes de começar qualquer exercicio de modelação, explorar as relações entre as variáveis



regressão e MLG

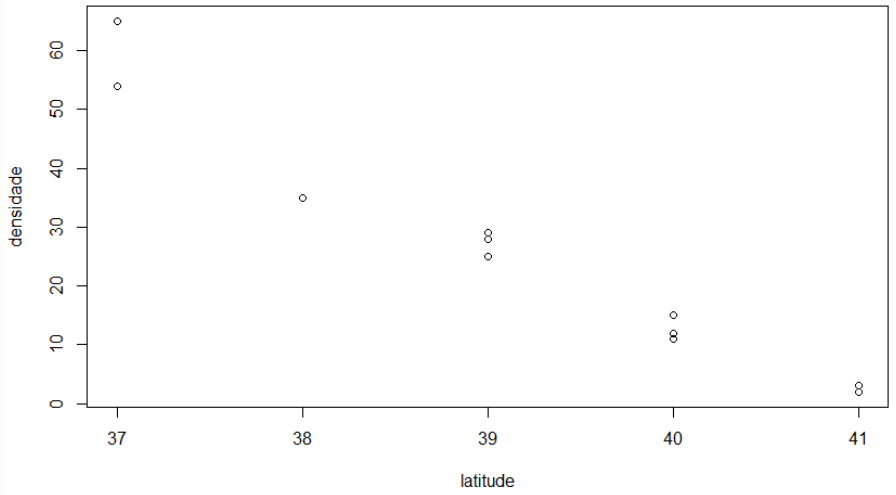
Um modelo de regressão multipla

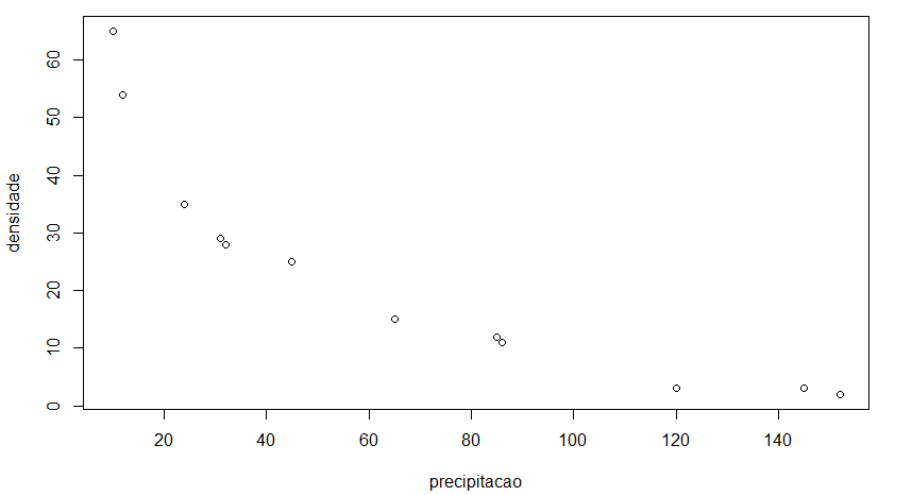
```
Call:
lm(formula = densidade ~ ., data = dens)

Residuals:
      1       2       3       4       5       6       7       8       9      10      11      12
 0.1644  1.9567 -1.6571  0.4803 -2.5667  1.0653 -1.0994  1.7283  1.2474 -2.5452  1.0432  0.1829

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  249.084140 116.048827   2.146   0.0984 .
latitude      -9.617558   2.526905  -3.806   0.0190 *
precipitacao   0.268172   0.081913   3.274   0.0307 *
cob.veg        0.008505   0.009097   0.935   0.4027
humidade      -0.215045   0.122166  -1.760   0.1532
insolacao      0.315824   0.180705   1.748   0.1554
estradas       1.042979   0.572918   1.820   0.1428
temperatura    6.788849   2.134940   3.180   0.0335 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.641 on 4 degrees of freedom
Multiple R-squared:  0.9938,    Adjusted R-squared:  0.983
F-statistic: 91.59 on 7 and 4 DF,  p-value: 0.0002996

> |
```

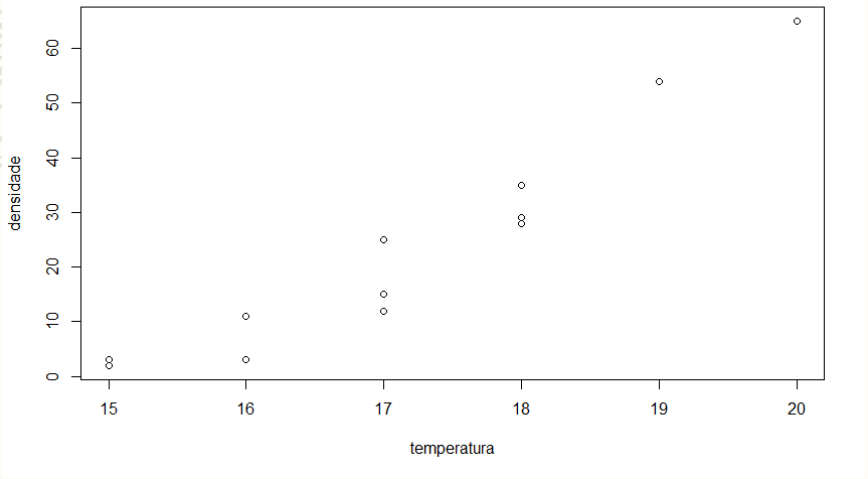
a densidade em função de todas as outras variáveis

Variáveis cujo coeficiente é significativamente diferente de 0, ou seja, que existem evidências de que influenciam a variável resposta

regressão e MLG

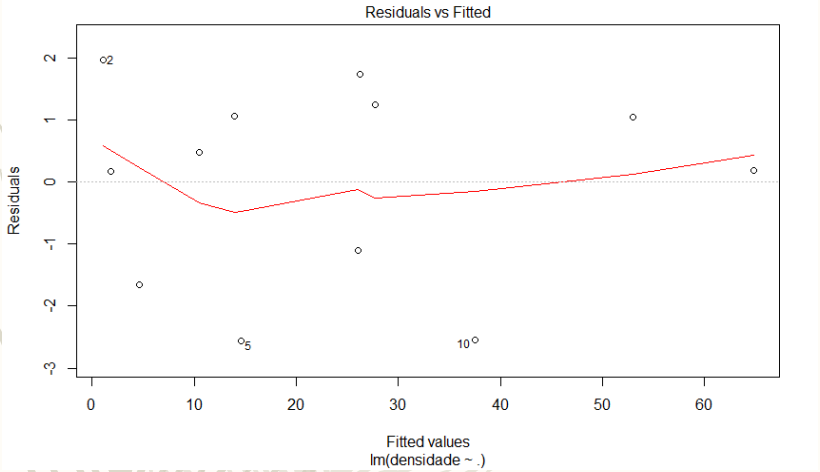


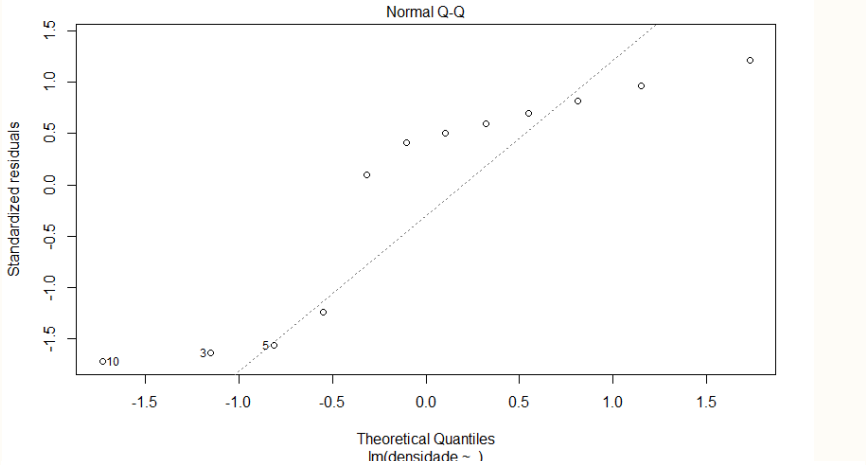


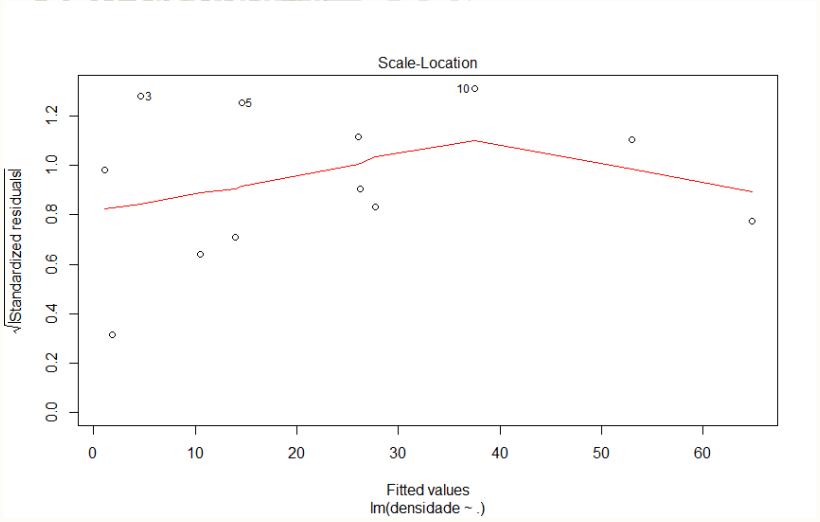


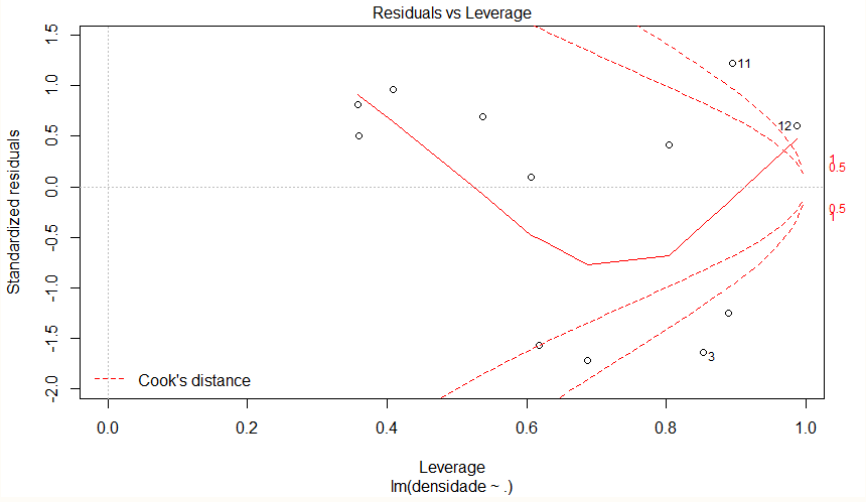
regressão e MLG

Diagnostic plots (obtêm-se fazendo o plot do modelo)

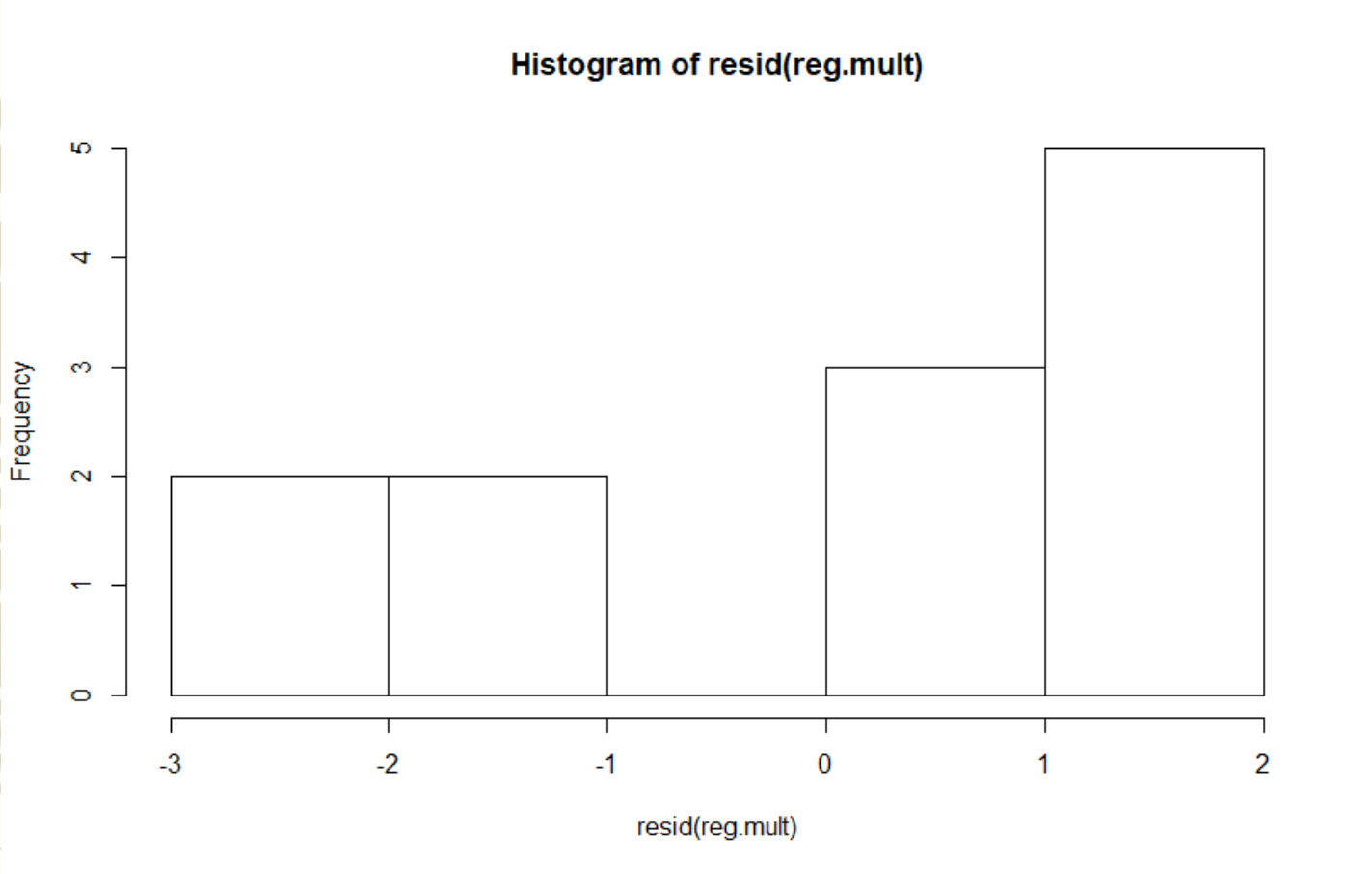








Histograma dos resíduos



A realidade é que, com um tão pequeno número de observações, é muito difícil interpretar os gráficos de diagnóstico – os desvios dos pressupostos teriam de ser gigantes para ser detetáveis. Mas de qualquer forma, comparemos estes valores com os do GLM correspondente.

regressão e MLG

Os mesmos dados, agora com um GLM poisson.

```
Call:
glm(formula = densidade ~ ., family = poisson, data = dens)

Deviance Residuals:
      1       2       3       4       5       6       7       8       9      10      11      12
-0.3845  0.2802  -0.2190  0.3128  0.1774  -0.2165  -0.0648  -0.1795  0.2761  -0.0831  -0.0069  0.0180

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  8.294842   7.845068   1.057   0.2904
latitude    -0.130397   0.156261  -0.834   0.4040
precipitacao -0.017535   0.008546  -2.052   0.0402 *
cob.veg     -0.000872   0.001231  -0.708   0.4787
humidade    -0.006591   0.008258  -0.798   0.4248
insolacao   -0.005802   0.013855  -0.419   0.6754
estradas     0.006784   0.046762   0.145   0.8847
temperatura  0.078208   0.182180   0.429   0.6677
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 194.24605  on 11  degrees of freedom
Residual deviance:   0.57038  on  4  degrees of freedom
AIC: 71.006

Number of Fisher Scoring iterations: 4

>
```

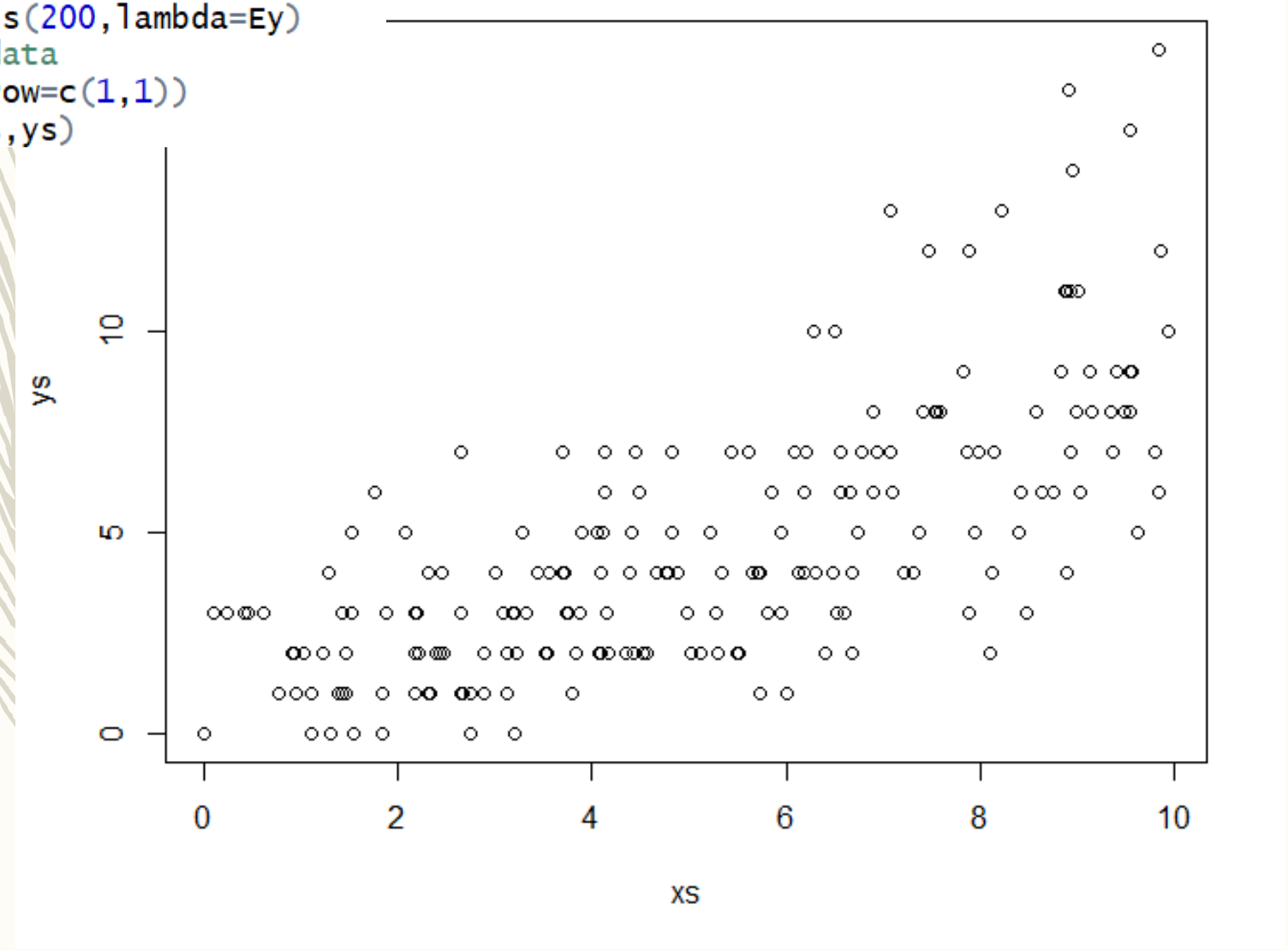
(Null Deviance - Residual Deviance) tem distribuição qui-quadrado com p graus de liberdade.

194.24-0.57 ≈ 193

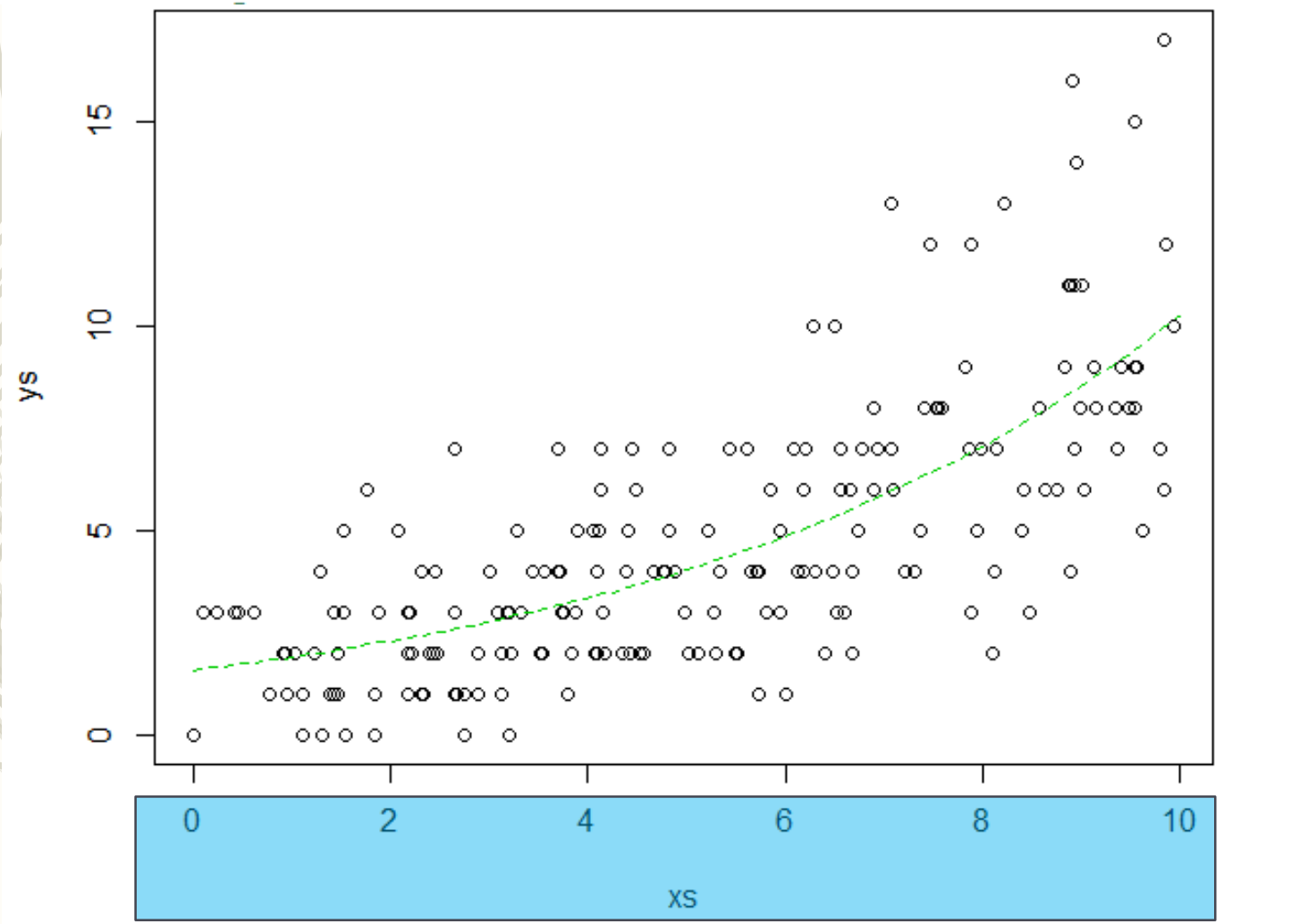
```
> 1-pchisq(193,7)
[1] 0
```

Como ver o modelo GLM estimado?

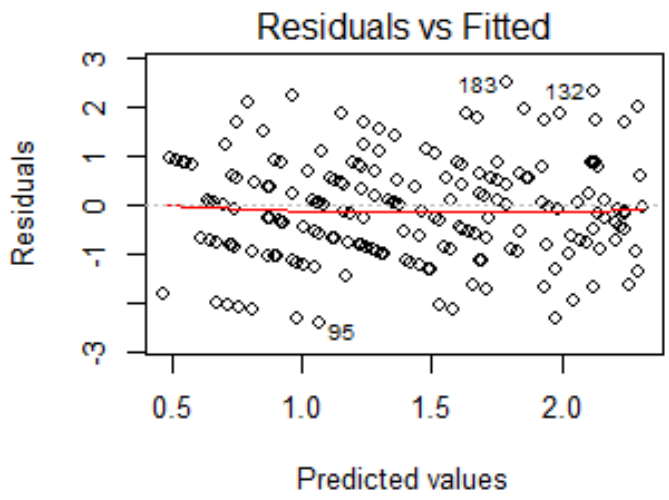
```r
#creating data for a glm
set.seed(123)
#deifine the covariate
xs=runif(200,0,10)
#get the mean value
Ey=exp(0.4+0.2*xs)
#generate response
ys=rpois(200,lambda=Ey)
#plot data
par(mfrow=c(1,1))
plot(xs,ys)
```

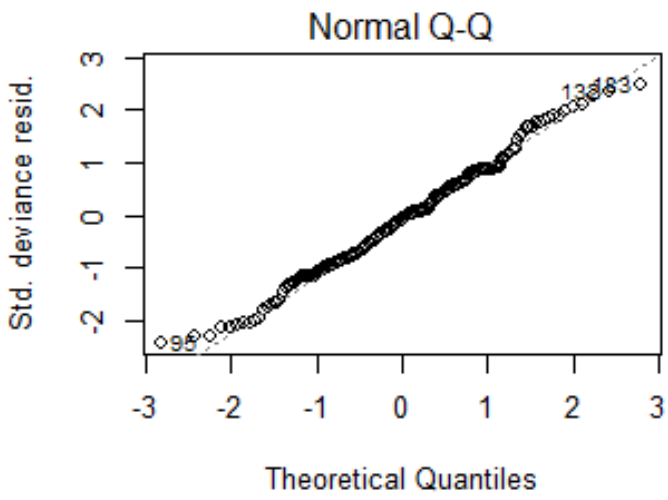


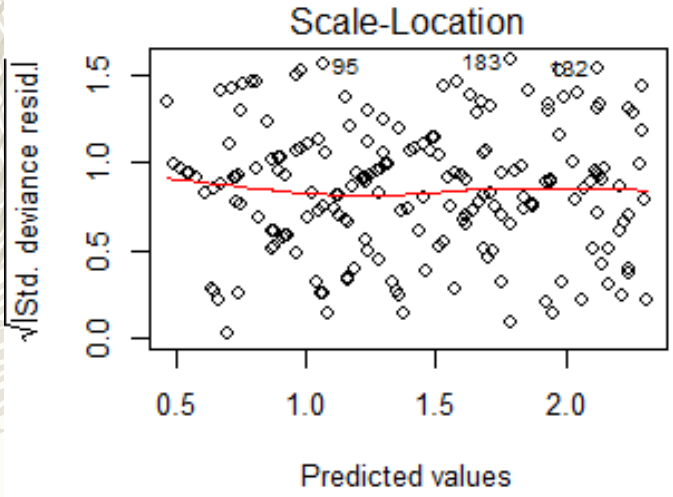
```r
#fit model
glm1=glm(ys~xs,family=poisson(link=log))
#get new data for prediction
newxs=seq(0,10,by=0.1)
#predict
predglm1=predict(glm1,newdata=data.frame(xs=newxs),type="response")
#add fitted model
plot(xs,ys)
lines(newxs,predglm1,lty=2,col=3)
```

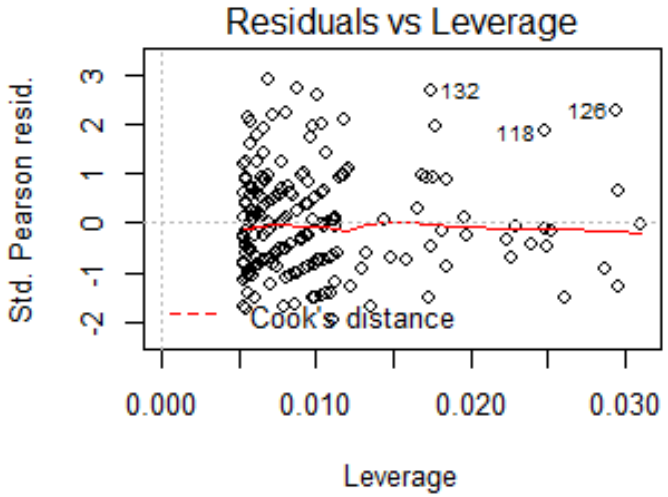


```r
#diagnostics plot
par(mfrow=c(2,2))
plot(glm1)
```

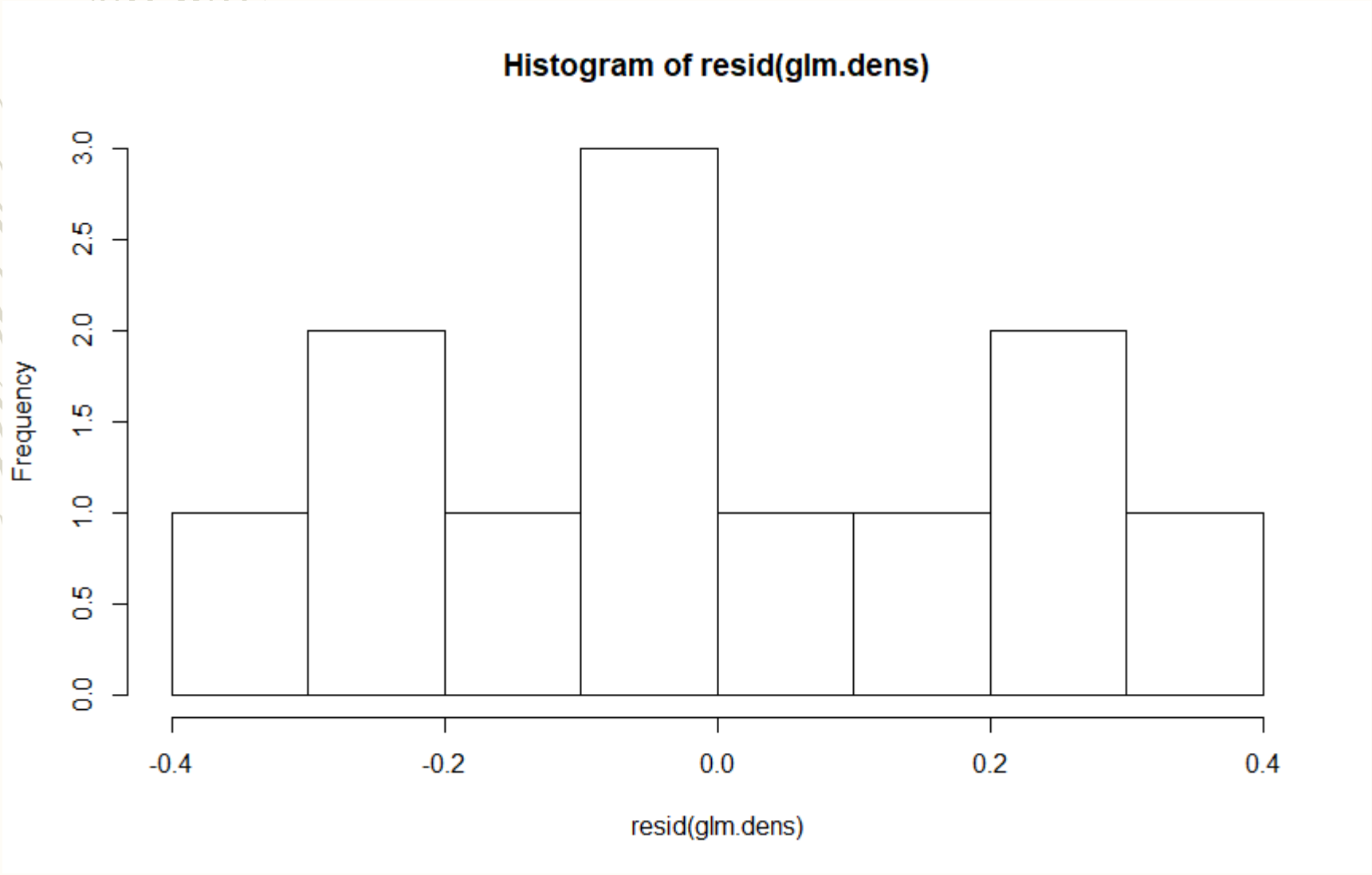








regressão e MLG



Modelos aditivos generalizados

Quando a relação entre a variável resposta e as covariáveis não for linear, podemos precisar de mais do que um GLM. Surgem assim os GAMs

CHAPTER 3

Introducing GAMs

Tal como nos GLMs

Função de ligação (*link function*)

Uma parte não suave

3.1 Introduction

A generalized additive model (Hastie and Tibshirani, 1986, 1990) is a generalized linear model with a linear predictor involving a sum of smooth functions of covariates. In general the model has a structure something like

$$g(\mu_i) = \mathbf{X}_i^* \boldsymbol{\theta} + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots \quad (3.1)$$

where

$$\mu_i \equiv \mathbb{E}(Y_i) \text{ and } Y_i \sim \text{ some exponential family distribution.}$$

Uma parte com funções suaves (smooth) das covariáveis

Wood, S. N. 2006 Generalized Additive Models: an introduction with R. CRC/Chapman & Hall

Sem entrar nos detalhes computacionais, que são extremamente complexos, o que um GAM faz é encontrar uma função suave que se ajusta aos dados, mas penalizando a complexidade do modelo.

As funções suaves mais comuns usadas na prática em GAMs são os splines (usadas por exemplo no package `mgcv`, função `gam`, que nós vamos usar para os implementar). Consistem num conjunto de funções base simples (ver figura abaixo) que multiplicadas por constantes e somadas conseguem reproduzir uma função suave.

O objectivo é encontrar as constantes (ou seja, os parâmetros) que resultam na melhor aproximação aos dados, minimizando o número de parâmetros!

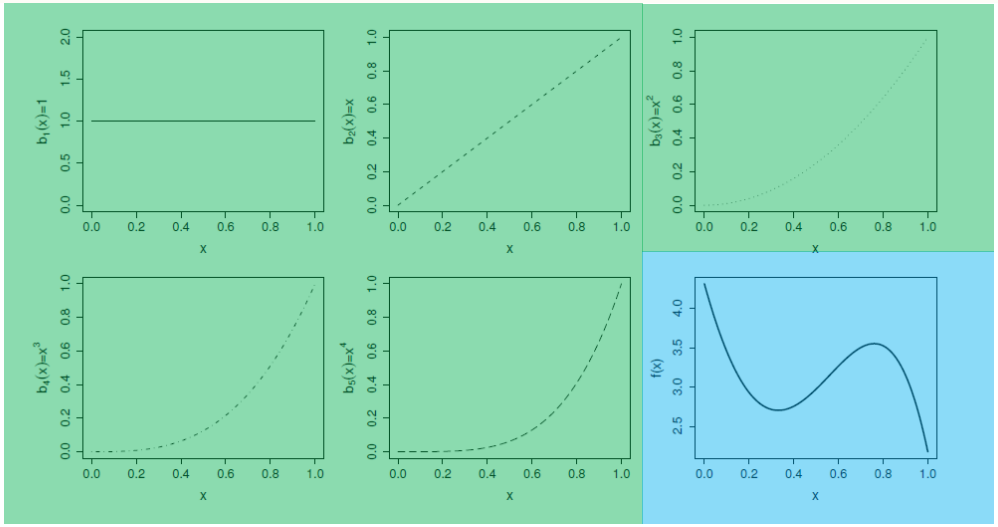


Figure 3.1 *Illustration of the idea of representing a function in terms of basis functions, using a polynomial basis. The first 5 panels (starting from top left), illustrate the 5 basis functions, $b_j(x)$, for a 4th order polynomial basis. The basis functions are each multiplied by a real valued parameter, β_j, and are then summed to give the final curve $f(x)$, an example of which is shown in the bottom right panel. By varying the β_j, we can vary the form of $f(x)$, to produce any polynomial function of order 4 or lower. See also figure 3.2*

Wood, S. N. 2006 Generalized Additive Models: an introduction with R. CRC/Chapman & Hall

Relação entre LM, GLM e GAM

Linear Model

$$E[y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Additive Model (1985)

$$E[y|\mathbf{x}] = \beta_0 + f_1(x_1) + \ldots + f_p(x_p)$$

Generalized Linear Model (1972)

$$g\left(E[y|\mathbf{x}]\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Generalized Additive Model (1990)

$$g\left(E[y|\mathbf{x}]\right) = \beta_0 + f_1(x_1) + \ldots + f_p(x_p)$$

parametric nonparametric

Lembrete: no primeiro tutorial sobre o R havia uma comparação simples entre um LM, um GLM e um GAM!

Relação entre LM, GLM e GAM

As principais diferenças entre modelos de regressão mais simples (LM) e os modelos lineares generalizados (GLM) ou modelos aditivos generalizados (GAM) são que os 2 últimos:

1. permitem contemplar distribuições não normais dos erros;

2. a relação entre a variável resposta e as variáveis independentes não precisa de ser linear (num GLM a relação é ainda linear, mas na escala da função de ligação)

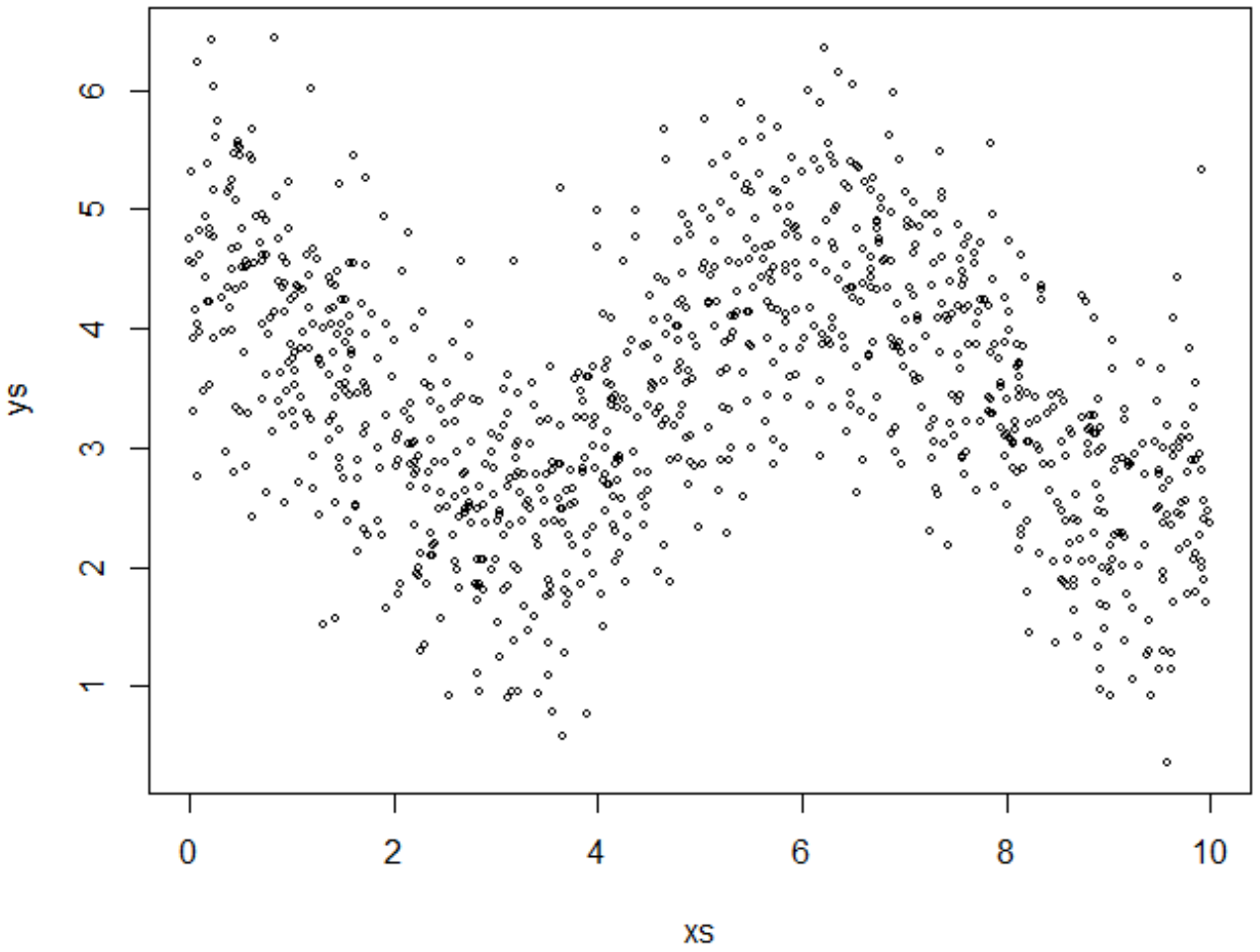
Assim, os GLM ajustam-se melhor a uma maior variedade de dados, e.g.

- contagens,
- proporções,
- presença/ausência,
- valores contínuos estritamente positivos,
- valores entre 0 e 1
- etc.

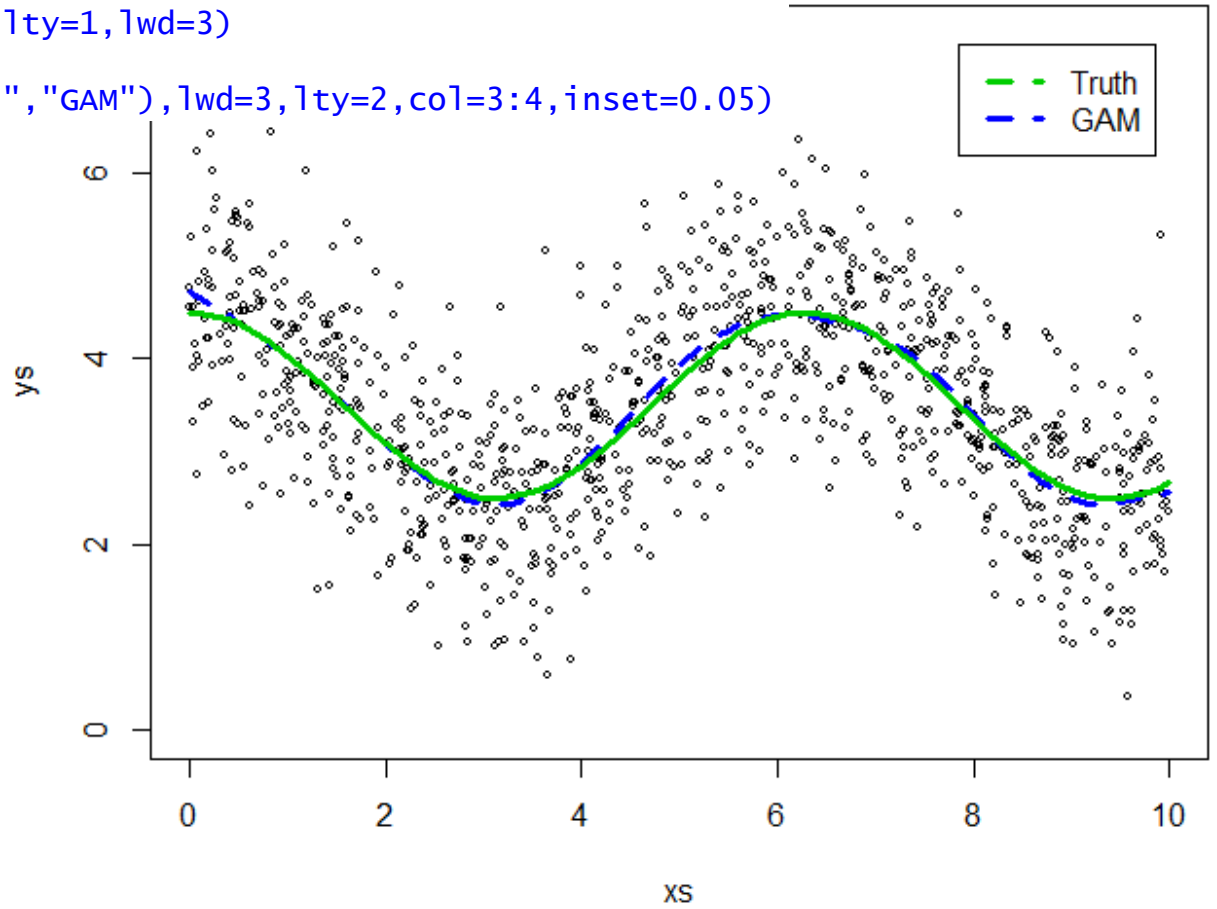
Os GAM fazem tudo isso mas também podem lidar com relações não lineares.

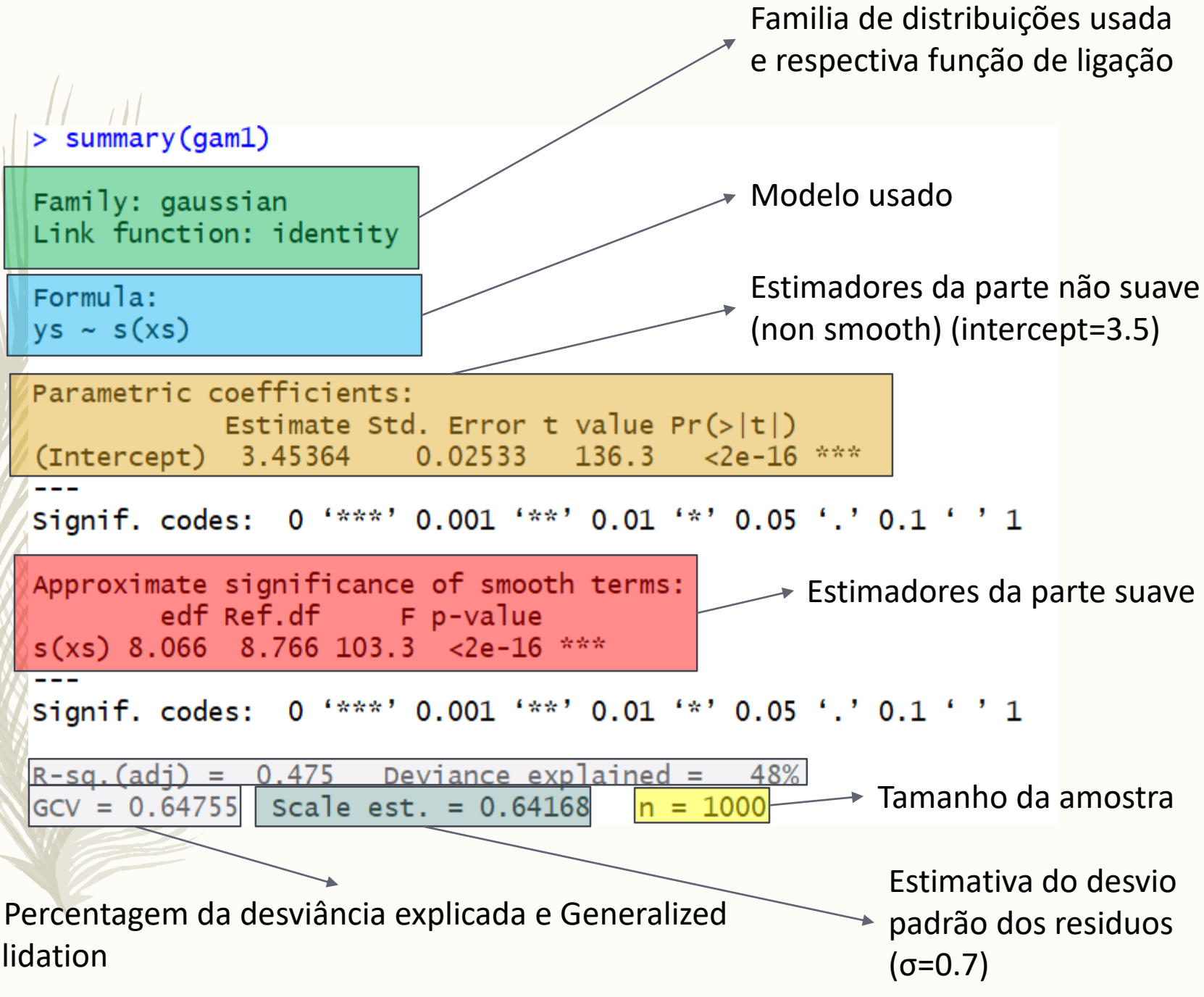
```
set.seed(123)
xs=runif(1000,0,10)
ys=3.5+cos(xs)+rnorm(1000,sd=0.8)
par(mfrow=c(1,1))
plot(xs,ys,cex=0.5)
```

Clearly non-linear relation, no LM or GLM will fit the data



```
library(mgcv)
#fit the model
gam1=gam(ys~s(xs),family=gaussian)
#get a reasonable range of xs for predictions
xpred=seq(0,10,by=0.1)
#predict, on the response scale
preds=predict(gam1,newdata=data.frame(xs=xpred),type="response")
par(mfrow=c(1,1))
#plot the data
plot(xs,ys,cex=0.5,ylim=c(0,7.5))
#add the predictions
lines(xpred,preds,col=4,lty=2,lwd=3)
#add the tue model
lines(xpred,3.5+cos(xpred),col=3,lty=1,lwd=3)
#add a legend
legend("topright",legend=c("Truth","GAM"),lwd=3,lty=2,col=3:4,inset=0.05)
```




> summary(gam1)
Family: gaussian
Link function: identity
Formula:
ys ~ s(xs)
Parametric coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.45364 0.02533 136.3 <2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Approximate significance of smooth terms:
edf Ref.df F p-value
s(xs) 8.066 8.766 103.3 <2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R-sq.(adj) = 0.475 Deviance explained = 48%
GCV = 0.64755 Scale est. = 0.64168 n = 1000
Familia de distribuições usada e respectiva função de ligação
Modelo usado
Estimadores da parte não suave (non smooth) (intercept=3.5)
Estimadores da parte suave
Tamanho da amostra
GOF: R2, Percentagem da desviância explicada e Generalized Cross Validation
Estimativa do desvio padrão dos residuos (σ=0.7)

Para mim, a mais importante classe de modelos para quem trabalha em ecologia são os modelos de regressão (e dentro destes, GLMs e GAMs).

I would call these the cornerstone of statistical ecology.