

# Modelação Ecológica

## AULA 9

15 November 2019 – 14:00-16:30 – room 2.3.37

Tiago A. Marques

# Material to work over

Getting more used to R

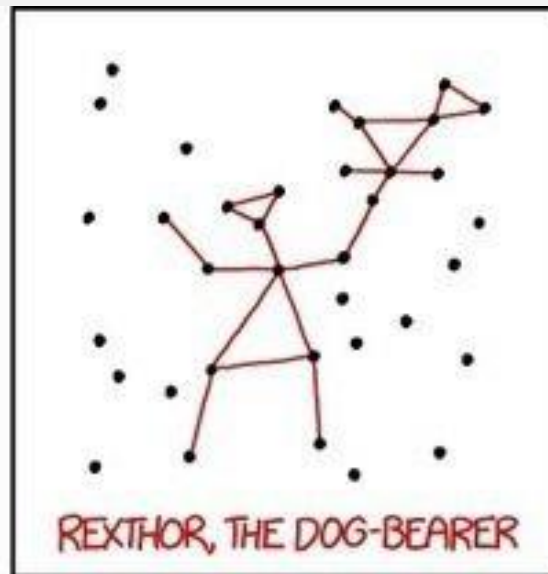
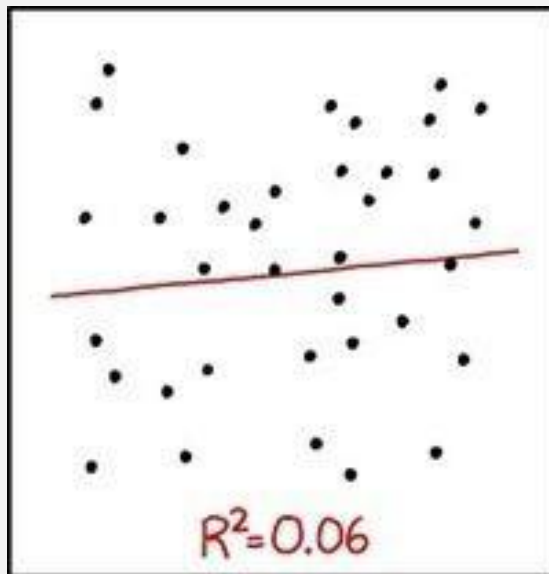
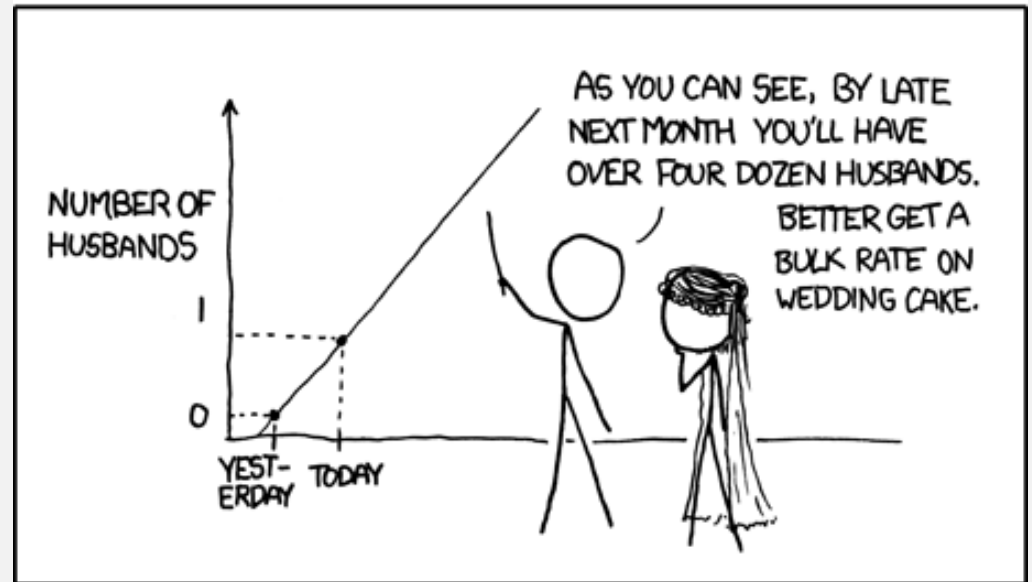
The screenshot shows a course management interface titled "Gestão de Páginas". On the left, a file tree is displayed under "Modelação Ecológica". The tree includes folders for "Aulas" (Aula1 to Aula9) and "Outros Recursos". Files are highlighted with colored boxes: "An extra set of exercises to practice R - (stolen from numerical)" (blue), "A task about merge" (blue), "FT1" and "FT2" (green), and "On Regression Models" (orange). On the right, the "On Regression Models" page is shown with a "Ficheiros" tab containing 4 files. Below the tab is a table listing the files.

#	Nome	Permissões
1	data.csv	Público
2	OnRegMods.Rmd	Público
3	OnRegMods.pdf	Público
4	OnRegMods.html	Público

On stats and analysis in R

On regression models

# MY HOBBY: EXTRAPOLATING



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

# ASSESSMENT



**MECOCO**

Component 20%

# MECOCO 20% - HAND IN 20<sup>TH</sup> DECEMBER 2019

## Gestão de Páginas

- ▼ Modelação Ecológica
  - Modelação Ecológica(Ecologia Marinha)
  - Modelação Ecológica(Ecologia e Gestão Ambiental)
- ▶ Aulas
- ▶ Outros Recursos
- ▼ Avaliação
  - MECOCO 20 % individual

+ Criar

## MECOCO 20 % individual

Página Ficheiros **4** Permissões Link

Adicionar Ficheiro

#	Nome
1	Assignment instructions Mecoco4ME.pdf
2	islands.bt
3	birds.bt
4	plants.bt

You are supposed to answer the following question:

What is/are the primary determinant(s) of alien species richness for birds / plants on oceanic islands:

- a) area per se,
- b) native species richness or
- c) human factors?

To do so you are given 3 data files:

Islands – a dataset with the characteristics of a set of islands for which there were measured the number of exotic species of birds and plants

Plants – the dataset with the number of exotic species and native of plants

Birds - the dataset with the number of exotic species and native species of birds



The goal is to find a suitable model that allows you to answer the above question. You should provide me a detailed report describing the entire process, from data import, through exploratory data analysis, modelling approach(es) and the answer to the above question, along with a discussion and final conclusions.

This is to be handed in by the 20<sup>th</sup> December 2019. As agreed upfront, this is individual work and it is worth 20% of your grade in Modelação Ecológica.





**Please remember, it is fundamental that you do not discuss with each other what you are doing. If you get stuck, you come and talk to me, not amongst yourselves. The entire goal is to see the range of models and modelling approaches that you use to address the question.**

**If you are found to not have respected this condition, I reserve the right to fail you in ME – this is a serious warning that you should not ignore nor take lightly.**



**"It doesn't *matter* that you never got caught!"**

# THEORETICAL WORK

Component 20%

# THE TASK

- Choose an R package, explore it, find a suitable dataset and implement an analysis with it, taking your conclusions about it
- Write a full report with the theoretical background on the type of model chosen, what it can be used for and how it has been used
- Choose a paper that has used said model/R package to anchor your discussion
- Discuss the merits of the model, and other existing implementations of similar models
- Examples of recommended packages (feel free to choose one from these!):
  - Distance, mrds, dsm, mrds – distance sampling
  - moveHMM, momentum – animal movement
  - lme4, pscl, nlme, rpart, randomForest, INLA – more complicated regression models
  - unmarked - distance sampling and occupancy analysis
  - RMark, secr, marked - capture recapture analysis
  - dismo, sdm - species distribution modelling, e.g. maxent for presence only data
  - rethinking, great – Bayesian models

## TIMELINE AND KEY GUIDELINES

By the 15<sup>th</sup> November 2019 – tell me which package and which paper you are using

Hand in 5 page max report 3<sup>rd</sup> December 2019

Presentation (10 mins) in class same day: every member of each group **MUST** talk

## CRAN Task View: Analysis of Ecological and Environmental Data

**Maintainer:** Gavin Simpson

**Contact:** ucfagls at gmail.com

**Version:** 2019-09-01

**URL:** <https://CRAN.R-project.org/view=Environmetrics>

### Introduction

This Task View contains information about using R to analyse ecological and environmental data.

The base version of R ships with a wide range of functions for use within the field of environmetrics. This functionality is complemented by a plethora of packages available via CRAN, which provide specialist methods such as ordination & cluster analysis techniques. A brief overview of the available packages is provided in this Task View, grouped by topic or type of analysis. As a testament to the popularity of R for the analysis of environmental and ecological data, a [special volume](#) of the *Journal of Statistical Software* was produced in 2007.

Those useRs interested in environmetrics should consult the [Spatial](#) view. Complementary information is also available in the [Multivariate](#), [Phylogenetics](#), [Cluster](#), and [SpatioTemporal](#) task views.

If you have any comments or suggestions for additions or improvements, then please contact the [maintainer](#) .

A list of available packages and functions is presented below, grouped by analysis type.

### General packages

These packages are general, having wide applicability to the environmetrics field.

- Package [EnvStats](#) is the successor to the S-PLUS module *EnvironmentalStats* , both by Steven Millard. A [user guide in the form of a book](#) has recently be released.

<https://cran.r-project.org/web/views/Environmetrics.html>

# INTERESTING PLACE WITH RESOURCES AND R CODE

<http://www.seec.uct.ac.za/stats-toolbox-seminars>

Species Distribution Modelling  
Occupancy models  
Distance sampling  
Handling spatial data  
Experimental and survey design  
Introduction to multivariate analyses  
Spatial capture-recapture (SCR) modelling  
Animal movement modelling with moveHMM  
Classification and regression trees  
Spatial occupancy models  
Hidden Markov Models for time series  
R Tidyverse  
Generalised Linear Mixed Models  
Single-season occupancy models using a Bayesian approach  
SDMs - using spatial information to supplement biased occurrence data  
Handling Spatial Data  
Generalised additive models (GAMs)  
Bayesian analysis  
Data exploration  
ggplot2 - the grammar of graphics  
Time series analysis  
Meta-analysis  
Cloud computing with R

## Stats Toolbox Seminars

Want to broaden your stats knowledge? Unsure of what you can do with your data? Still developing your proposal?

Learn about stats methods that can be applied to ecological and environmental data. On the last Thursday of every month at 13:00 we will introduce a method, e.g. distance sampling or occupancy models, that can be added to your stats toolbox.

**Click on the links in the left side menu for presentation slides and R scripts from previous Stats Toolbox Seminars.**

See below for details of our Stats Toolbox Seminars for this year.




## SEEC STATS TOOLBOX 2019 SCHEDULE

Date	Topic	Speaker
28 February	The grammar of graphics – ggplot2	Kirsten Packer
28 March	Population status assessment tools	Henning Winker
25 April	Time series analyses	Birgit Erni
30 May	Meta-analysis	Vernon Visser
25 July	Cloud computing and R with Amazon Web Services (AWS)	Ian Durbach

REVIEW |  Full Access |

## Navigating through the R packages for movement

Rocio Joo , Matthew E. Boone, Thomas A. Clay, Samantha C Patrick, Susana Clusella-Trullas, Mathieu Basille

First published: 06 October 2019 | <https://doi.org/10.1111/1365-2656.13116>

**Texto Integral @ b-on**

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi:10.1111/1365-2656.13116



PDF



TOOLS



SHARE

# PRACTICAL WORK

Component 45%



Read the guidelines...!

Pay attention to it



(ler o guião com muita atenção!)

Entrega: 31<sup>th</sup> January 2019

## Modelação Ecológica

2019/2020

### Guião para a elaboração do trabalho prático

#### Objectivo

O trabalho prático visa aplicar os conhecimentos das sessões teórico-práticas da disciplina a um estudo de caso, inovador em que seja adequado utilizar modelação ecológica.

#### Autorias

O trabalho poderá ser realizado em grupos de até 4 alunos

#### Formato, estrutura e conteúdo:

O trabalho deverá ser baseado num relatório dinâmico.

O trabalho deverá ter 4 secções principais, as quais deverão corresponder grosseiramente a 4 páginas (o trabalho não deverá exceder as 6 páginas, havendo penalização caso seja incluída informação não relevante). Deverá ser estruturado da seguinte forma:

#### Título

#### Autorias

#### Enquadramento e objetivo (1 pág.)

Informação necessária à compreensão da problemática em questão. Dadas as limitações à extensão do trabalho, deverá haver poder e capacidade de síntese e ser dado ênfase no domínio da modelação e não tanto na questão ecológica (i.e. para o problema proposto, quais os tipos de modelos que têm sido utilizados, e quais os resultados anteriores). Indicação clara de quais os objetivos do trabalho

#### Métodos (1 pág.)

Apresentação do modelo com o devido enquadramento em termos de referências bibliográficas. Deve seguir-se as diferentes etapas comuns ao desenvolvimento de modelos e apresentado um esquema conceptual.

#### Resultados (1 pág.)

Apresentação dos principais resultados da modelação, usando como suporte preferencial tabelas e gráficos e/ou resultados de análises estatísticas, consoante o caso. Podem ser apresentados os diferentes cenários e respetivas previsões, se adequado.

#### Discussão (1 pág.)

Interpretação e comparação dos resultados com outros trabalhos similares, privilegiando a componente de modelação à ecológica. Crítica metodológica e perspetivas de desenvolvimento futuro.

#### Referências bibliográficas (não contabilizadas nas 4 páginas)

Listagem das referências citadas no texto seguindo a formatação de uma revista de referência na área da disciplina.

Podem apresentar anexos, por exemplo código ou resultados intermédios não necessários no documento principal.

#### Prazo de entrega:

A definir hoje!

#### Modo de entrega:

Ficheiro em formato “pdf” e respetivo “.Rmd”, bem como os dados, enviado por e-mail para [tiago@fc.ul.pt](mailto:tiago@fc.ul.pt) (até às 24h da data de entrega).

#### Apresentação oral do trabalho:

As apresentações serão no dia ????? de ????, às ????

A apresentação deverá ter no máximo 10 minutos e seguir a estrutura do trabalho escrito.

#### Critérios de avaliação:

Tema do trabalho e caso de estudo em particular, carácter inovador e nível de dificuldade

Qualidade do enquadramento efetuado

Rigor na seleção e aplicação dos métodos

Forma de apresentação dos resultados – critério na escolha do conteúdo e forma quanto a esta secção

Riqueza bibliográfica, qualidade da interpretação e autocrítica demonstrada na Discussão

Relevância do número e tipo de publicações citadas

Apresentação oral do trabalho

# BACK TO BUSINESS

The concept of maximum likelihood... and all that !

# The concept of maximum likelihood

and maximum likelihood estimator (MLE)

Given a model, we can calculate the probability of the data:  $P(\text{data}|\text{model},\theta)$

Given a model and the data, we can evaluate what are the parameter values which are more likely:  $P(\theta|\text{data},\text{model})$  – these are maximum likelihood estimators (MLE)

In reality, much of what we do in statistics is based on likelihoods and the corresponding maximum likelihood estimators!

For a linear model, the minimum squares estimates of  $a$  (the intercept) and of  $b$  (the slope) are MLEs!

The sample mean, the estimator of the mean in a Gaussian or Poisson is a MLE

etc...

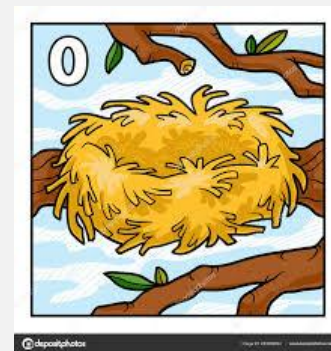
For almost all of the statistical models you know, there will be an underlying likelihood!

Imagine the following situation: a biologist goes is interested in the reproductive success of a species of bird. In particular, he is interested in estimating the probability of a couple of birds nesting actually laying eggs. To do he looks for nests and the records whether there are eggs in the nest.

Define  $X$  as the variable representing the presence of eggs in the nest.  $X$  takes the value 1 if there are eggs in a nest and 0 otherwise. Therefore:

$X=1$  with probability  $\theta$

$X=0$  with probability  $1-\theta$



Note that  $\theta$  must be a value between 0 and 1, and that  $\theta$  is the parameter for the model that we assume for  $X$ . Our objective is to estimate  $\theta$ .

The biologist collects the data from 5 nests, and records

$$\underline{X}=(1,0,1,0,0)$$

$$P(X=1) = \theta$$

$$P(X=0) = 1-\theta$$

What is the probability of observing this sample (often forgotten but key, nests are independent!)

$$\begin{aligned} P(\underline{X}) &= P(X=1) \times P(X=0) \times P(X=1) \times P(X=0) \times P(X=0) = \\ &= \theta (1-\theta) \theta (1-\theta)(1-\theta) \\ &= \theta^2 (1-\theta)^3 \end{aligned}$$

Note that  $P(\underline{X})$  can be seen either as a function of the data, conditional on  $\theta$ ,  $P(\underline{X}|\theta)$ , or a function of  $\theta$ , conditional on the data,  $P(\theta|\underline{X})$ . The former is what we call the likelihood.

Imagine that you know  $\theta = 0.3$    $P(X)=0.3^2 \times 0.7^3=0.03087$

Let's recall, we have a sample  $\underline{X}=(1,0,1,0,0)$ , and define  $n_1$  as the number of 1's and  $n_0$  the number of 0's in our sample, we have a likelihood function given by

$$L(\theta | X) = \theta^{n_1} (1 - \theta)^{n_0}$$

What is the value of  $\theta$  that maximizes the likelihood function?

Using a grid search approach

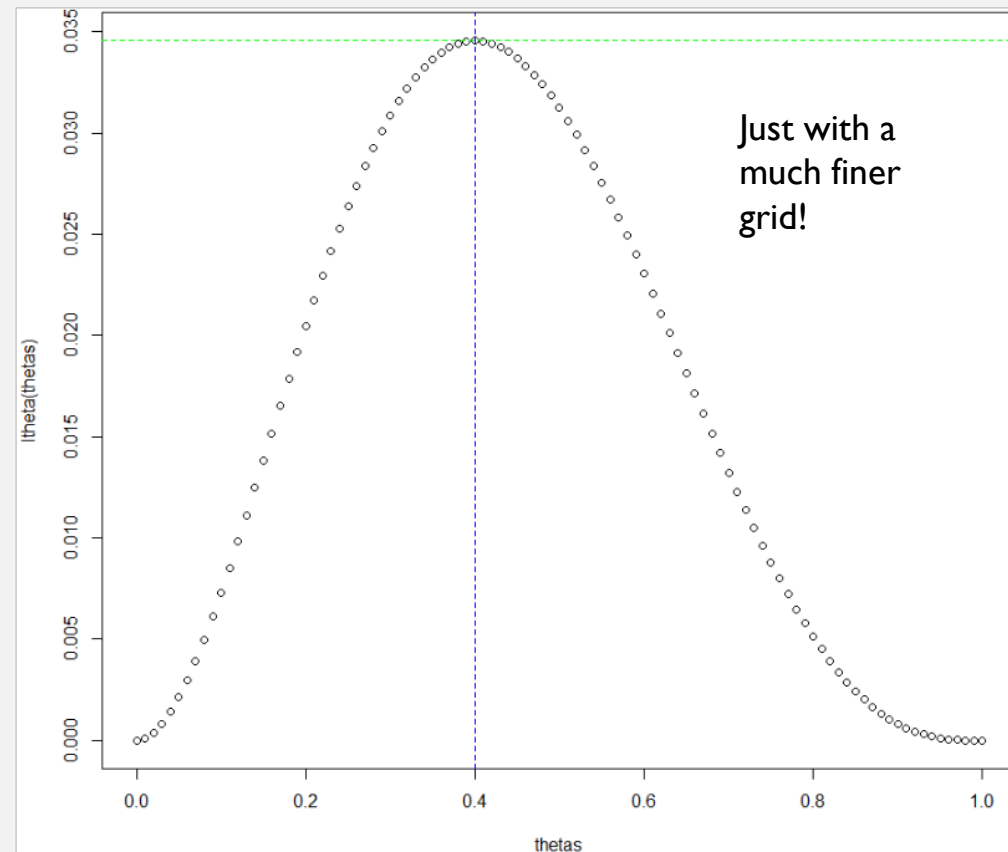


	$\theta$	$L(\theta   \underline{X})$
1	0.05	0.0021
2	0.15	0.0138
3	0.25	0.0264
4	0.35	0.0336
5	0.45	0.0337
6	0.55	0.0276
7	0.65	0.0181
8	0.75	0.0088
9	0.85	0.0024
10	0.95	0.0001

```
ltheta=function(theta,nl=2,n0=3){  
  lik=(theta)^nl*(1-theta)^n0  
  return(lik)}
```

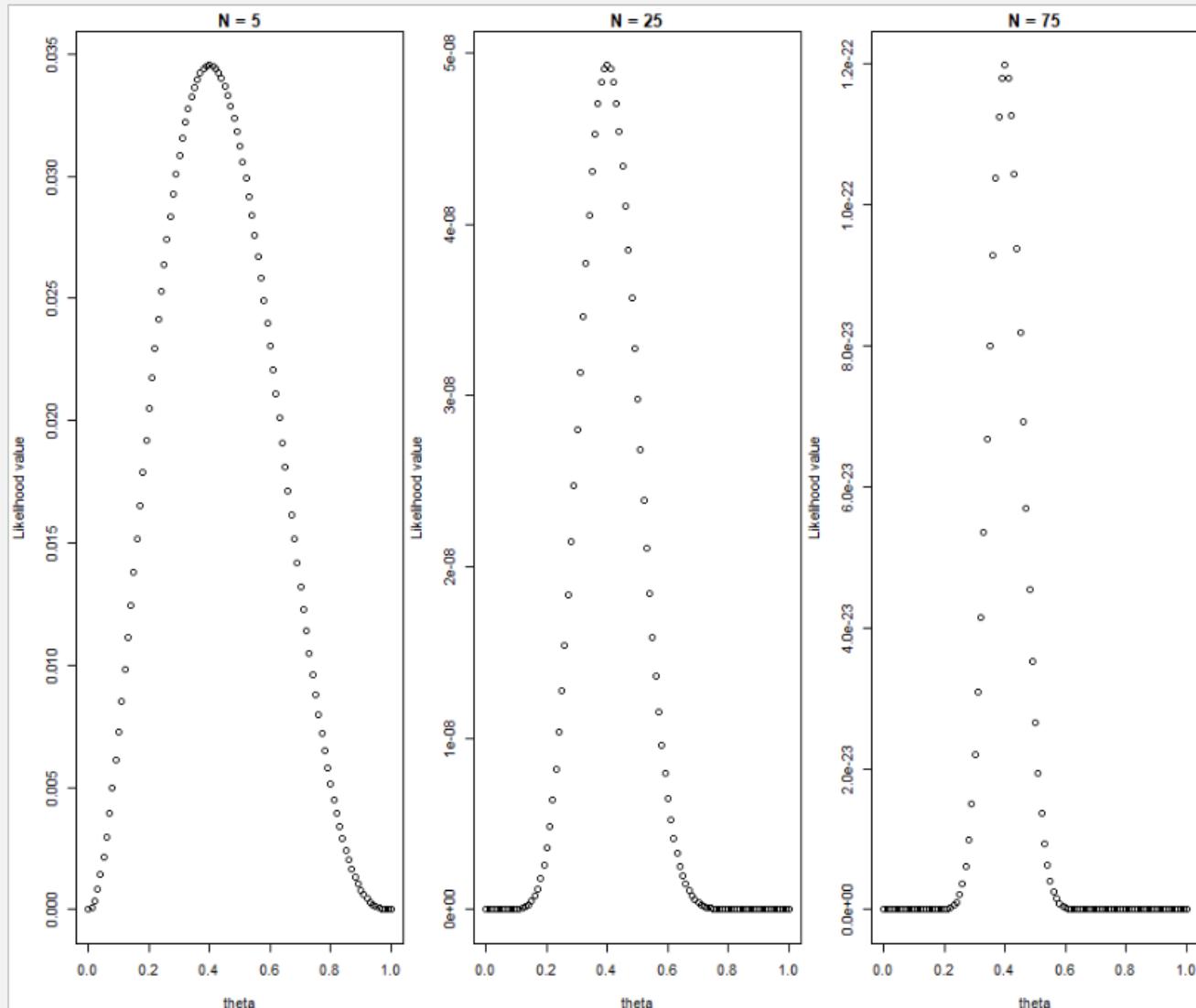
```
par(mfrow=c(1,1),mar=c(4,4,0.2,0.2))  
thetas=seq(0,1,by=0.01)  
plot(thetas,ltheta(thetas),ylab="Likelihood value",xlab="theta")
```

```
#index of the maximum  
maxind=ltheta(thetas)==max(ltheta(thetas))  
#plot the maximum of the function  
abline(h=ltheta(thetas)[maxind],lty=2,col=3)  
#plot the theta that maximizes the function  
abline(v=thetas[maxind],lty=2,col=4)
```





If we have more data, we can estimate the maximum likelihood parameters with higher precision (note the likelihood profile!)



In practice, we do not use trial and error but numerical procedures to maximize the likelihood!

## Code for previous plot!

```
par(mfrow=c(1,3),mar=c(4,4,1.5,0.2))
plot(thetas,ltheta(thetas),ylab="Likelihood
value",xlab="theta",main="N = 5")
plot(thetas,ltheta(thetas,n1=10,n0=15),ylab="Likelihood
value",xlab="theta",main="N = 25")
plot(thetas,ltheta(thetas,n1=10*3,n0=15*3),ylab="Likelihood
value",xlab="theta",main="N = 75")
```

Exemplos  
de  
maximização numérica  
de uma  
verosimilhança

# FOR THE BIRD NEST EXAMPLE



In R, for a function of a single parameter, we can use function `optimize`

Function `optim` is suited for multiple parameters

`nlm` (from package `stats`) can also be used

```
liktheta=function(theta,data){  
  loglik=sum(log(theta^sum(data==1))+sum(log((1-theta)^sum(data==0))))  
  return(loglik)  
}
```

```
optimize(liktheta,interval=c(0.01,0.99),data=c(0,1,0,1,0),maximum=TRUE)
```

By default `optimize`  
minimizes functions

Function to maximize  
over the parameter space

Range of values to look over (i.e.  
parameter space)

The data

```
> optimize(liktheta,interval=c(0.01,0.99),data=c(0,1,0,1,0),maximum=TRUE)  
$maximum  
[1] 0.399996  
  
$objective  
[1] -3.365058
```

Challenge: estimate the maximum likelihood estimate if the data was 25 presences and 34 absences

```
newdata=c(rep(1,25),rep(0,34))  
optimize(liktheta,interval=c(0.01,0.99),data= newdata,maximum=TRUE)
```

```
> newdata=c(rep(1,25),rep(0,34))  
> optimize(liktheta,interval=c(0.01,0.99),data= newdata,maximum=TRUE)  
$maximum  
[1] 0.4237282  
  
$objective  
[1] -40.20656
```

Can you estimate this parameter in any other way? (turn your brain on!!!)

```
> 25/(25+34)  
[1] 0.4237288
```

This is just a binomial proportion: the MLE of a proportion is the empirical proportion (not that surprising, right?)

**FOR A  
GAUSSIAN**

#now get a likelihood for a gaussian

```
liknorm=function(pars,data){  
  media=pars[1]  
  desvio=pars[2]  
  minusloglik=-sum(log(dnorm(xs,mean=media,sd=desvio)))  
  return(minusloglik)  
}
```

```
xs=rnorm(100,mean=2,sd=0.7)  
optim(par=c(1,1),fn=liknorm,data=xs)
```

```
> xs=rnorm(100,mean=2,sd=0.7)  
> optim(par=c(1,1),fn=liknorm,data=xs)  
$par  
[1] 1.9020340 0.6960986
```

Not so easy to do the grid search now, but still possible... challenge: do it at home (level of difficulty: hard!)



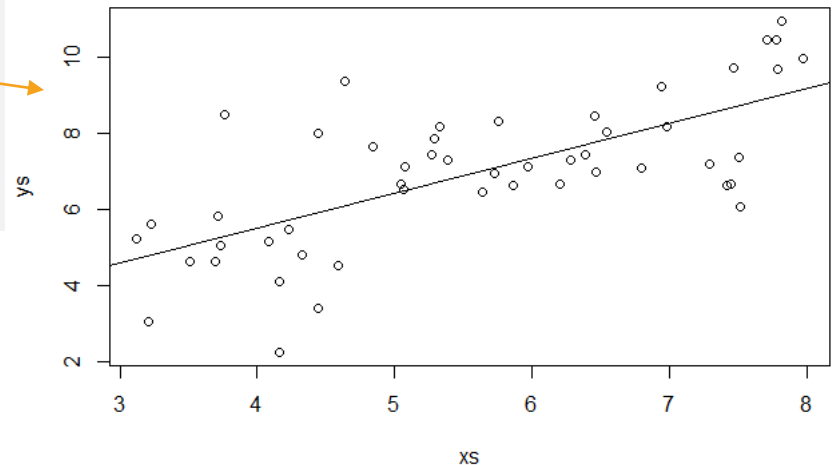
# FOR A LINEAR MODEL

Estimating the parameters of  
a straight line  
by  
maximum likelihood:  
Comparing R  
with  
doing it “manually”

Model:  $y_i = a + b \times x_i + e_i$

where  $e_i$  is Gaussian with mean 0 and standard deviation  $\sigma$

```
set.seed(123)
xs=runif(50,3,8)
ys=2.4+0.8*xs+rnorm(50,mean=0,sd=1.5)
plot(xs,ys)
mod1=lm(ys~xs)
summary(mod1)
abline(mod1)
```



```
> summary(mod1)
```

Call:

```
lm(formula = ys ~ xs)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.3837	-0.8368	-0.0985	0.8239	3.2792

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	1.8429	0.7925	2.326	0.0243	*
xs	0.9145	0.1369	6.678	2.29e-08	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.411 on 48 degrees of freedom

Multiple R-squared: 0.4816, Adjusted R-squared: 0.4708

F-statistic: 44.6 on 1 and 48 DF, p-value: 2.293e-08

$$\text{Model: } y_i = a + b \times x_i + e_i$$

where  $e_i$  is Gaussian with mean 0 and standard deviation  $\sigma$

This also means that that

$$e_i = Y_i - (a + b \times x_i)$$

is Gaussian with mean 0 and standard deviation  $\sigma$

```

liklm=function(pars,data){
  #data must be a data.frame with columns y and x
  a=pars[1]
  b=pars[2]
  sigma=pars[3]
  ps=dnorm(data$y-(a+b*data$x),mean=0,sd=sigma)
  #minus loglik
  loglik=-sum(log(ps))
  return(loglik)
}

```

```

optim(par=c(2,1,1),fn=liklm,data=data.frame(y=ys,x=xs))
summary(mod1)

```

```
> optim(par=c(2,1,1),fn=liklm,data=data.frame(y=ys,x=xs))
```

```
$par
[1] 1.8425878 0.9145521 1.3819778
```

```
$value
[1] 87.12513
```

```
$counts
function gradient
          92          NA
```

```
$convergence
[1] 0
```

```
$message
```

```
> summary(mod1)
```

```
Call:
lm(formula = ys ~ xs)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-3.3837 -0.8368 -0.0985  0.8239  3.2792
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.8429     0.7925   2.326  0.0243 *
xs           0.9145     0.1369   6.678 2.29e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

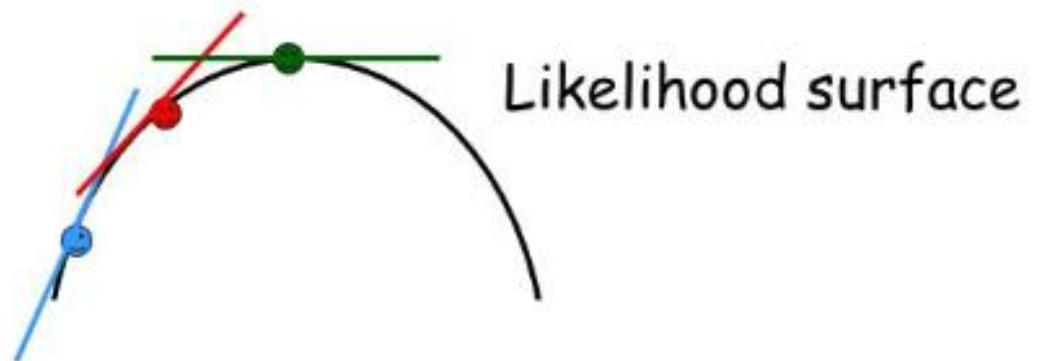
```
Residual standard error: 1.411 on 48 degrees of freedom
Multiple R-squared:  0.4816,    Adjusted R-squared:  0.4708
F-statistic:  44.6 on 1 and 48 DF,  p-value: 2.293e-08
```

# Local Optimization - Gradient Methods

- Derivative-based (Newton-Raphson) methods:

$$l(\theta | y) = p(y | \theta)$$

$$\frac{dl}{d\theta} = \dots = 0$$



General approach: Vary parameter estimate systematically and search for zero slope in the first derivative of the likelihood function...(using numerical methods to estimate the derivative, and checking the second derivative to make sure it is a maximum, not a minimum)

Why are these cool? Because these are easily extendable such that now we could have the parameters being dependent of the value of covariates.

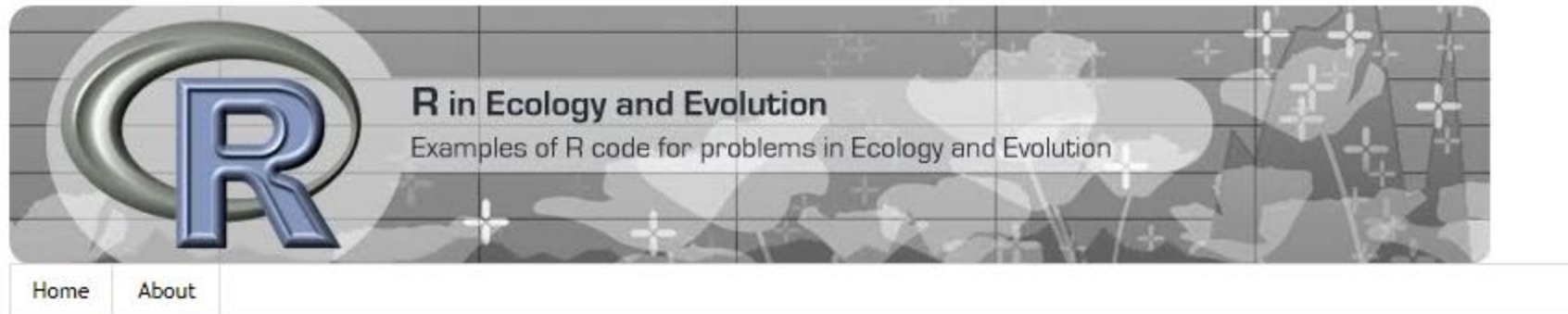
e.g. in the nest example, the  $\theta$  might be a function of

- the size of the father
- the size of the mother
- the habitat around the nest
- the distance to a river
- the distance to a road,
- etc



And we can make a more complicated likelihood that would allow us to estimate how these different variables might affect the parameter value – lots of ecological insight to be gain based on data and models – but that is no longer part of Ecological Modeling 101 😊 !

# Generalized Linear Models



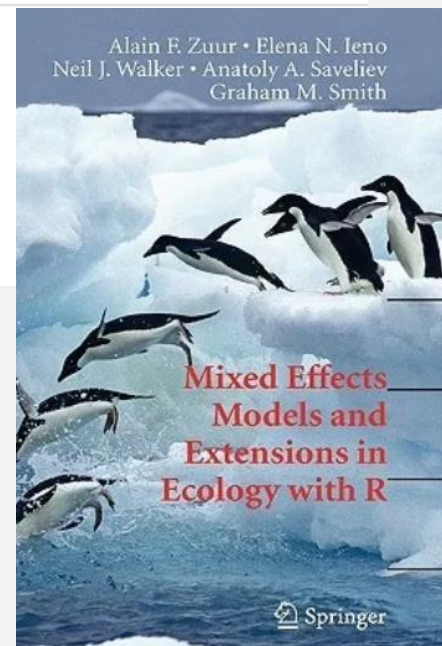
Sunday, May 14, 2017

A gentle introduction to Generalized Linear Models in R

What are generalized linear models?

<http://r-eco-evo.blogspot.com/2017/05/generalized-linear-models.html>

<http://spatialecology.weebly.com/r-code--data/category/glm>







## Regression and GLM

---

### GLM (generalized linear models)

- These generalize the linear model by allowing different response types, which implies that the errors no longer need to be Gaussian
- The violation of one (or several) of the assumptions of the linear model is common, most likely being the rule and not the exception with ecological data. Therefore GLM's are extremely useful within Ecological Modelling.



# GLM (generalized linear models)

A GLM has 3 key components:

- A probability distribution family for the response (from the exponential family)
- A linear predictor (just as before for the linear model)
- A link function, that links the mean value of the response to the linear predictor

# The link function

$$\mathbf{E}(\mathbf{Y}) = \boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta})$$

As an example, if you have a log link function (a common choice, e.g. to enforce a positive response), then that means you have

$$E(Y) = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$$

And naturally this also means that

$$\log(E(Y)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Note that, irrespectively of what the parameter values might be, the mean value of the response will always be positive with a log link!



## Regression and GLM

---

# Logistic Regression (the simplest GLM example)

- A logistic regression has the objective to model/predict a variable resulting from the success of failure, presence absence, of an experiment, based on a set of independent variables (these can be continuous or categorical)
- This is an extremely useful and popular model:

*Explain what are the differences between the habits of healthy people vs unhealthy people*

*Modeling the presence absence of species*

*Modeling the intentions of voters in a given candidate in the elections*

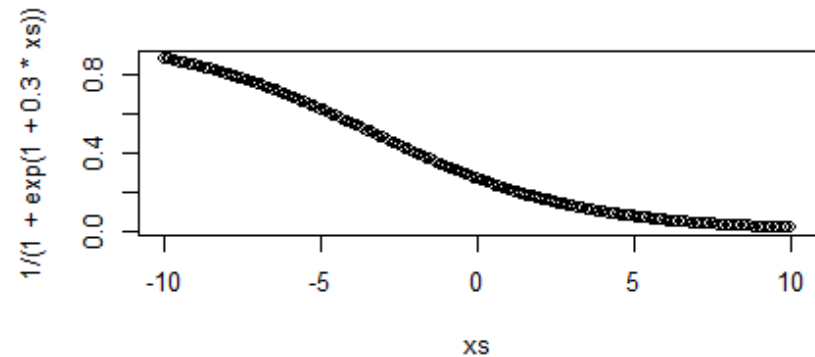
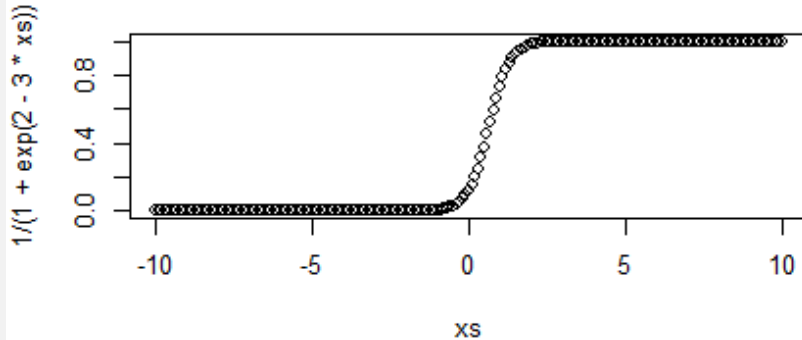
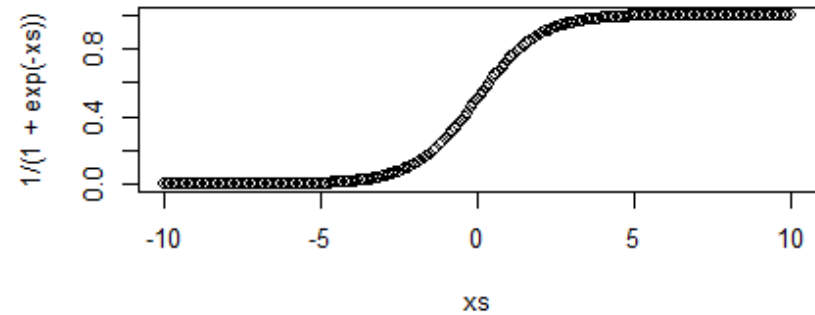
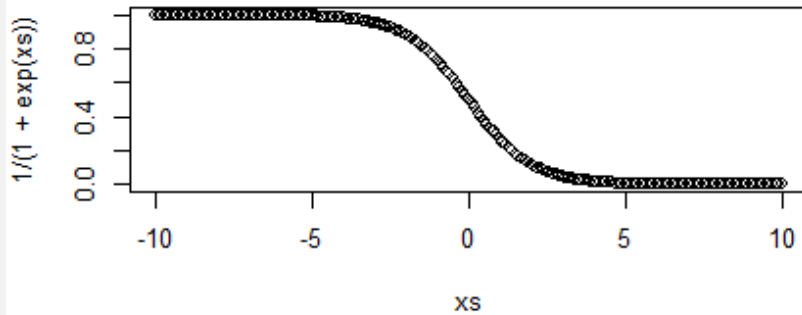
*Modelling mortality/survival of individuals as a function of environmental covariates*



## Regressão logística

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

$$\text{logit}^{-1}(\alpha) = \text{logistic}(\alpha) = \frac{1}{1 + \exp(-\alpha)} = \frac{\exp(\alpha)}{\exp(\alpha) + 1}$$



$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_k x_{k,i} \longrightarrow p_i = \frac{1}{1 + \exp(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}$$

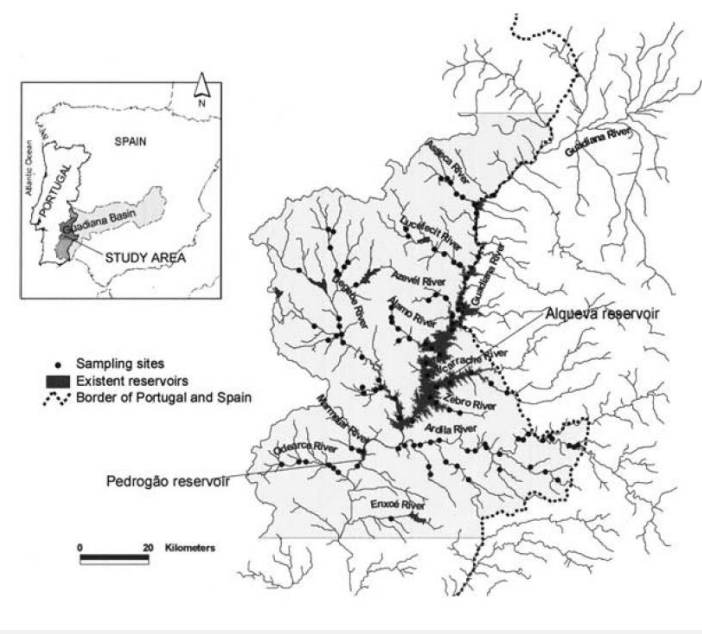
Full Access

## Selection of Priority Areas for Fish Conservation in Guadiana River Basin, Iberian Peninsula

Selección de Áreas Prioritarias para la Conservación de Peces en la cuenca del Río Guadiana, Península Ibérica

A. F. FILIPE, T. A. MARQUES, P. TIAGO, F. RIBEIRO, L. MOREIRA DA COSTA, I. G. COWX, M. J. COLLARES-PEREIRA

First published: 30 January 2004 | <https://doi.org/10.1111/j.1523-1739.2004.00620.x> | Cited by: 71



The conservation value of each area  $j$  ( $j = 1, \dots, N$ ) ( $VA_j$ ) across the study region was calculated as the sum of the products of the probability of occurrence of each species at each area and the corresponding species conservation value ( $VS_k$ ):

$$VA_j = \sum_{k=1}^S (P_{kj} \times VS_k),$$

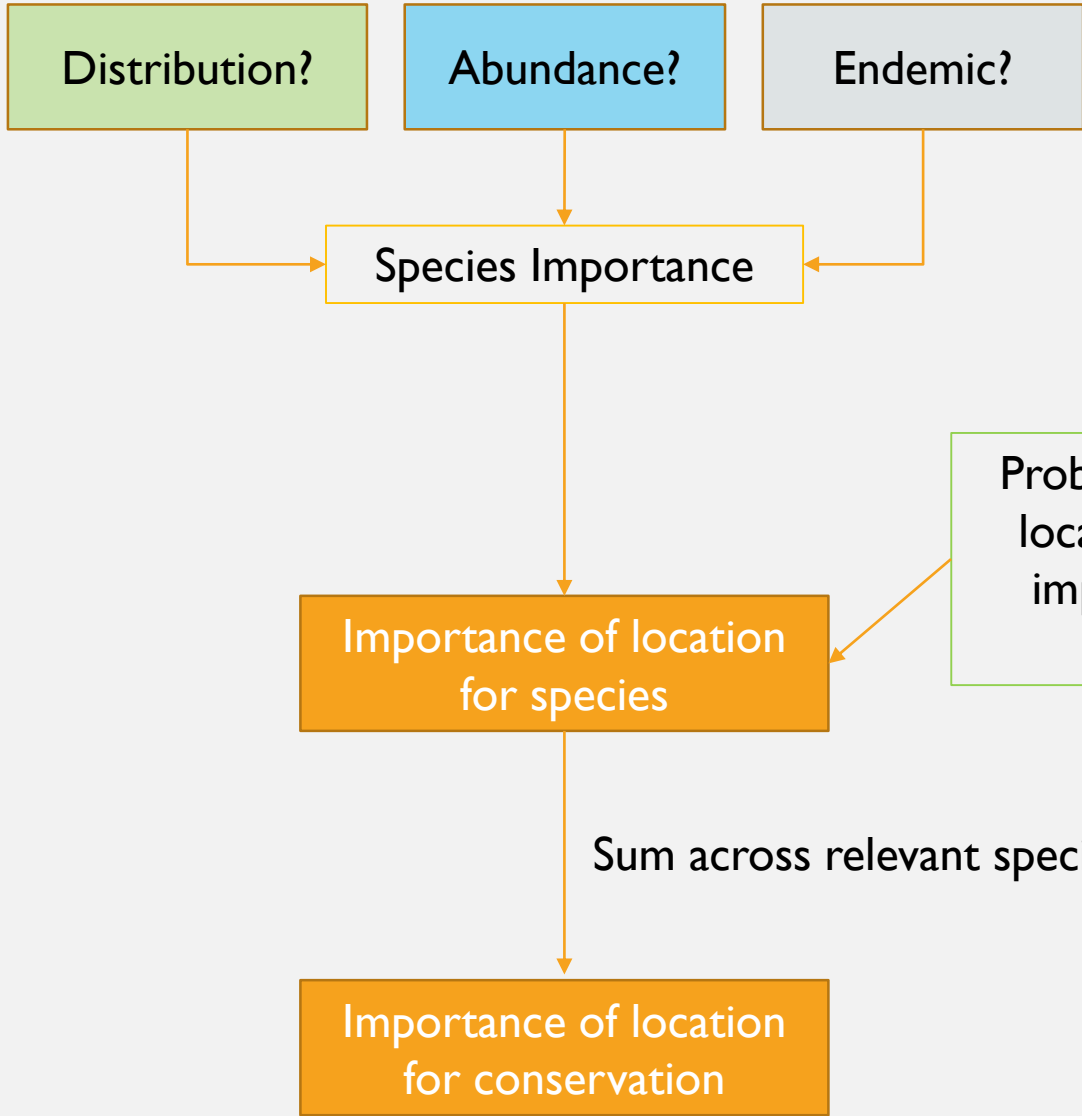
where  $P_{kj}$  is the probability of occurrence of species  $k$  in area  $j$  and  $VS_k$  is the conservation value of species  $k$ . The conservation value of area  $j$  varies between 0 and 100.

than occurrence alone. Therefore, the conservation value for species  $k$  ( $VS_k$ ) is

$$VS_k = \left( a \frac{1/O_k}{\sum_{i=1}^S 1/O_i} + b \frac{1/\ln T_k}{\sum_{i=1}^S 1/\ln T_i} + c \frac{1/E_k}{\sum_{i=1}^S 1/E_i} \right) \times 100,$$

where  $S$  is the number of species considered,  $O_k$  is the total number of sampling sites where species  $k$  occurred in all samples,  $T_k$  is the total number of captured individuals of species  $k$  in all samples,  $E_k$  is the endemic value of species  $k$  according to its distribution range (the species with lowest value has the most restricted distribution), and  $a$ ,  $b$ , and  $c$  are weighting factors that may vary according to the importance placed on conserving distribution ( $a$ ), abundance ( $b$ ), or endemicity ( $c$ ).

Logistic regression modelling presence absence of species as a function of environmental covariates (e.g. river width, distance to sea, elevation, temperature, etc)



A species is more important as it becomes local, low in density, and is endemic

Probability of species being in a given location – the higher it is the more important that place is for a given species

## Implementação de um GLM

As duas componentes estruturais a seleccionar e que vão ditar o tipo de modelo a implementar são:

- Modelo probabilístico de distribuição dos erros
- Função de ligação





# regressão e MLG

## GLM (generalized linear models)

Common distributions with typical uses and canonical link functions

Distribution	Support of distribution	Typical uses	Link name	Link function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\beta = \mu$
Exponential	real: $(0, +\infty)$	Exponential-response data, scale parameters	Inverse	$\mathbf{X}\beta = \mu^{-1}$
Gamma				
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\beta = \mu^{-2}$
Poisson	integer: $[0, +\infty)$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\beta = \ln(\mu)$
Bernoulli	integer: $[0, 1]$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\beta = \ln\left(\frac{\mu}{1-\mu}\right)$
Binomial	integer: $[0, N]$	count of # of "yes" occurrences out of N yes/no occurrences		
Categorical	integer: $[0, K)$	outcome of single K-way occurrence		
	K-vector of integer: $[0, 1]$ , where exactly one element in the vector has the value 1			
Multinomial	K-vector of integer: $[0, N]$	count of occurrences of different types (1 .. K) out of N total K-way occurrences		

Como escolher a família de distribuições a usar num GLM (ou GAM): tem a ver com o tipo de dados, em particular com os valores que a **variável resposta** pode tomar.

- Dados Contínuos: Gaussiana
- Dados Contínuos apenas positivos ou com variância crescente: Gamma
- Dados de contagens: Poisson
- Dados de contagens com variância maior que a média: Binomial Negativa
- Dados de presença/ausência: Binomial
- Contagens com variância menor que a media (raro): Binomial
- Numero de sucessos em  $n$  provas: Binomial

Existem ainda outras famílias mais gerais, como a Quasi-Poisson e Quasi-Binomial, ou a distribuição de Tweedie.