Aula 17 Goodies*

* Goodies related to animals, plants and numbers…

A talk delivered at

CURRENT TRENDS IN ECOLOGICAL STATISTICS ARE DETACHED FROM ECOLOGISTS' STATISTICAL TEACHING

"...Ainda ensinamos estatística como há 20 anos atrás..." – Maria do Rosário Oliveira, 2019

TIAGO A. MARQUES

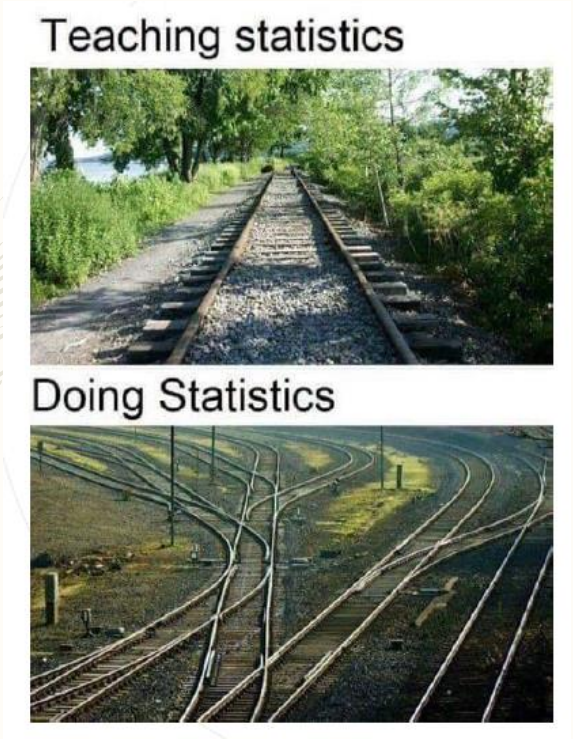
CREEM
Centre for Research into Ecological and Environmental Modelling

University of St Andrews

CEAUL
Centro de Estatística e Aplicações
Universidade de Lisboa

Ciências ULisboa | Biologia Animal

9TH NOVEMBER 2019

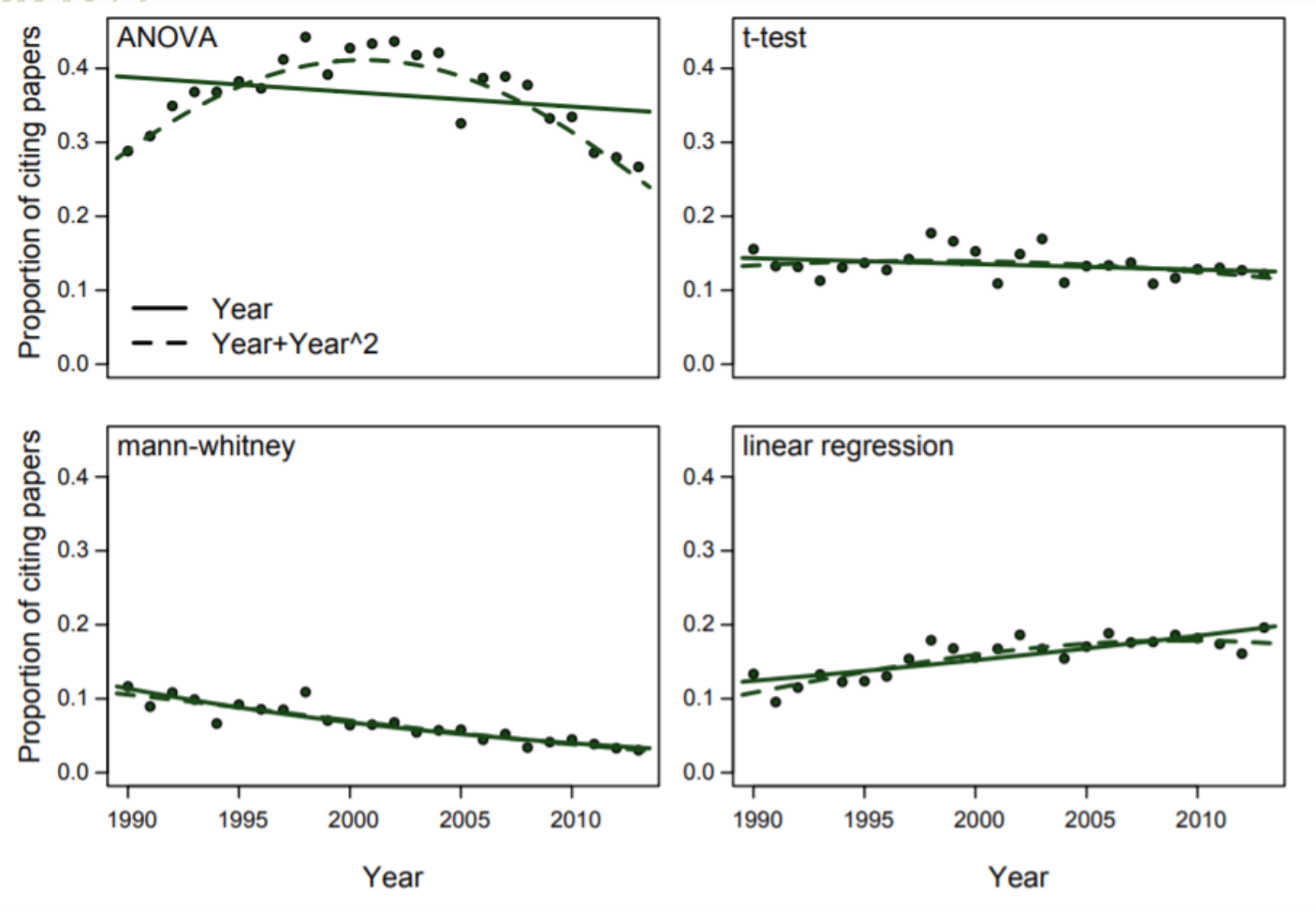


Walter J. Radermacher

esa ECOSPHERE

The mismatch between current statistical practice and doctoral training in ecology

JUSTIN C. TOUCHON[1] AND MICHAEL W. MCCOY[2,†]



Gestão de Páginas

- Ecologia Numérica
  - Ecologia Numérica(Tecnologias de Informação)
  - Teoricas
  - Práticas
    - Week1
    - Week 2
    - Week 3
    - Week 4
    - Week 5
    - Week 6
    - Week 7
    - Week 8
    - Week 9
      - Extra FT7b
  - PDFs
  - Outros Recursos

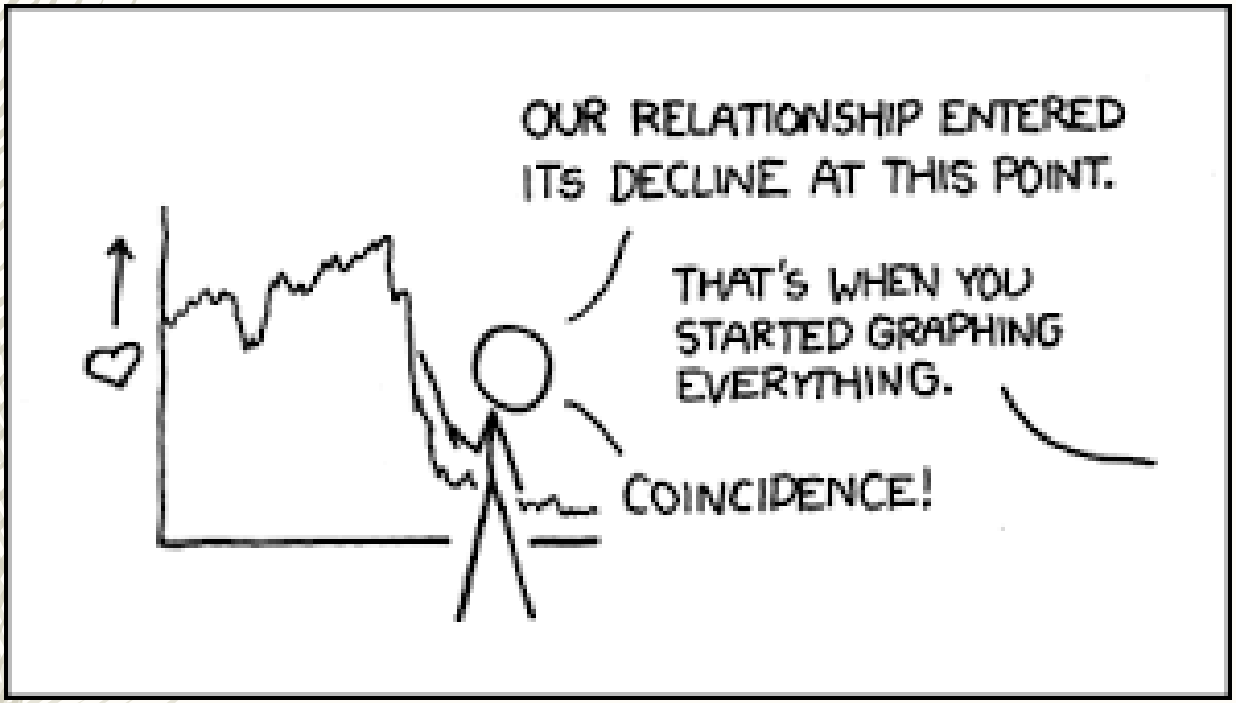
✚ Criar

PDFs

Página | **Ficheiros 4** | Permissões | Link

Adicionar Ficheiro

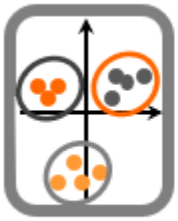
#	Nome
1	Numerical Ecology with R *Borcardetal2001EcologyUseR.pdf*
2	The R Book.pdf
3	Publication bias: What are the challenges and can they be overcome? *jpn-37-149.pdf*
4	The mismatch between current statistical practice and doctoral training in ecology *Touchon&McCoy_2016_MismatchStatsPractive&TeachingEcology.pdf*


OUR RELATIONSHIP ENTERED ITS DECLINE AT THIS POINT.
THAT'S WHEN YOU STARTED GRAPHING EVERYTHING.
COINCIDENCE!

AULAS TP'S

Esta semana estamos a trabalhar na FT7.pdf

Depois disso, vocês deverão realizer e entregar a FT7b.pdf



Ecologia Numérica

Componente Teórica - Prática

Ficha de trabalho

7b

Esta é uma ficha de trabalho mais longa que as anteriores, para ser realizada de forma semi-independente (ou seja, sem que seja necessário um apoio directo nas aulas TPs). Pode ser realizada de forma individual ou, caso assim o desejem, a pares. Pretende integrar os conhecimentos adquiridos ao longo da primeira metade da cadeira de Ecologia Numérica, com enfase nos modelos de regressão linear (*Linear Model*, LM) e modelos lineares generalizados (*Generalized Linear Models*, GLMs), e chegando a explorar um modelo aditivo generalizado (*Generalized Additive Models*, GAM).

Entrega: envio por e-mail de um relatório dinâmico em R Markdown, enviando o pdf e o respectivo .Rmd

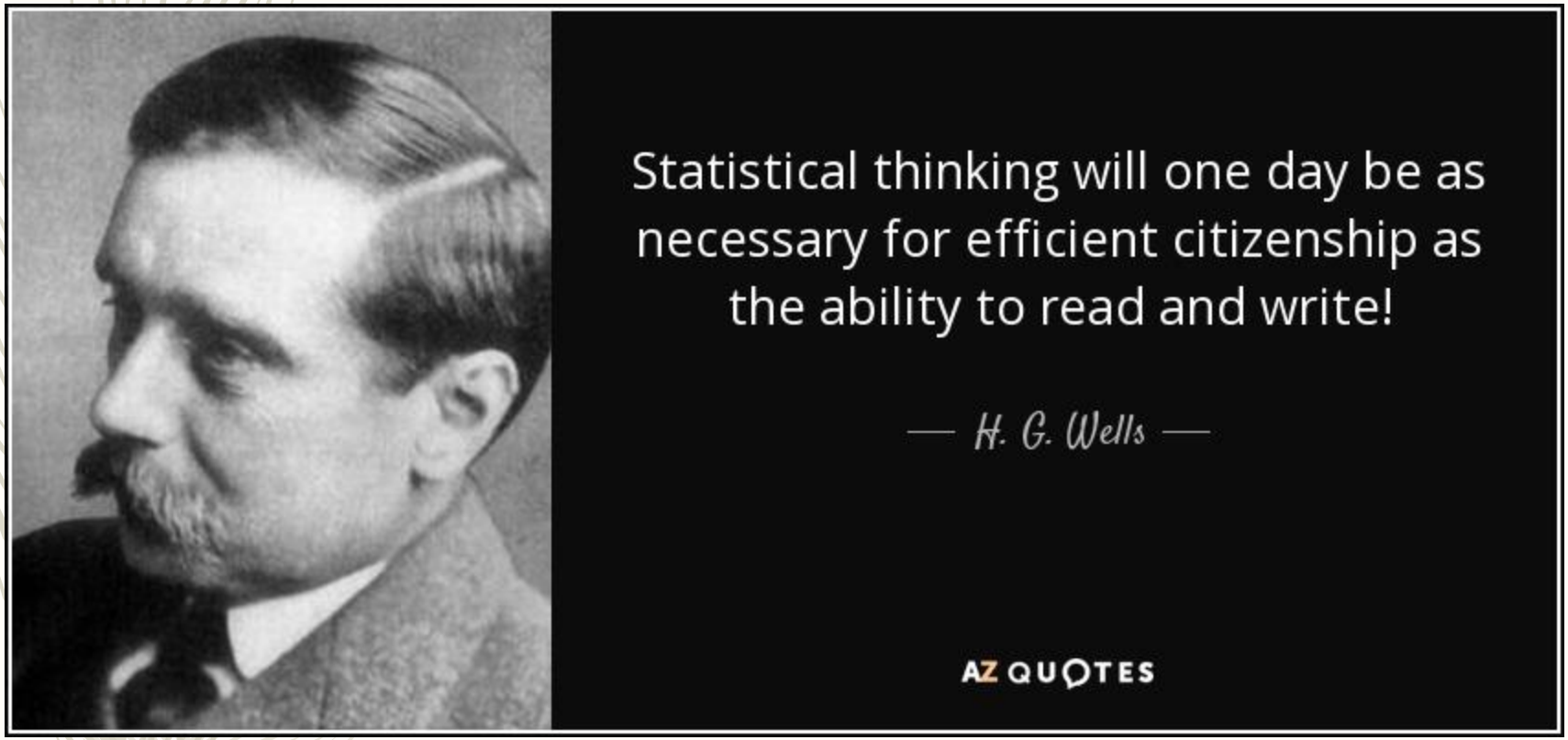
Data limite para entrega: 24 de Novembro

Nome dos ficheiros: FT7A*****.Rmd ou FT7A1*****A2*****.Rmd e FT7A***** Ou FT7A1*****A2*****.Rmd, onde ***** corresponde ao(s) número(s) de aluno correspondente(s) ao(s) autor(es).

As resoluções das fichas TPs são consideradas material essencial, que deverão estudar com cuidado.

Qualquer material que esteja abordado nessas resoluções é considerado relevante, e como tal, como material potencialmente examinável.

Ecologia Numérica - Aula Teórica 17 – 12-11-2019



https://www.azquotes.com/quote/685542

Desenvolvimento de um GLM

A sequência de procedimentos para a implementação de um GLM é, em geral, a seguinte:

- Análise e selecção de variáveis candidatas;
- Escolha de um modelo base (e.g. lm, ou glm com indicação da distribuição do erro e função de ligação)
- Avaliação do modelo e geração de sub-modelos
- Comparação dos sub-modelos
- Selecção do modelo final
- Validação e/ou aplicação

Como escolher a familia 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.

Ver Hoef, J. M. & Boveng, P. L. 2007 Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology* **88**: 2766-2772

Gestão de Páginas

- Ecologia Numérica
  - Ecologia Numérica(Tecnologias de Informação)
  - Teoricas
  - Práticas
    - Week1
    - Week 2
    - Week 3
    - Week 4
    - Week 5
    - Week 6
    - Week 7
    - Week 8
    - Week 9
      - Extra FT7b
    - PDFs
  - Outros Recursos

+ Criar

PDFs

Página | Ficheiros 5 | Permissões | Link

Adicionar Ficheiro

#	Nome
1	Numerical Ecology with R *Borcardetal2001EcologyUseR.pdf*
2	The R Book.pdf
3	Publication bias: What are the challenges and can they be overcome? *jpn-37-140.pdf*
4	The mismatch between current statistical practice and doctoral training in ecology *Touchon&McCoy_2016_MismatchStatsPractive&TeachingEcology.pdf*
5	Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? *Hoef&Boveng2007.pdf*

regressão e MLG

Genericamente, os vários modelos produzidos podem ser comparados com base nas seguintes ferramentas:

	Vantagens	Desvantagens
Testes da qualidade do ajustamento ***Goodness-of-fi,*** **R^2, *Deviance***	Representa a variância explicada	Muito influenciados pela parametrização (sub ou sobre)
Testes de hipóteses	Estrutura hierárquica	Robustez e potência reduzidas
Teoria da informação AIC, BIC	Compromisso ajustamento / complexidade	Muito genérica
Validação cruzada (*Cross-validation*)	Avalia a qualidade preditiva	Computacionalmente intensiva

Por exemplo *leave-one-out* cross validation: para cada observação (1) retiramos essa observação, (2) re-ajustamos o modelo (3) prevemos a observação que ficou de fora (4) somamos os erros de predição

O conceito de máxima verosimilhança

Maximum likelihood estimator (MLE)

Dado um modelo, podemos calcular a probabilidade dos dados $P(\text{dados}|\text{modelo},\theta)$

Dados um modelo e os dados, podemos avaliar quais os valores dos parâmetros mais prováveis: $P(\theta|\text{dados},\text{modelo})$ - são os estimadores de máxima verosimilhança (MLE)

Na realidade, por trás de muito daquilo que fazemos em estatística estão estimadores de máxima verosimilhança!

Os estimadores dos mínimos quadrados do a (ordenada na origem) e do b (declive da reta) são MLE!
A média amostral, o estimador da média numa Gaussiana ou numa Poisson é um MLE
O estimador do desvio padrão é um MLE (corrigido!)
etc…

Para qualquer modelo (dos que vocês conhecem) existe por trás uma verosimilhança!

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 θ
X=0 with probability 1-θ

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

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!)

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

Note that $P(\underline{X})$ can be seen either as a function of the data, conditional on θ, $P(\underline{X}| \theta)$, or a function of θ , 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 n1 as the number of 1's and n0 the number of 0's in our sample, we have a likelihood function given by

$L(\theta|X)= \theta^{n1} (1- \theta)^{n0}$

What is the value of θ that maximizes the likelihood function?

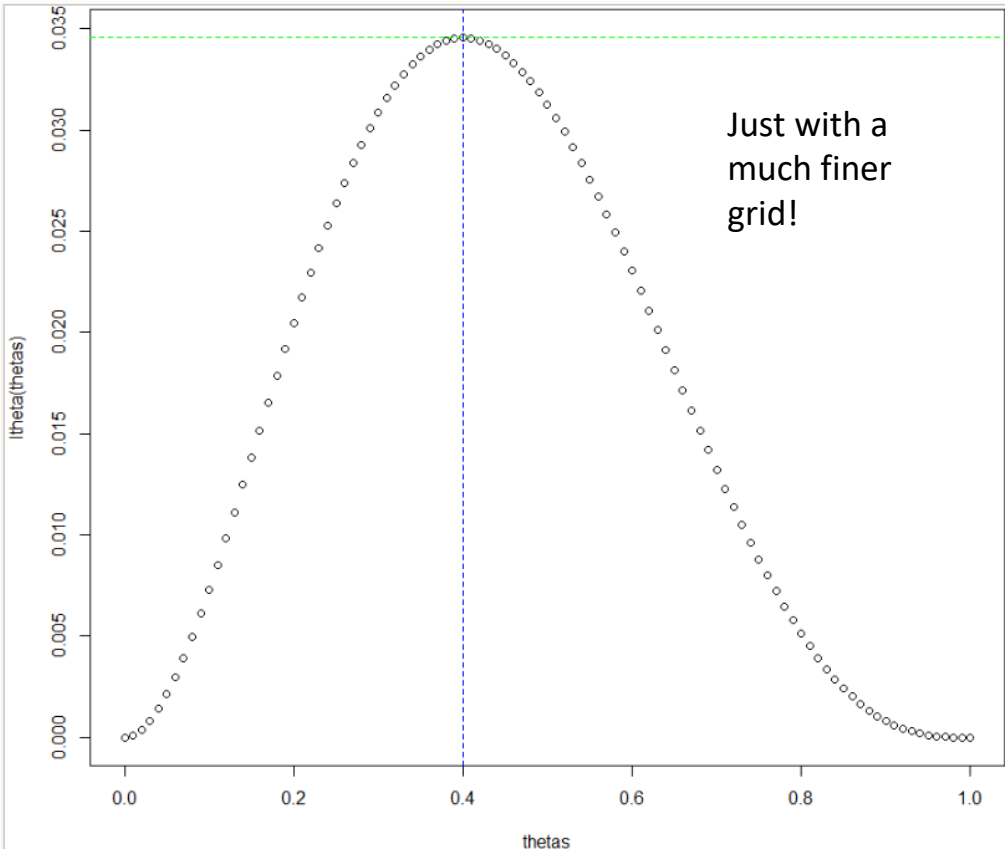
Using a grid search approach

	θ	$L(\theta\|\underline{X})$
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

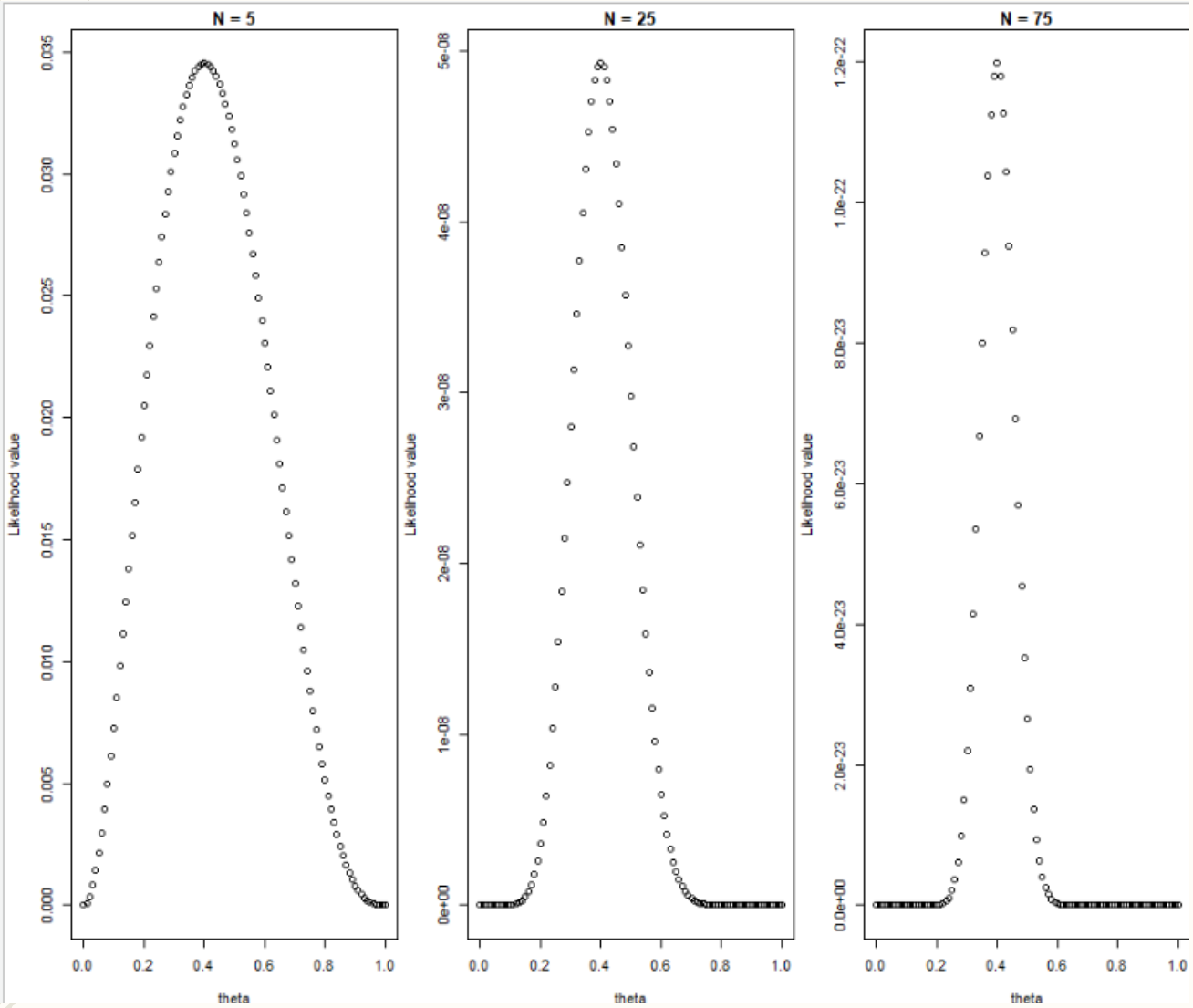
```
ltheta=function(theta,n1=2,n0=3){
  lik=(theta)^n1*(1-theta)^n0
  return(lik)}

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")
```

A deviance ... (em PT desviância...)

D=2(LL saturated model -LL proposed model)

é (o dobro d') a diferença das razões log-verosimilhança (ou *log-likelihood*, LL) de um modelo comparativamente ao modelo sobre-parametrizado (modelo saturado, i.e. com tantos parâmetros como observações!).

Esta diferença reflecte a qualidade do ajustamento.

Nos GLM é frequente comparar a deviance residual com a deviance do modelo nulo (apenas uma média).

regressão e MLG

Null Deviance = 2(LL(Saturated Model) - LL(Null Model))
df = df_Sat - df_Null

Residual Deviance = 2(LL(Saturated Model) - LL(Proposed Model))
df = df_Sat - df_Res

Saturated Model – modelo que considera que cada observação dos dados tem o seu parâmetro próprio (modelo com *n* parâmetros).

Null Model – considera apenas um parâmetro para todas as observações (modelo com 1 parâmetro, i.e. uma média global).

Proposed Model – *p* parâmetros mais um termo relativo à intercepção, ou seja *p+1* parâmetros.

regressão e MLG

Null Deviance = 2(LL(Saturated Model) - LL(Null Model))

Se a Null Deviance for pequena, significa que um modelo com apenas 1 parâmetro explica bem os dados.

Residual Deviance = 2(LL(Saturated Model) - LL(Proposed Model))

Se a Residual Deviance for pequena significa que o modelo proposto com p variáveis explica bem os dados.

A diferença entre deviances é em geral testada através de uma estatística de teste que segue uma distribuição qui-quadrado.

(Null Deviance - Residual Deviance) tem distribuição qui-quadrado com p-1 graus de liberdade.

Null Deviance = 2(LL(Saturated Model) - LL(Null Model)) df = N-1

Residual Deviance = 2(LL(Saturated Model) - LL(Proposed Model)) df = N-p

Residual Deviance should be small (compared to the Null Deviance)

Null Deviance - Residual Deviance $\longrightarrow$ χ^2 df = (N-1)-(N-p)=p-1

Deviance test H0: simpler model is enough to describe the data

AIC - Akaike information criterion

AIC= 2k - 2LL (k=número de parâmetros, LL log likelihood)

É uma medida da qualidade relativa (i.e. da parcimónia) do ajustamento de um modelo estatístico a um determinado conjunto de dados.

Uma parte penaliza a complexidade, a outra valoriza o ajustamento

Constitui uma ferramenta para a seleção de modelos (model selection).

O AIC baseia-se na teoria da informação – fornece uma estimativa relativa da quantidade de informação perdida quando um determinado modelo é usado.

O AIC não dá uma medida absoluta da qualidade do modelo e por isso apenas serve para a comparação entre modelos – não diz nada sobre se o melhor é bom!

Aho et al. 2014 Model selection for ecologists: the worldviews of AIC and BIC *Ecology* **95**: 631-636

regressão e MLG

Há diversas metodologias de seleção de modelos.

Frequentemente, os modelos podem ter um número muito elevado de termos, devido às interações entre variáveis explicativas.

As metodologias mais utilizadas são procedimentos iterativos de avaliação multi-etápica, com base em critérios como o AIC ou indicadores equivalentes, uma vez que outros (e.g. R^2, deviance) são extremamente sensíveis em relação ao número de parâmetros e ao número de observações.

regressão e MLG

	densidade	latitude	precipitacao	cob.veg	humidade	insolacao	estradas	temperatura
1	2	41	152	256	25	21	1	15
2	3	41	145	354	35	22	3	15
3	3	41	120	564	41	25	5	16
4	11	40	86	324	12	32	4	16
5	12	40	85	125	24	41	3	17
6	15	40	65	90	25	51	5	17
7	25	39	45	35	26	65	9	17
8	28	39	32	145	28	70	4	18
9	29	39	31	25	42	85	5	18
10	35	38	24	10	32	95	2	18
11	54	37	12	38	31	94	4	19
12	65	37	10	64	8	102	2	20

Variável resposta, ou dependente, que tentamos explicar à custa das variáveis independentes