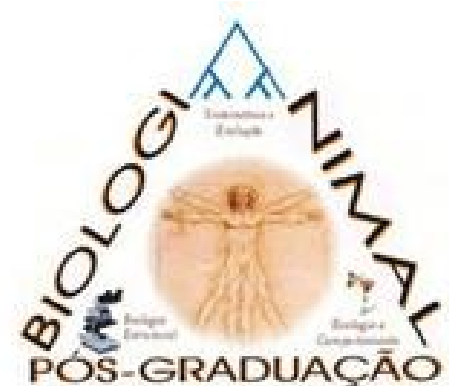


Estatística aplicada à ecologia usando o



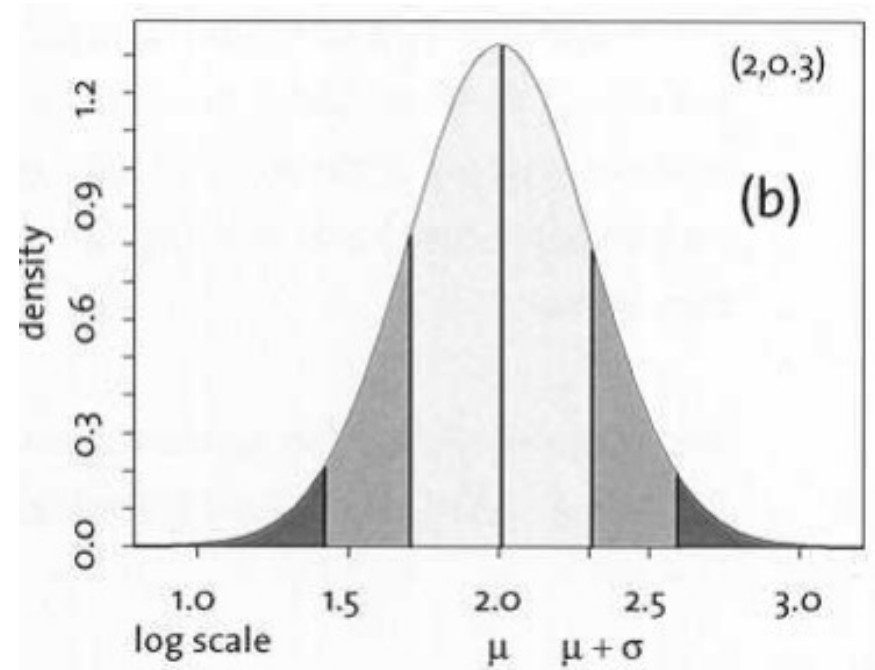
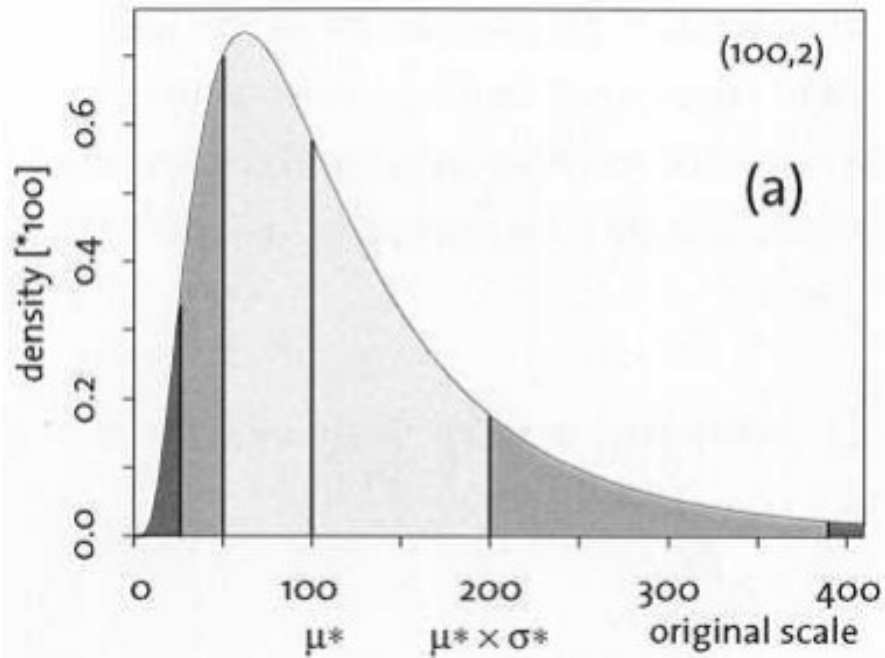
Modelos lineares generalizados (GLM)

Docentes: Fernando Rodrigues da Silva
Diogo Borges Provete
Thiago Gonçalves Souza



Modelo Linear Generalizado

É uma flexível generalização da regressão linear.



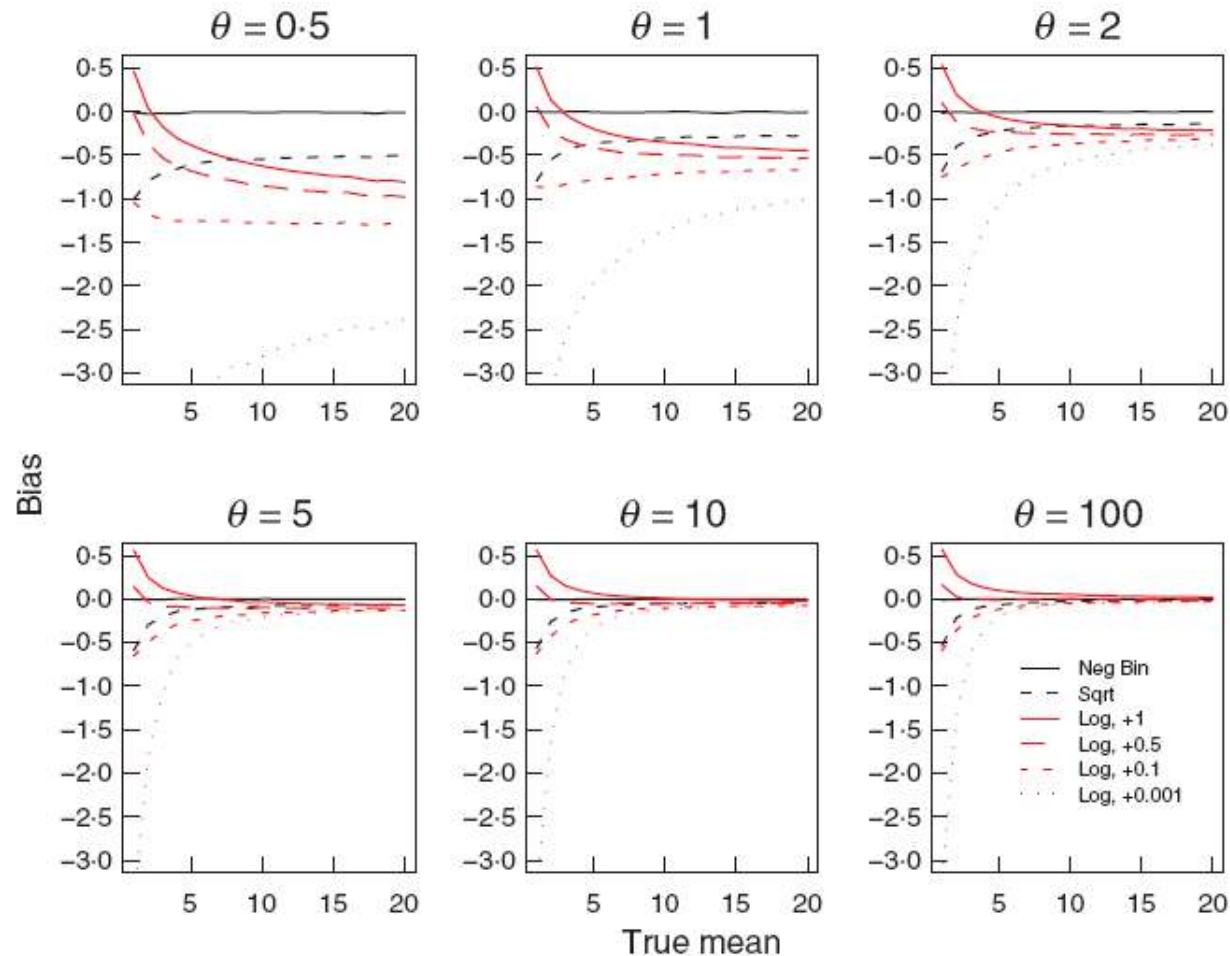
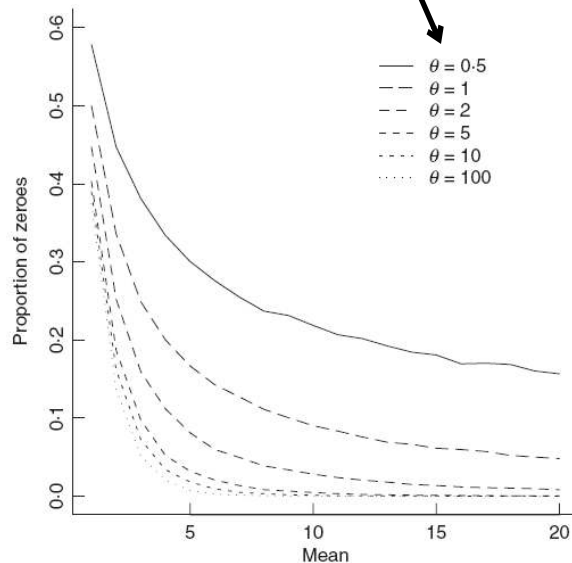
Do not log-transform count data

Robert B. O'Hara^{1*} and D. Johan Kotze²

¹Biodiversity and Climate Research Centre, Senckenberganlage 25, D-60325 Frankfurt am Main, Germany and

²Department of Environmental Sciences, PO Box 65, University of Helsinki, Helsinki FI-00014, Finland

Quanto maior,
menor a dispersão



A protocol for data exploration to avoid common statistical problems

Alain F. Zuur^{1,2}, Elena N. Ieno^{1,2} and Chris S. Elphick³

¹Highland Statistics Ltd, Newburgh, UK; ²Oceanlab, University of Aberdeen, Newburgh, UK; and ³Department of Ecology and Evolutionary Biology and Center for Conservation Biology, University of Connecticut, Storrs, CT, USA

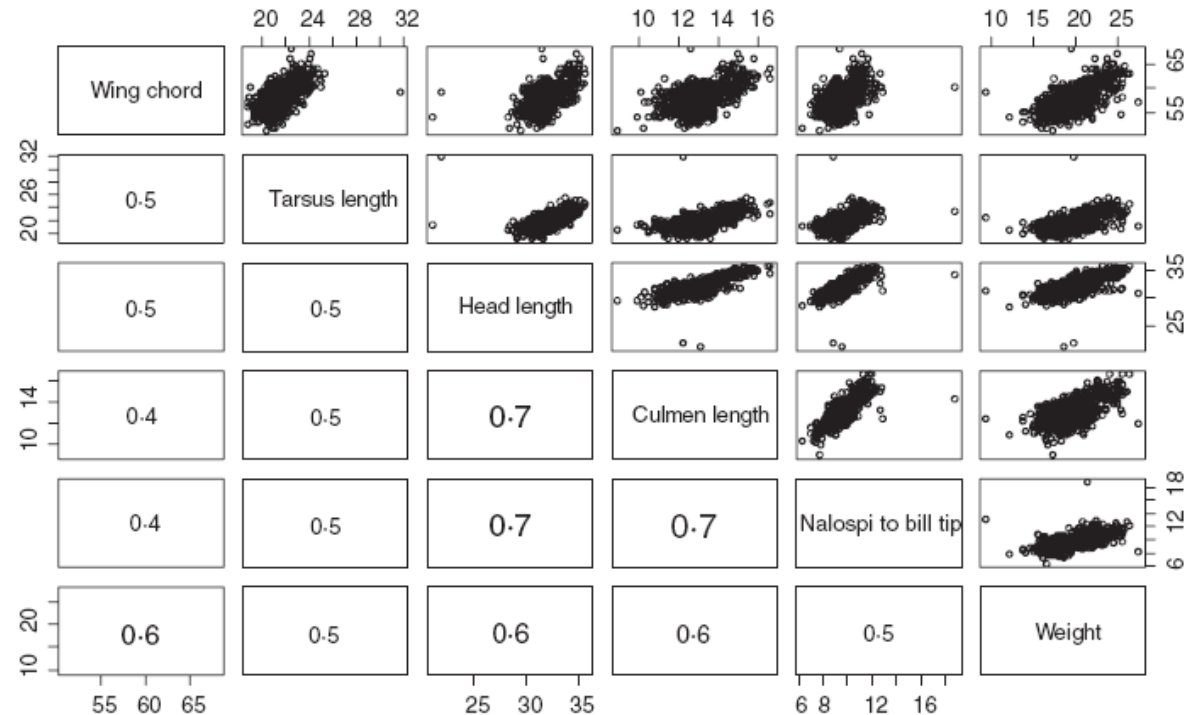
1	Formulate biological hypothesis Carry out experiment & collect data	
	Data exploration	
2	1. Outliers Y & X	<i>boxplot & Cleveland dotplot</i>
	2. Homogeneity Y	<i>conditional boxplot</i>
	3. Normality Y	<i>histogram or QQ-plot</i>
	4. Zero trouble Y	<i>frequency plot or corrgram</i>
	5. Collinearity X	<i>VIF & scatterplots correlations & PCA</i>
	6. Relationships Y & X	<i>(multi-panel) scatterplots conditional boxplots</i>
	7. Interactions	<i>coplots</i>
	8. Independence Y	<i>ACF & variogram plot Y versus time/space</i>
3	Apply statistical model	

A protocol for data exploration to avoid common statistical problems

Alain F. Zuur^{1,2}, Elena N. Ieno^{1,2} and Chris S. Elphick³

¹Highland Statistics Ltd, Newburgh, UK; ²Oceanlab, University of Aberdeen, Newburgh, UK; and ³Department of Ecology and Evolutionary Biology and Center for Conservation Biology, University of Connecticut, Storrs, CT, USA

1	Formulate biological hypothesis Carry out experiment & collect data
	Data exploration
	1. Outliers Y & X <i>boxplot & Cleveland dotplot</i>
	2. Homogeneity Y <i>conditional boxplot</i>
	3. Normality Y <i>histogram or QQ-plot</i>
2	4. Zero trouble Y <i>frequency plot or correlogram</i>
	5. Collinearity X <i>VIF & scatterplots correlations & PCA</i>
	6. Relationships Y & X <i>(multi-panel) scatterplots conditional boxplots</i>
	7. Interactions <i>coplots</i>
	8. Independence Y <i>ACF & variogram plot Y versus time/space</i>
3	Apply statistical model



A consequência da colinearidade é erros padrões muito elevados. Mesmo que o R^2 seja bastante alto, nenhum dos coeficientes da regressão é estatisticamente significativo.

A protocol for data exploration to avoid common statistical problems

Alain F. Zuur^{1,2}, Elena N. Ieno^{1,2} and Chris S. Elphick³

¹Highland Statistics Ltd, Newburgh, UK; ²Oceanlab, University of Aberdeen, Newburgh, UK; and ³Department of Ecology and Evolutionary Biology and Center for Conservation Biology, University of Connecticut, Storrs, CT, USA

Covariate	<i>P</i> -value (full model)	VIF	<i>P</i> -value (collinearity removed)	<i>P</i> -value (reduced model)
% <i>Juncus gerardii</i>	0.0203	44.9953	0.0001	0.00004
% Shrub	0.9600	2.7818	0.0568	0.0727
Height of thatch	0.9989	1.6712	0.8263	
% <i>Spartina patens</i>	0.0640	159.3506	0.3312	
% <i>Distichlis spicata</i>	0.0527	53.7545	0.2538	
% Bare ground	0.0666	12.0586	0.8908	
% Other vegetation	0.0730	5.8170	0.9462	
% <i>Phragmites australis</i>	0.0715	3.7490	0.2734	
% Tall sedge	0.2160	4.4093	0.4313	
% Water	0.0568	17.0677	0.6942	
% <i>Spartina alterniflora</i> (short)	0.0549	121.4637	0.2949	
% <i>Spartina alterniflora</i> (tall)	0.0960	159.3828		
Maximum vegetation height	0.2432	6.1200		
Vegetation stem density	0.7219	3.2064		

VARIANCE INFLATOR FACTOR (VIF) = raiz quadrada (1 / (1 - R_j²))

Quinn & Keough (2002) = VIF 10

ZUUR et al (2009) VIF 3

Modelo Linear Generalizado

Relaciona a distribuição aleatória da variável dependente no experimento (a **função de distribuição**) com a parte sistemática (não aleatória) (ou **preditor linear**) através de uma função chamada **função de ligação**.

GLM consiste em três etapas:

- 1. Função de distribuição** - Uma hipótese sobre a distribuição da variável resposta Y_i . Isso também define a média e a variância de Y_i .
- 2. Preditor Linear** - especificação da parte sistemática. Esta é uma função das variáveis explicativas.
- 3. Função de Ligação** - A relação entre o valor médio de Y_i e da parte sistemática. Esta é também chamada de ligação entre a média e a parte sistemática

Modelo Linear Generalizado



Preditor Linear

*função
distribuição*

*função de
ligação*

variavel dependente ~ variaveis predictoras, family = poisson, link = log

<i>Family</i>	<i>Canonical Link</i>	<i>Range of Y_i</i>	<i>$V(Y_i \eta_i)$</i>
Gaussian	Identity	$(-\infty, +\infty)$	ϕ
Binomial	Logit	$\frac{0, 1, \dots, n_i}{n_i}$	$\frac{\mu_i(1 - \mu_i)}{n_i}$
Poisson	Log	$0, 1, 2, \dots$	μ_i
Gamma	Inverse	$(0, \infty)$	$\phi \mu_i^2$
Inverse-Gaussian	Inverse-square	$(0, \infty)$	$\phi \mu_i^3$

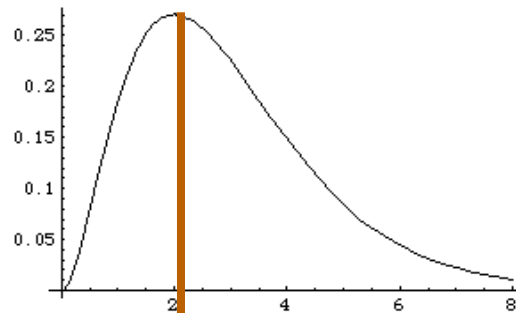
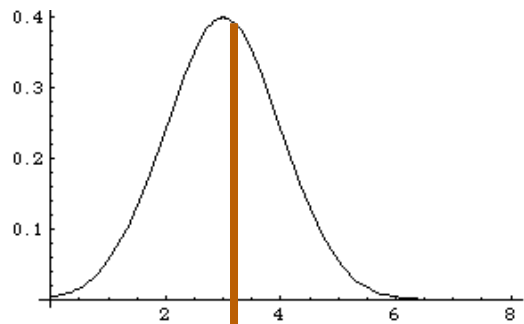
Maximum Likelihood Estimation (MLE)

Estimar os parâmetros dos modelos significa achar os parâmetros que fazem o modelo se ajustar melhor aos dados coletados.

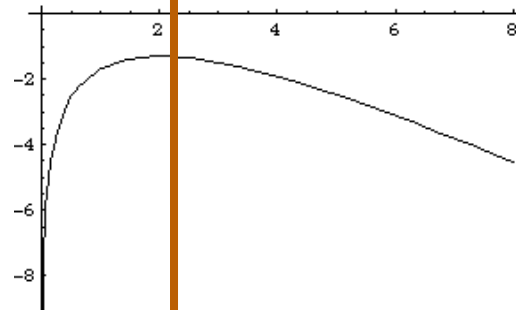
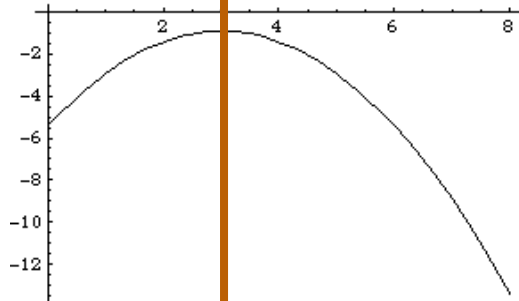
$$n_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{k1}$$

Máxima verossimilhança (*maximum likelihood estimates*) dos parâmetros estima aqueles valores dos parâmetros que fazem os dados observados mais prováveis de terem acontecido.

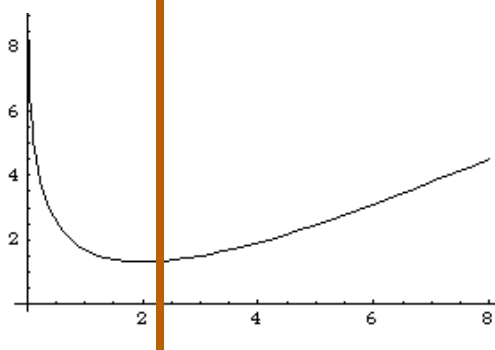
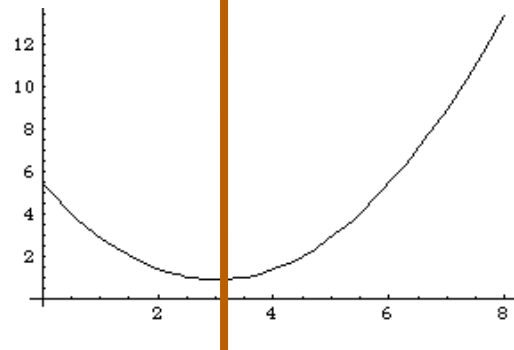
Maximum Likelihood Estimation (MLE)



**Função densidade
de probabilidade**



**Maximum
Likelihood**



Log likelihood

Likelihood Rate Test (LRT)

A LRT compara dois modelos aninhados, testando se os parâmetros aninhados do modelo mais complexo difere significativamente do valor nulo.

variavel dependente ~ A1 + A2 + A3, family = poisson, link = log

variavel dependente ~ A1 + A2, family = poisson, link = log

variavel dependente ~ A1, family = poisson, link = log

A diferença entre as *deviance* dos modelos apresenta uma distribuição chi-square com $p_1 - p_2$ graus de liberdade.

Analysis of Deviance Table

Model 1: TOT.N ~ OPEN.L + D.PARK

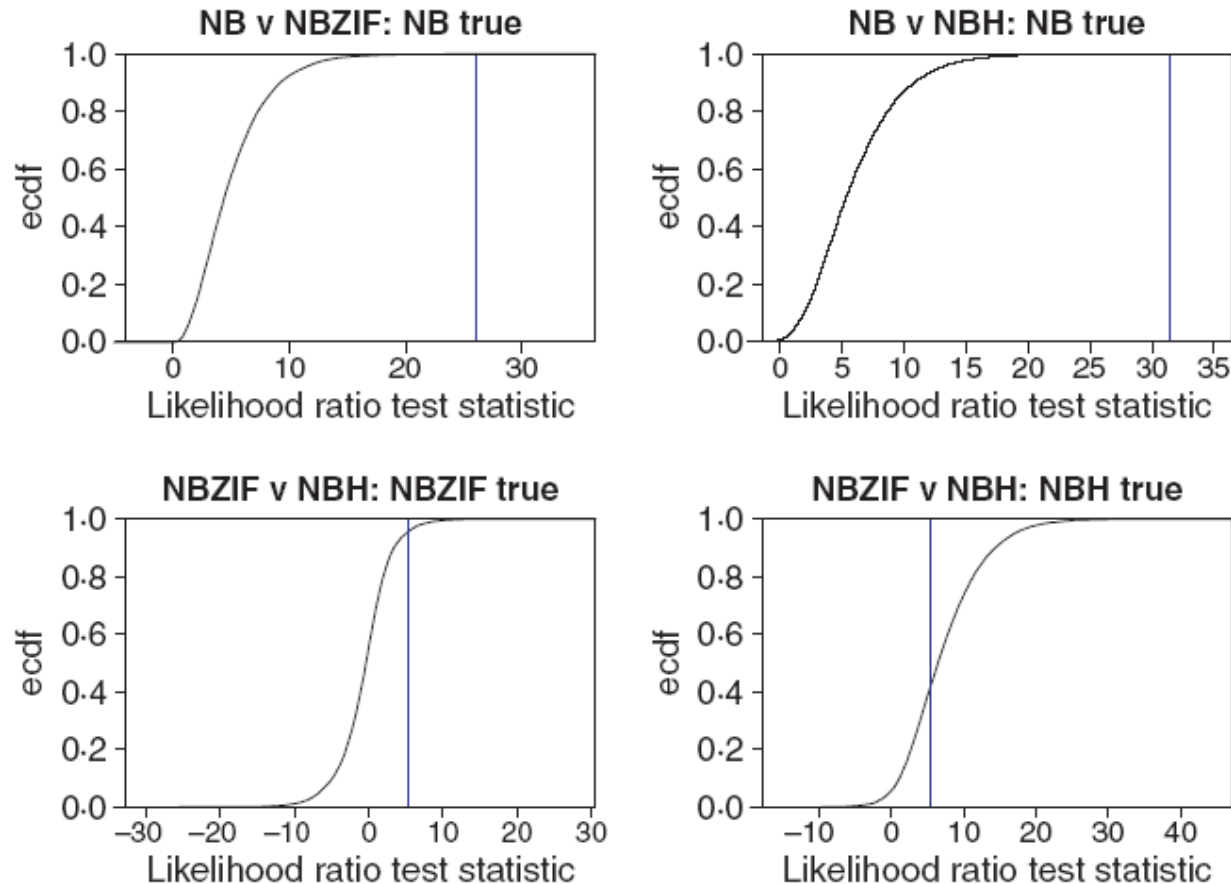
Model 2: TOT.N ~ OPEN.L

	Resid.	Df	Resid.	Dev	Df	Deviance	P(> Chi)
1		49		384.70			
2		50		994.65	-1	-609.95	< 2.2e-16 ***

A unified approach to model selection using the likelihood ratio test

Fraser Lewis^{1*}, Adam Butler² and Lucy Gilbert³

¹Scottish Agricultural College, King's Buildings, West Mains Road, Edinburgh EH9 3JG, UK; ²Biomathematics and Statistics Scotland, The King's Buildings, Edinburgh EH9 3JZ, Scotland, UK; and ³Macaulay Institute, Craigiebuckler, Aberdeen AB15 8QH, UK



REVIEW

Multimodel inference in ecology and evolution: challenges and solutions

C. E. GRUEBER, S. NAKAGAWA, R. J. LAWS & I. G. JAMIESON

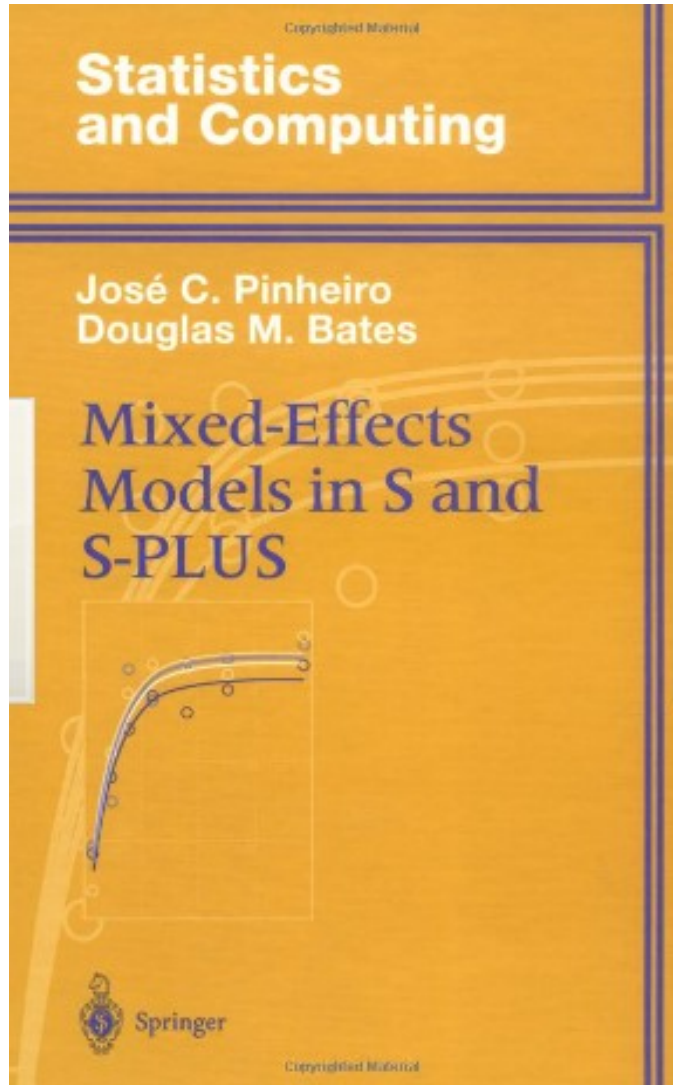
*Department of Zoology, University of Otago, Dunedin, New Zealand***Table 1** Information criteria for model selection.

Information criterion	Formula*	References
Akaike Information Criterion	$AIC = -2 \cdot \ln L + 2k$	Akaike (1973)
AIC – small sample size correction	$AIC_C = -2 \cdot \ln L + \frac{2k(k+1)}{n-k-1}$	Hurvich & Tsai (1989)
Quasi-AIC	$QAIC = \frac{-2 \ln L}{\hat{c}} + 2k$	Lebreton <i>et al.</i> (1992)
Conditional AIC	$cAIC = -2 \cdot \ln L + 2k_C$	Vaida & Blanchard (2005); Liang <i>et al.</i> (2008)
Bayesian Information Criterion	$BIC = -2 \cdot \ln L + k \ln(n)$	Schwarz (1978)
Deviance Information Criterion	$DIC = -2 \cdot \ln L + 2k_D$	Spiegelhalter <i>et al.</i> (2002)

Correção para o grau de liberdade

Correção para o número de parâmetros

Mixed – Effects Models



variavel dependente ~ variaveis predictoras ,
random = ~ 1 | variável

PODE LIDAR COM VARIÂNCIA DOS DADOS

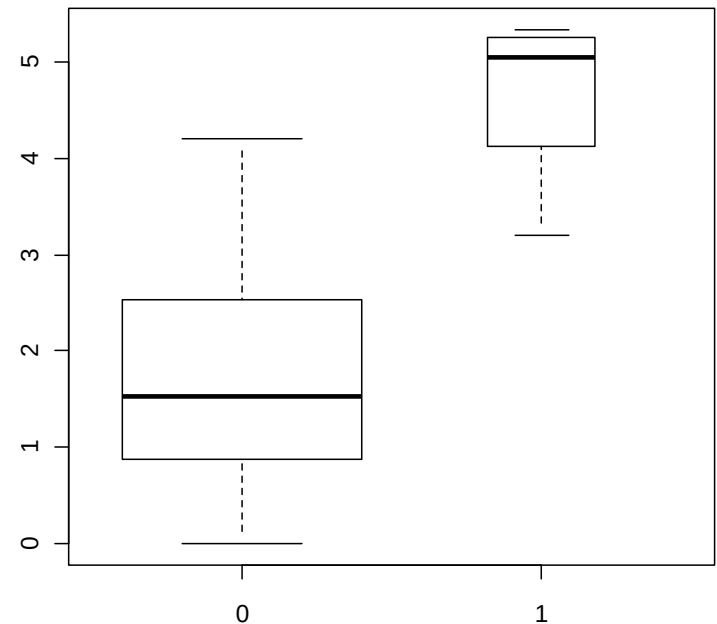
varIdent

varFixed

varExp

varPower

varConstPower



Mixed – Effects Models

variavel dependente ~ variaveis preditoras , random = ~ 1 | variável

**PODE LIDAR COM CORRELAÇÃO
TEMPORAL**

corCompSymm

AR-1

ARMA

**PODE LIDAR COM CORRELAÇÃO
ESPACIAL**

corSpher

corLin

corRatio

corGaus

corExp

Generalized linear mixed models: a practical guide for ecology and evolution

Benjamin M. Bolker¹, Mollie E. Brooks¹, Connie J. Clark¹, Shane W. Geange², John R. Poulsen¹, M. Henry H. Stevens³ and Jada-Simone S. White¹

¹ Department of Botany and Zoology, University of Florida, PO Box 118525, Gainesville, FL 32611-8525, USA

² School of Biological Sciences, Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand

³ Department of Botany, Miami University, Oxford, OH 45056, USA

Table I. Techniques for GLMM inferences, their advantages and disadvantages and the software packages that implement them

Method	Advantages	Disadvantages	Software
Wald tests (Z , χ^2 , t , F)	Widely available, flexible, OK for quasiliikelihood (QL)	Boundary issues; poor for random effects; t and F require residual df	GLIMMIX, NLMIXED (SAS), glmmPQL (R)
Likelihood ratio test	Better than Wald tests for random effects	Bad for fixed effects without large sample sizes; boundary effects; inappropriate for QL	NLMIXED (SAS), lme4 (R)
Information criteria	Avoids stepwise procedures; provides model weights and averaging; QAIC applies to overdispersed data	Boundary effects; no p value; requires residual df estimate for AIC_c	GLIMMIX, NLMIXED (SAS), lme4 (R)
Deviance information criterion	Automatically penalizes model complexity	Requires MCMC sampling	WinBUGS