

Princípios da modelagem ecológica e seleção de modelos

Leonardo Liberali Wedekin

LET – Laboratório de Ecologia Teórica

Instituto de Biociências - USP

Resumo

1. Uso de modelos na ecologia
2. Seleção de modelos
3. O modelo de regressão linear e suas extensões

Modelagem ecológica



“A natureza é um livro escrito em linguagem matemática”
Galileu Galilei

Modelagem ecológica

“Toda inferência na ecologia é feita através de algum modelo. Para alguma observação fazer sentido, todo mundo precisa de um modelo, independente se o observador sabe disto ou não”

(Kéry, 2010)

Inferência:

Obter conclusões estatísticas através de dados.

Fazer afirmações sobre uma população através de uma amostra.

“A modelagem na ciência permanece, no mínimo parcialmente, uma arte.”

(McCullagh & Nelder, 1989)

Modelagem ecológica

Algumas dicotomias (Bolker, 2008):

Teórico		Aplicado
Descritivo		Preditivo
Matemático		Estatístico
Analítico		Computacional
Dinâmico	X	Estático
Contínuo		Discreto
Determinístico		Estocástico
População		Indivíduo

Modelagem ecológica

Objetivos de um modelo incluem:

- *Previsão*

- *Explanação*

- *Generalização*

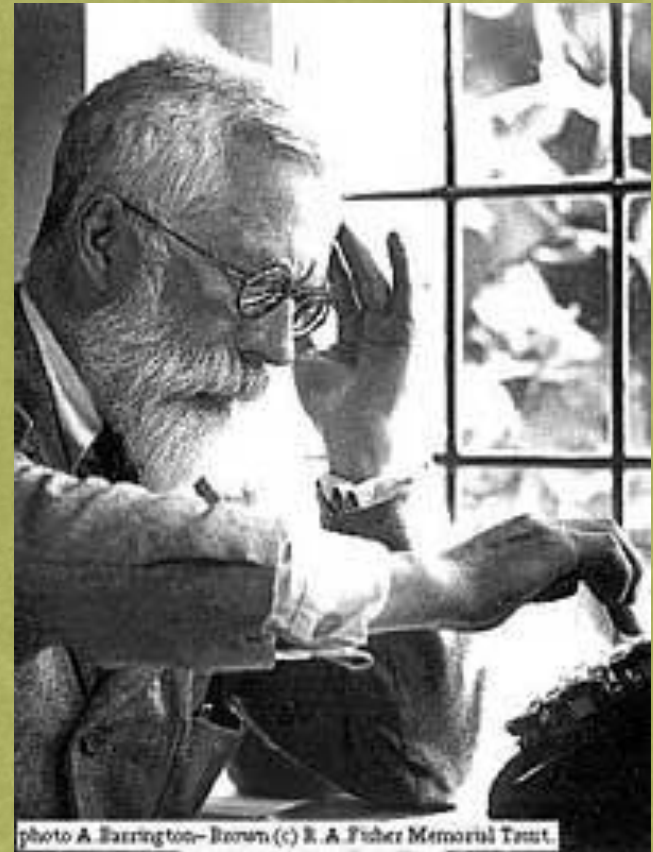
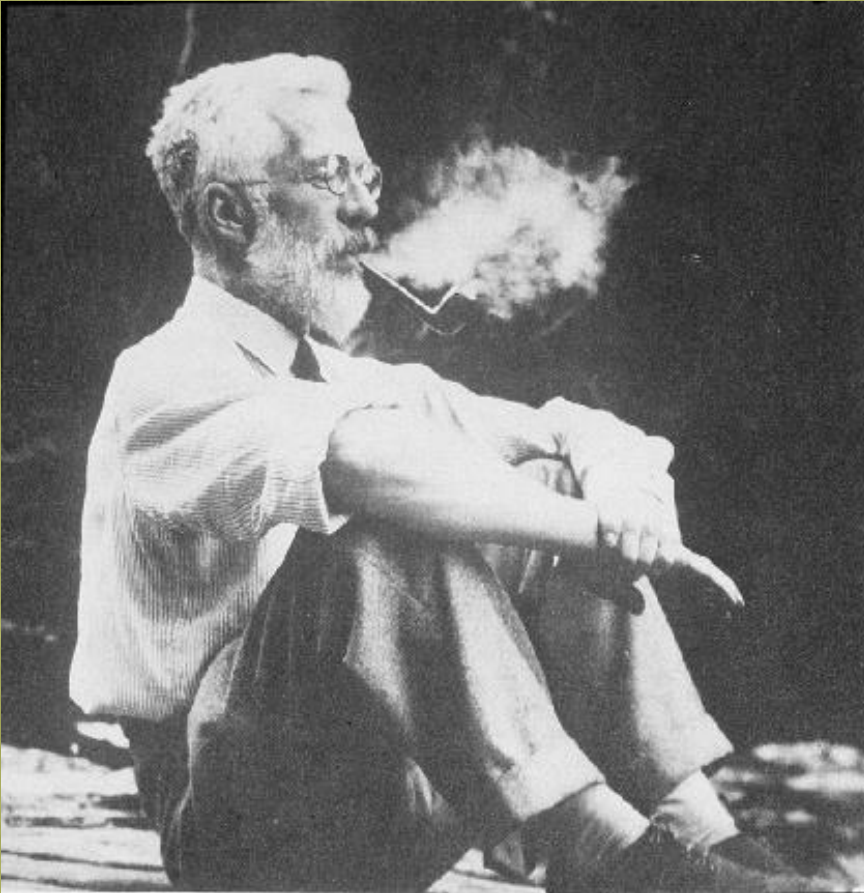
Modelagem ecológica

“... o objetivo dos métodos estatísticos é a **redução dos dados**. Uma quantidade de dados, que geralmente pelo seu grande tamanho é impossível de entrar na mente, é para ser substituído por relativamente poucas quantidades que devem representar adequadamente o todo, ou que, em outras palavras, deve conter o máximo possível, ou idealmente **toda a informação relevante contida nos dados originais**”

(Fisher, 1922)

Modelagem ecológica

Ronald A. Fisher (1890-1962)



PERGUNTA NECESSÁRIA

Qual modelo deve ser usado para melhor aproximar a realidade, baseado em dados de boa qualidade e relevantes para a questão?

(Burnham & Anderson, 2001)

TRÊS PRINCÍPIOS GERAIS GUIAM

(Burnham & Anderson, 2001)

- PARCIMÔNIA
- MÚLTIPLAS HIPÓTESES DE TRABALHO
- FORÇA DA EVIDÊNCIA

Modelagem ecológica

“quia frustra fit per plura quod potest fieri per pauciora”

“porque é em vão fazer com mais o que se pode fazer com menos”

William of Ockham, Inglaterra, Século XIV



Navalha de Ockham elimina
tudo que é desnecessário para
explicar um fenômeno

Múltiplas hipóteses de trabalho

Ao invés de uma única hipótese de trabalho (como no teste de hipóteses), trabalha-se com múltiplas hipóteses ou modelos biologicamente plausíveis

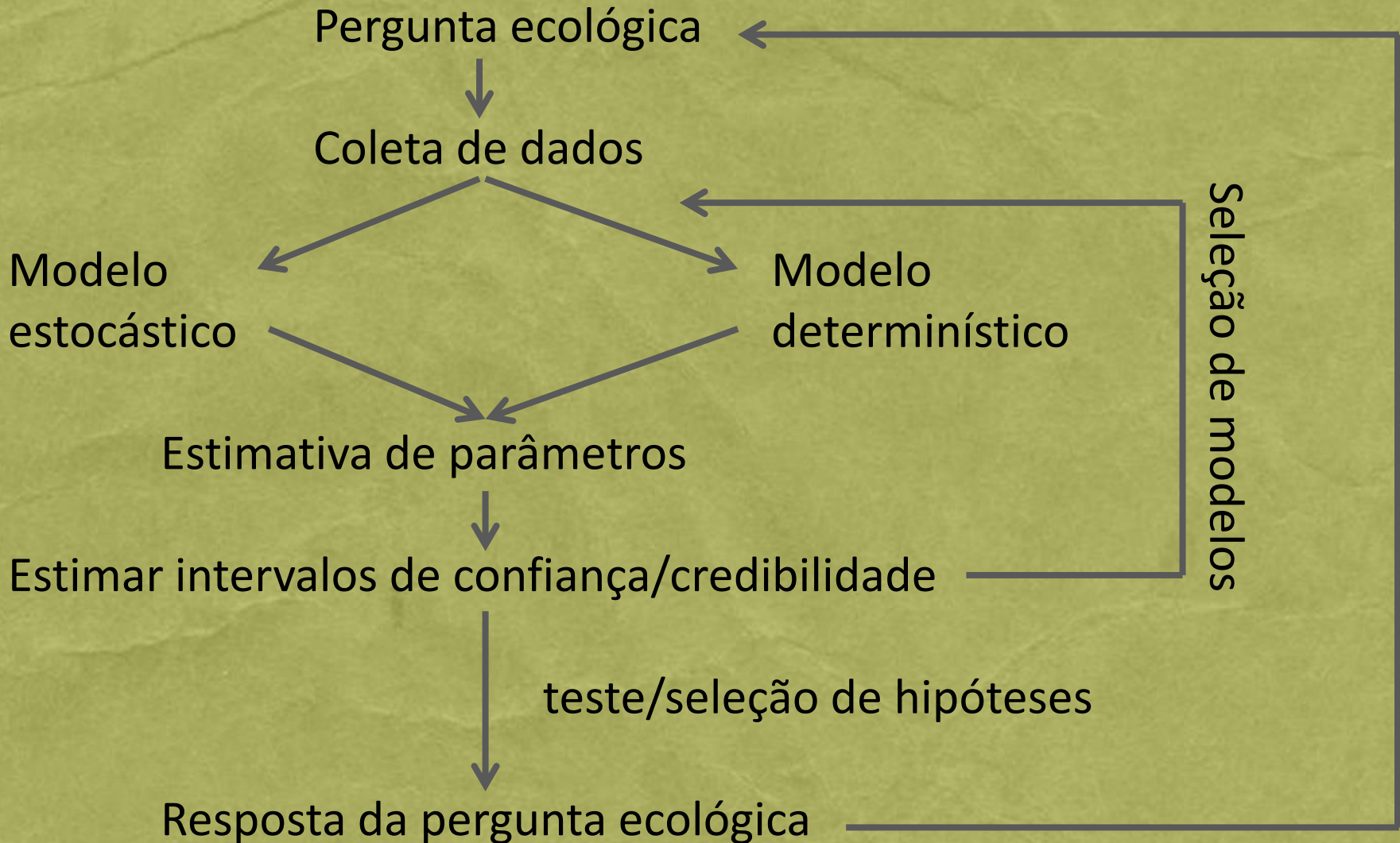
Múltiplas hipóteses de trabalho

Esta abordagem permite que hipóteses novas ou mais elaboradas sejam adicionadas continuamente nos modelos, enquanto hipóteses sem fundamento sejam gradualmente abandonadas.

Força de evidência

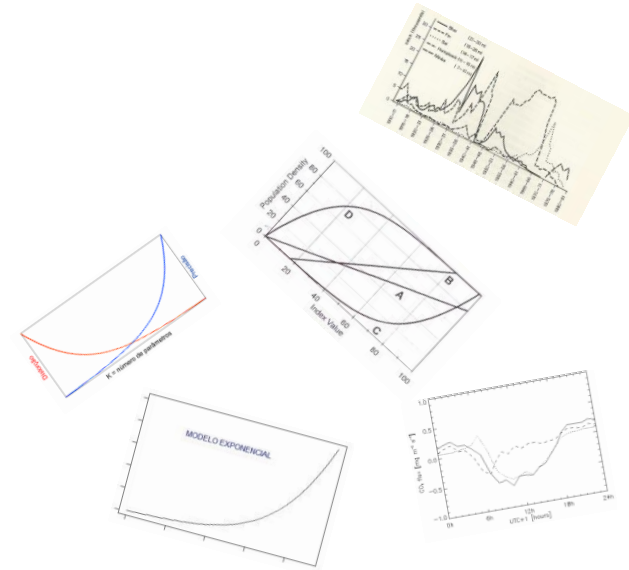
Não existe dicotomia, mas diferentes forças de evidência para cada hipótese / modelo

Modelagem ecológica




Modelagem ecológica

Evitar dragagem de dados...



Modelagem ecológica

Abordagens diferentes para inferência estatística
(Bolker, 2008):

- 
- Frequentista clássica
 - Verossimilhança
 - Bayesiana

Estimativa por máxima verossimilhança

Função de verossimilhança: calcula-se uma distribuição de probabilidades que é função dos parâmetros condicionada aos dados.

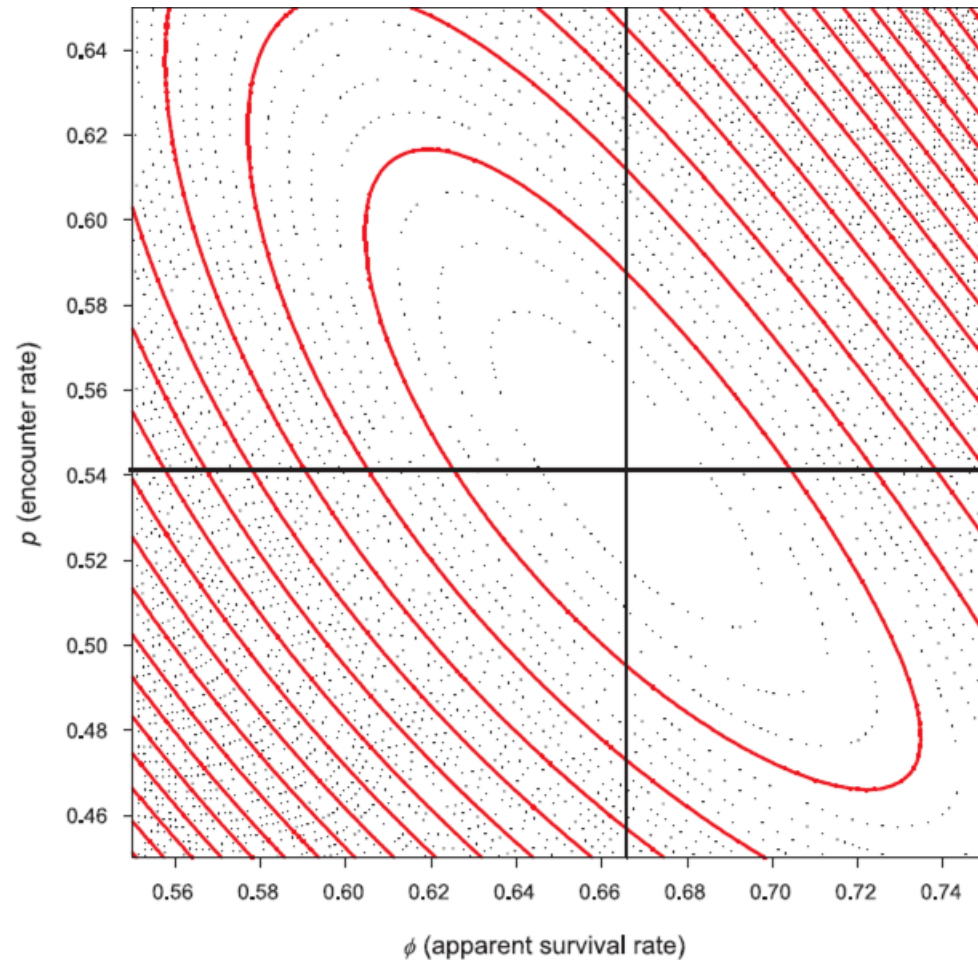
Modelagem ecológica

Estimativa por máxima verossimilhança

Dado o modelo, para quais valores de parâmetros os dados são mais prováveis?

Os valores que maximizam a função de verossimilhança são as melhores estimativas para os parâmetros.

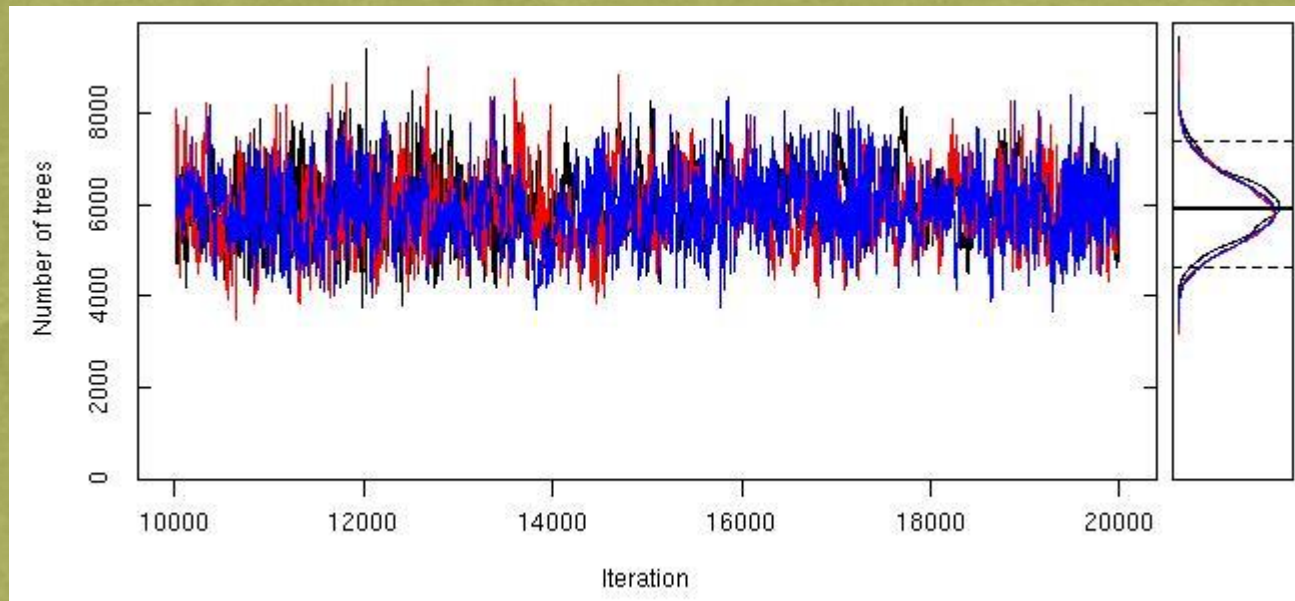
Modelagem ecológica



Modelagem ecológica

Estimativa por métodos Bayesianos

MCMC – Markov Chain Monte Carlo



2. Seleção de modelos

- Nos métodos atuais grande ênfase é dada à seleção de modelos, sendo um aspecto crítico da análise de dados;
- Estimativa dos parâmetros assume uma menor importância relativa, enquanto identificar processos biológicos significantes na área de estudo comparando diferentes modelos assume uma maior importância.

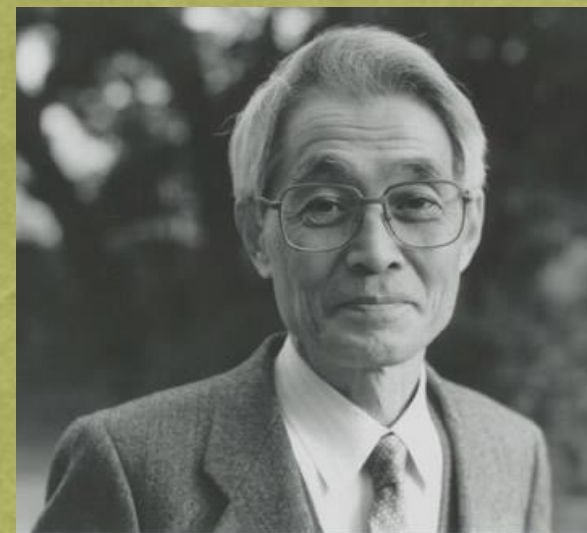
Seleção de modelos

- S. Kullback & R. A. Leibler (1951) descreveram uma medida de distância entre a realidade conceitual (f) e um modelo que aproxima esta realidade (g) – distância Kullback-Leibler (K-L)
 - $I(f, g)$ = informação perdida quando um modelo é usado para aproximar a realidade ou distância entre o modelo e a realidade

Seleção de modelos

Estabeleceu a relação entre a distância K-L e a máxima verossimilhança dentro de uma abordagem de otimização (Akaike, 1974).

Hirotsugu Akaike
(1927 - 2009)



Critério de Informação de Akaike

$$\text{AIC} = -2 \log(\mathcal{L}(\hat{\theta} \mid \text{data})) + 2K.$$



Valor da log
verossimilhança

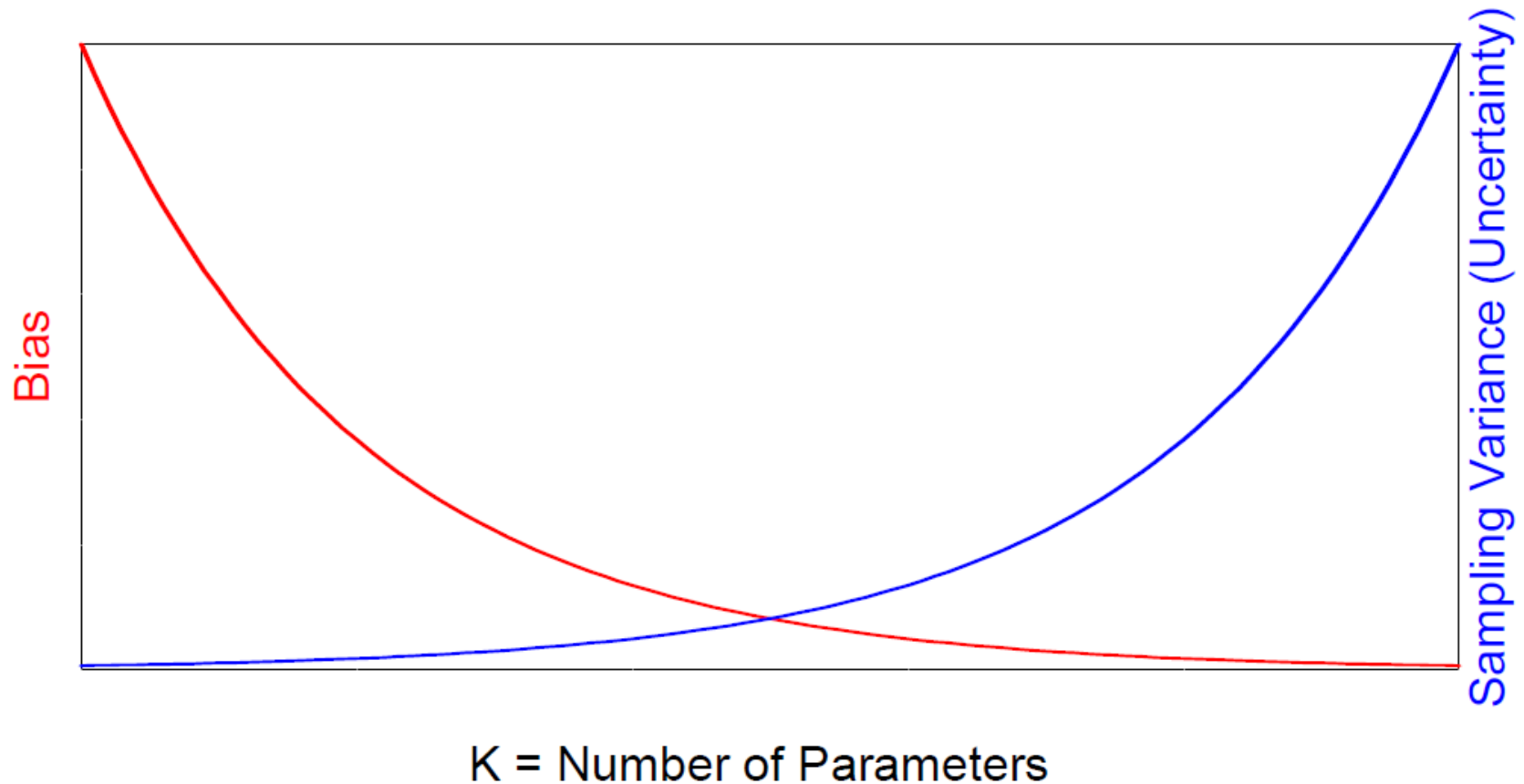


Número de
parâmetros

Critério de informação não é teste...

Não existem conceitos como: poder do teste, p-valor ou significância.

Seleção de modelos



Seleção de modelos

- Quanto menor o AIC, menor a distância entre o modelo e a realidade
- AIC pode ser computado para cada modelo
- Assim, devemos comparar e escolher o modelo com menor valor de AIC

Delta AIC

$$\Delta_i = \text{AIC}_i - \min \text{AIC}$$

- Valores de AIC re-escalados de forma que o modelo com menor AIC passe a ter $\text{AIC} = 0$
 - $\Delta_i \leq 2$: suporte substancial ao modelo
 - $\Delta_i = 4 - 7$: menor suporte considerável ao modelo
 - $\Delta_i > 10$: sem suporte

AICc

- Correção para modelos com amostra pequena relativa ao número de parâmetros
- Como AIC e AICc convergem com amostras grandes **AICc sempre deve ser usado**

Peso de Akaike (*Akaike weight*)

- Indica a força de evidência de um modelo
- Pode ser interpretado com a probabilidade daquele modelo ser o melhor dentre os modelos candidatos

$$w_i = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{r=1}^R \exp(-\frac{1}{2}\Delta_r)}$$

Inferência multi-modelos (*model averaging*)

Inferência é feita a partir do conjunto de modelos ponderando as estimativas pelos pesos de Akaike de cada modelo

Teste de razão de verossimilhança **(*Likelihood Ratio Test* - LRT)**

Teste de hipótese para modelos aninhados, ou seja, um modelo contém o outro, ou é um caso simplificado do outro

Seleção de modelos

Teste de razão de verossimilhança (LRT)

$\ln L(\text{parâmetros de um modelo geral})$

vs.

$\ln L(\text{parâmetros de um modelo reduzido})$

Tamanho da diferença entre log-verossimilhanças indica se o modelo reduzido deve ser preferido

Seleção de modelos

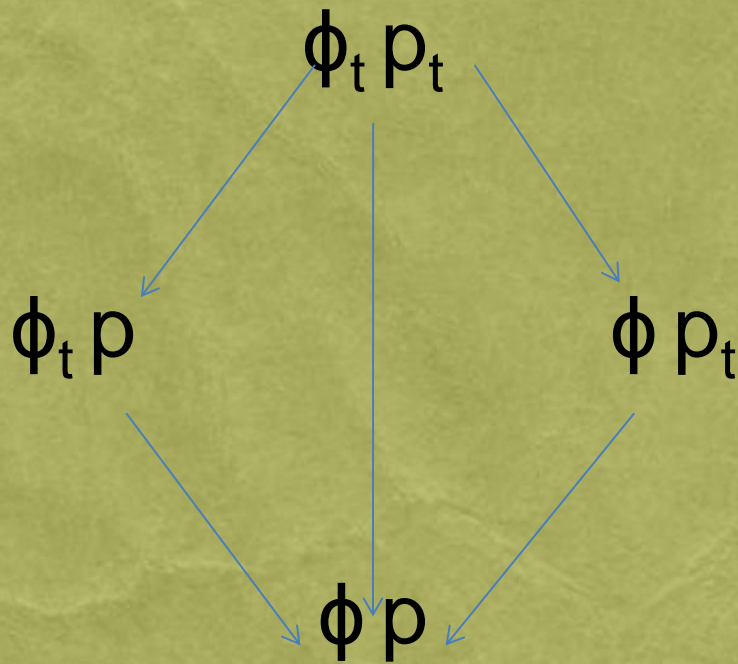
Teste de razão de verossimilhança (LRT)

Diferença possui distribuição de Qui-Quadrado com graus de liberdade igual à diferença de parâmetros entre os dois modelos

Diferença significativa indica que modelo geral deve ser preferido, enquanto diferença não significativa indica que o modelo mais simples deve ser escolhido

Seleção de modelos

Teste de razão de verossimilhança (LRT)



3. Modelo linear e suas extensões

$$y = \alpha + \beta x$$

Modelo linear clássico

$$Y = \alpha + \beta X + \varepsilon$$

Onde:

α = intercepto / onde a linha cruza o eixo y

β = coeficiente regressão / inclinação (+ ou -)

Y = variável resposta/dependente

X = variável explicatória/independente

ε = erro residual

Premissas

- Relação linear entre X e Y
- Normalidade
- Variância constante
- Independência

Regressão múltipla

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_i X_i + \varepsilon$$

Onde:

β_0 = intercepto / onde a linha cruza o eixo y

β_i = coeficiente regressão / inclinação (+ ou -)

Y = variável resposta/dependente

X_i = variável explicatória/independente

ε = erro residual

Modelo linear generalizado (GLM)

$$E(Y) = f(\beta_0 + \beta_1 X)$$

Onde:

E = distribuição assumida da variável resposta (estrutura do erro)

$f(z)$ = função de ligação (*link function*)

$(\beta_0 + \beta_1 X)$ = componente linear

> Permite outras distribuições de erro relaxando premissas do modelo linear clássico e conferindo flexibilidade ao modelo

Modelo linear e suas extensões

Funções de ligação (*link function*)

- Variável resposta assume outras distribuições da família exponencial, como por exemplo:
 - Números positivos e inteiros (Poisson)
 - Resposta binária 0 ou 1 (Binomial)

Error	Canonical link
normal	<i>identity</i>
poisson	<i>log</i>
binomial	<i>logit</i>
Gamma	<i>reciprocal</i>

Modelo linear / ANOVA

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Onde:

Y = variável resposta contínua

β_0 = intercepto

β_1 = inclinação da reta

x = variável código (*dummy*) = 0 ou 1

ε = erro residual

Modelos fatoriais

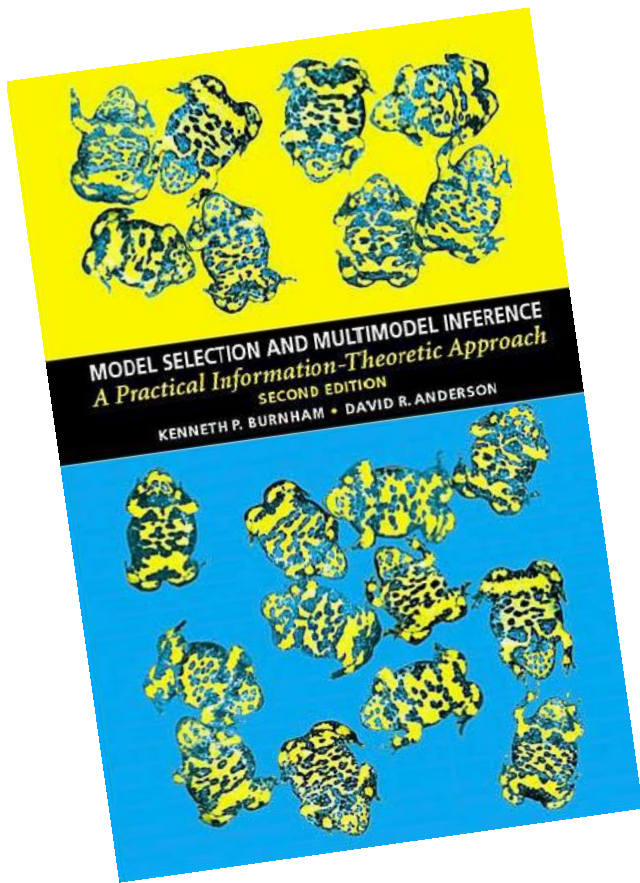
Matriz de desenho

$$\mathbf{y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1k} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2k} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1k} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{2k} \end{bmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Modelo aditivo generalizado (GAM)

- Extensão não paramétrica dos GLM;
- O que determina o formato da função são os dados e não funções ou distribuições específicas definidas *a priori*;
- Vários métodos de “suavização” (*smooth*) e grau de suavização (graus de liberdade).

Leituras adicionais



Burnham & Anderson (2002)

Leituras adicionais



Review

TRENDS in Ecology and Evolution Vol.19 No.2 February 2004

Full text provided by www.sciencedirect.com



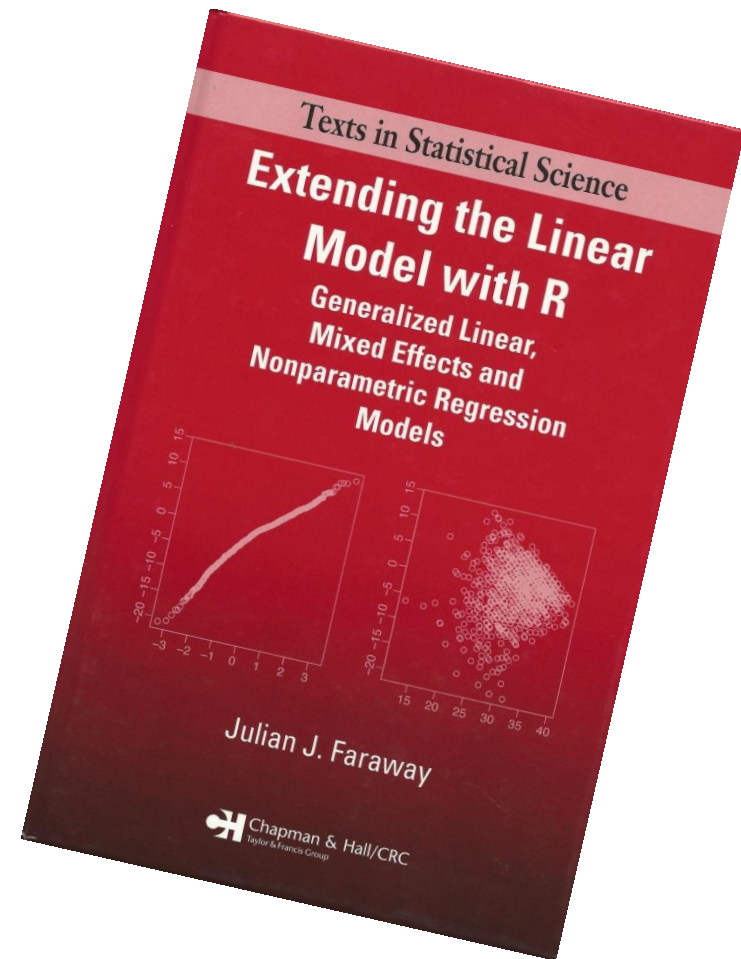
Model selection in ecology and evolution

Jerald B. Johnson¹ and Kristian S. Omland²

¹Conservation Biology Division, National Marine Fisheries Service, 2725 Montlake Boulevard East, Seattle, WA 98112, USA

²Vermont Cooperative Fish & Wildlife Research Unit, School of Natural Resources, University of Vermont, Burlington, VT 05405, USA

Leituras adicionais



Faraway (2006)