Seleção de modelos

Paulo Inácio Prado e João L.F. Batista

BIE5781 - Pós-Graduação em Ecologia USP

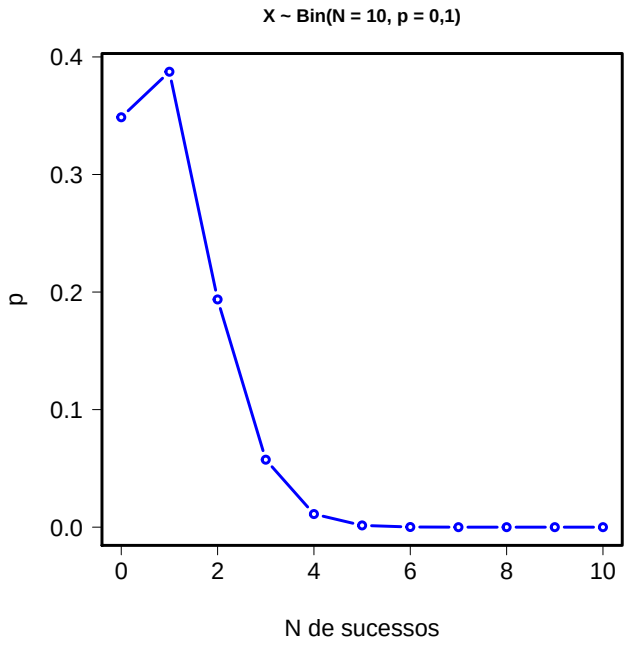
Novembro de 2020

Objetivos

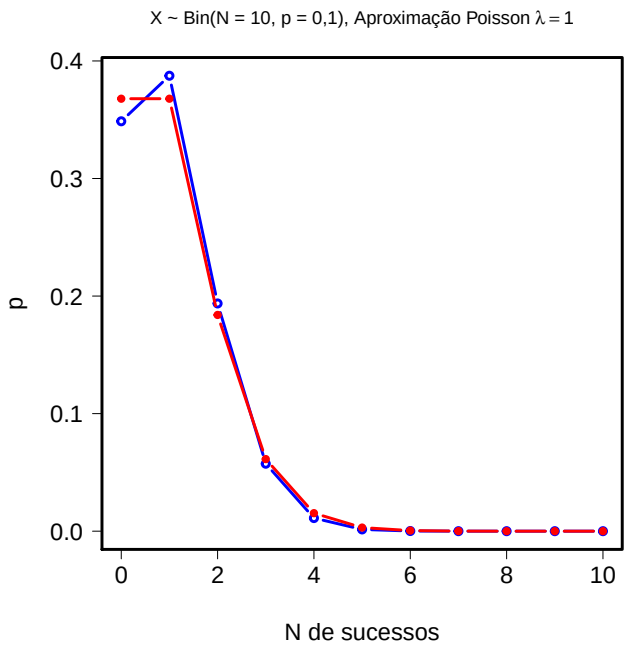
Os objetivos desta aula são:

1. Definir AIC e medidas correlatas;
2. Apresentar os fundamentos teóricos do AIC e seus contextos de uso
3. Exemplificar o uso dessas medidas em seleção de modelos e variáveis

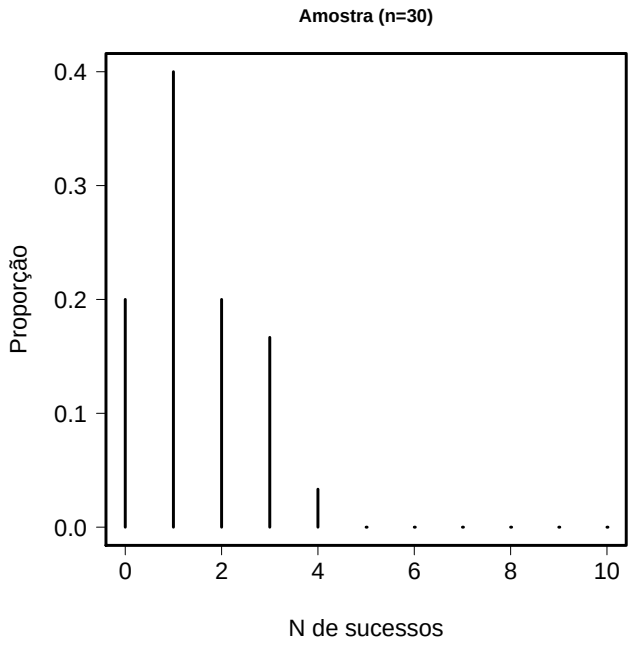
Modelos como aproximações



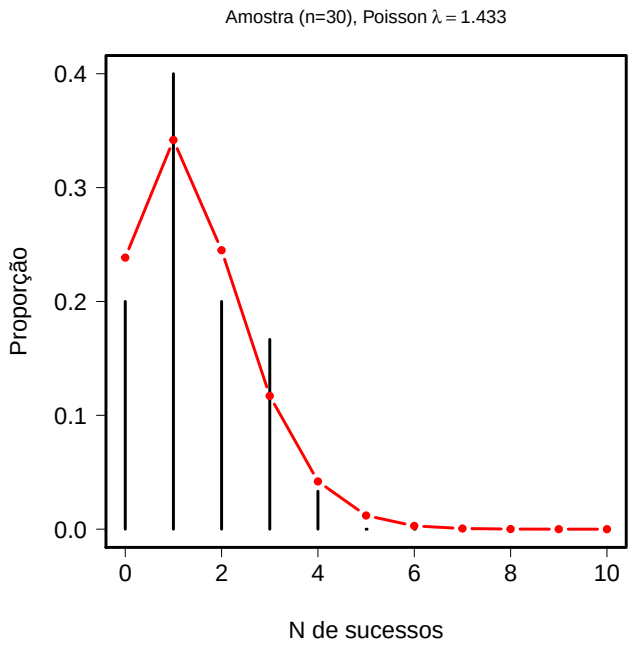
Modelos como aproximações



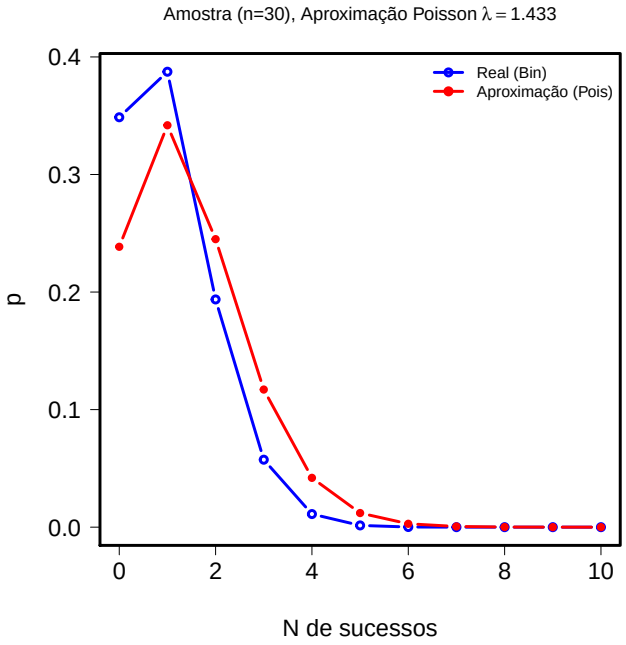
Na vida real aproximação é feita de uma amostra



Na vida real aproximação é feita de uma amostra



Na vida real aproximação é feita de uma amostra



As duas incertezas que causam divergência entre modelos estatísticos

Incerteza do Modelo de Aproximação

- Não conhecemos o modelo "real", temos apenas uma amostra.
- Portanto todos os modelos ajustados à amostra devem ser considerados aproximações a um modelo desconhecido.
- O que sabemos sobre este modelo desconhecido? O que a amostra nos revela.

As duas incertezas que causam divergência entre modelos estatísticos

Incerteza de Estimação

- Conhecemos os modelos de aproximação
- Mas não conhecemos *o valor dos parâmetros* da melhor aproximação que cada um pode dar.
- O valor dos parâmetros dos modelos são *estimados* da amostra.

O Critério de Informação de Akaike: uma medida de divergência ou perda de informação

Em relação aos dados

Perda de informação ou discrepância no ajuste do modelo proposto aos dados.

Assintoticamente em relação ao modelo real

Perda de informação ou discrepância no ajuste do modelo proposto ao modelo que gerou os dados.

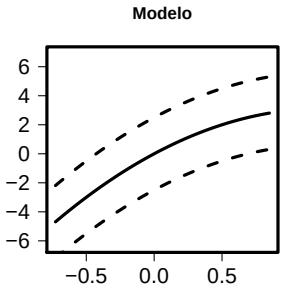
Viés de estimação

Log-verossimilhança é estimador enviesado de divergência

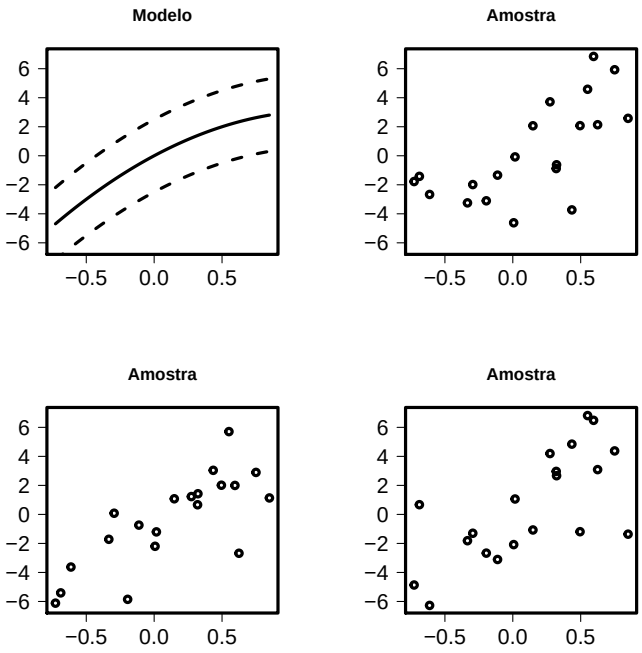
- Akaike percebeu que quando as duas amostras:
  - para representar o modelo verdadeiro
  - para estimar os parâmetros
- Eram a *mesma amostra*, então a máxima log-verossimilhança era uma estimativa viciada da divergência ao modelo real.
- O vício era *positivo* e proporcional ao número de parâmetros do modelo (K):
- Estimativa corrigida:

$$\mathbf{L}\{g(\,\widehat{\theta}\,|\,x)\} - K, \qquad \widehat{\theta} = \text{ MLEs}$$

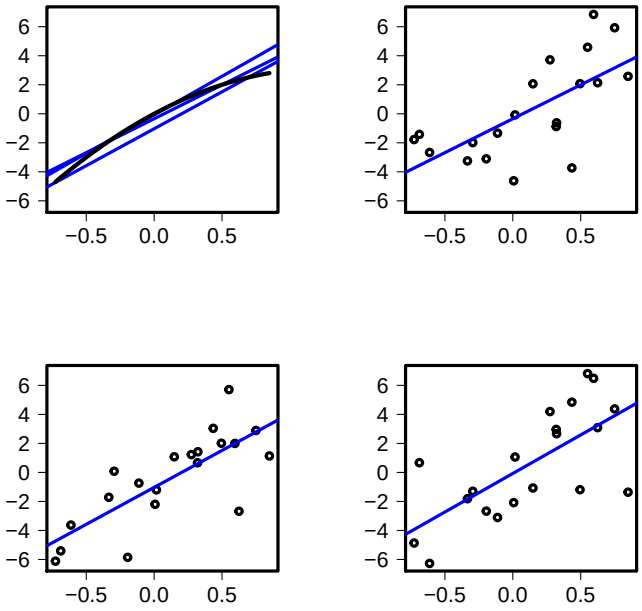
Por que penalizar parâmetros: o viés da mesma amostra



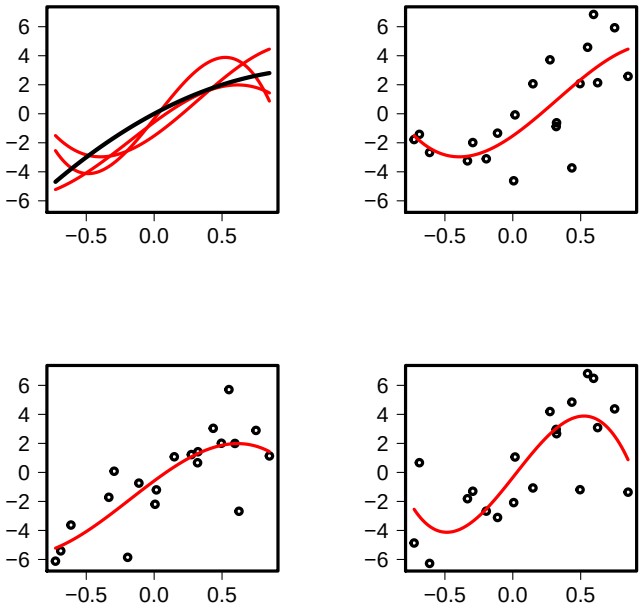
Por que penalizar parâmetros: o viés da mesma amostra



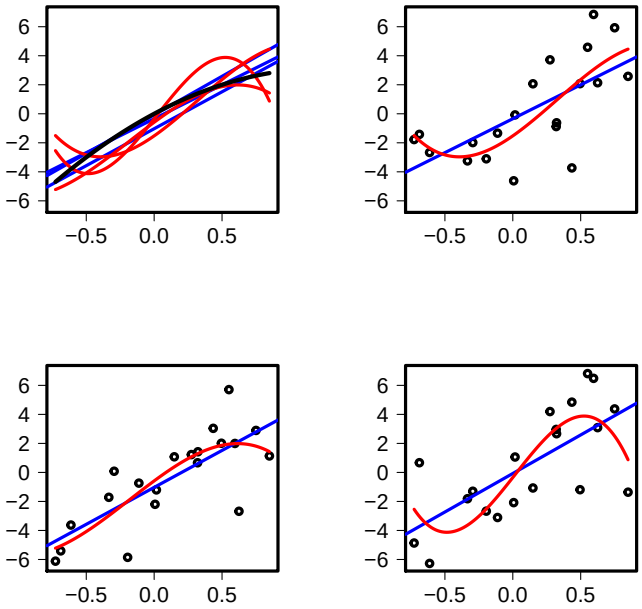
Por que penalizar parâmetros: o viés da mesma amostra



Por que penalizar parâmetros: o viés da mesma amostra



Por que penalizar parâmetros: o viés da mesma amostra



Definição do AIC

$$\text{AIC} = -2\,\mathbf{L}\{g(\widehat{\theta}|x)\} + 2\,K$$

Onde:

$$\begin{aligned} \widehat{\theta} &= \text{MLEs} \\ K &= \text{número de parâmetros} \\ \mathbf{L}\{g(\widehat{\theta}|x)\} &= \text{função de log-verossimilhança} \end{aligned}$$

ΔAIC

- O AIC é uma medida de distância (ou de perda de informação) **relativa**
- Em um conjunto de modelos, o de menor AIC será o mais plausível.
- Regra canônica: modelos com diferenças de AIC ≤ 2 são igualmente plausíveis.
- Para facilitar a comparação, calculamos o ΔAIC:

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

- O modelo mais plausível terá $\Delta_i = 0$

Correção para amostras pequenas

$$\text{AICc} = -2\,\mathbf{L}\{g(\widehat{\theta}|x)\} + 2\,K\left(\frac{n}{n-K-1}\right)$$

Onde:

- $n =$ n de observações
- amostras pequenas: $n/K < 40$

Akaike weights

$$w_i = \frac{e^{-1/2\Delta_i}}{\sum e^{-1/2\Delta_i}}$$

- Expressam o suporte relativo de cada modelo, em uma escala que soma um.
- Dependem do conjunto de modelos comparados:
  - se modelos são incluídos ou retirados os pesos mudarão;
  - os pesos tendem a ser menores se há muitos modelos na comparação.

Notas importantes sobre seleção de modelos

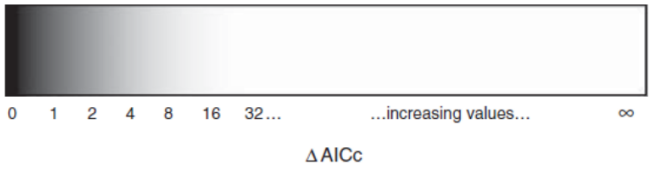
- Empates acontecem: indicam que os dados não contêm evidência suficiente para identificar o melhor modelo;
- seleção de modelos não é um teste estatístico;
- seleção de modelos não mede a qualidade do ajuste;
- a seleção de modelos está restrita aos modelos comparados;
- o AIC não pode ser usado para comparar modelos ajustados a conjuntos de dados diferentes;
- pela mesma razão, o AIC não pode ser usado para comparar modelos ajustados a dados transformados e não transformados.

O que é um suporte conclusivo?

Livro de Burnham & Anderson 2008, p.70

Δ_i	Level of Empirical Support of Model i
0-2	Substantial
4-7	Considerably less
> 10	Essentially none.

Burnham *et al.* Behav Ecol Sociobiol (2011) 65:23–35



Seleção de modelos no R

Funções do pacote bbmle

```
> m1 <- lm(y1 ~ x)
> AIC(m1)

[1] 97.58727

> AICc(m1, nobs=length(y1))

[1] 99.08727
```

Seleção de modelos no R

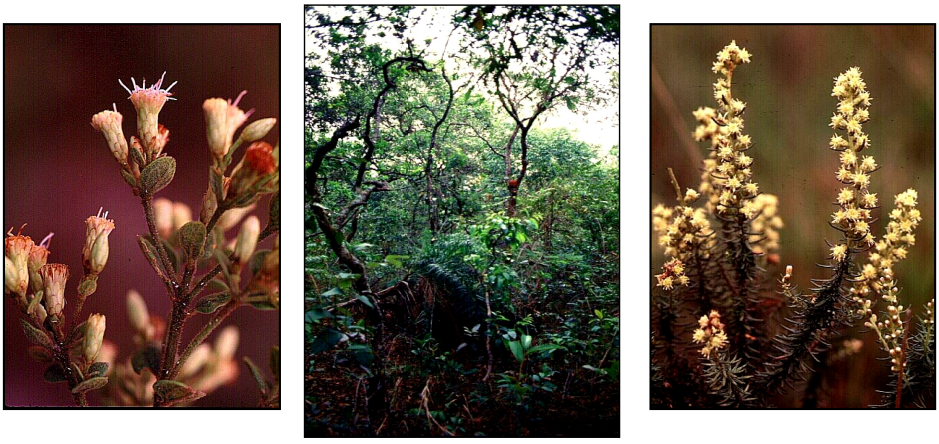
Funções do pacote bbmle

```
> m2 <- lm(y1 ~ x + I(x^2))
> m3 <- lm(y1 ~ x + I(x^2) + I(x^3))
> AICctab(m1, m2, m3, nobs=length(y1),
+         base=T, weights=T)

   AICc  dAICc df weight
m1  99.1   0.0 3  0.45
m2  99.1   0.1 4  0.44
m3 101.9   2.8 5  0.11
```

Exemplo de Seleção de modelos

Asteráceas x capins invasores no cerrado de SP, Almeida-Neto et al. *Plant Ecology* 2010



Exemplo de Seleção de modelos

Pergunta

A ocupação de áreas de cerrado por capins invasores afeta a diversidade de Asteráceas nativas no cerrado?

Hipóteses

- Redução gradual da diversidade
- Limiar de perda da diversidade
- Perturbação intermediária

Exemplo de Seleção de modelos

Dados

- 30 áreas de cerrado no estado de SP
- 45 parcelas de 35 x 5 m por área:
  - Contagem de indivíduos de cada espécie de Asteraceae nos transectos
  - Cobertura de gramíneas invasoras nos transectos

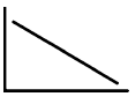
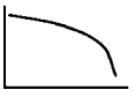
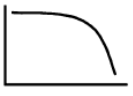
Variáveis

Resposta : número de plantas da família Asteraceae nos transectos (especialistas e generalistas)

Preditora : índice de cobertura de gramíneas exóticas nos transectos

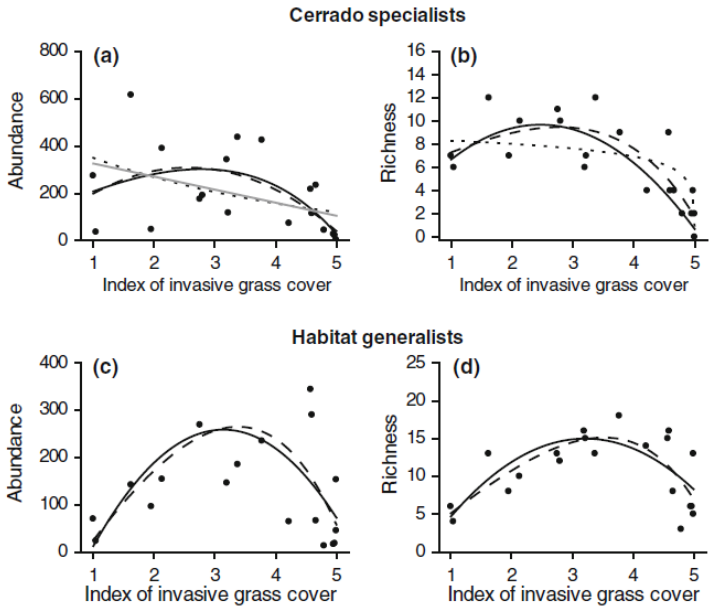
Exemplo de Seleção de modelos

Modelos concorrentes

Model	Abbreviation	Function	Outline graph
Linear	Linear	$y = a - bx$	
Negative log	Neg-log	$y = \log(a - bx)$	
Exponential decrease from an asymptote	Exp-decr	$y = a - e^{b.x}$	
Quadratic	Quadr	$y = a + bx + cx^2$	
Asymmetric hump	Asym-hump	$y = a + bx - e^{c.x}$	

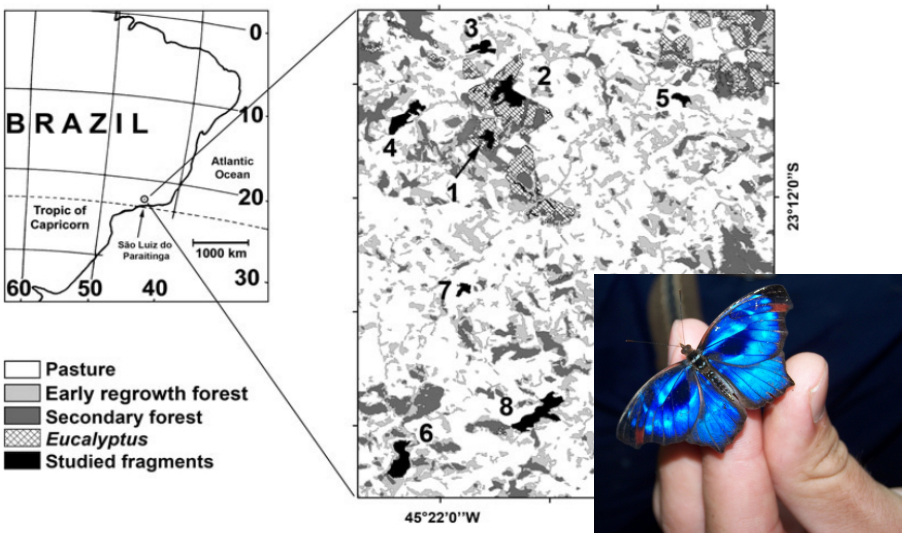
Exemplo de Seleção de modelos

Modelos selecionados



Exemplo de Seleção de variáveis

Efeito da paisagem sobre abundância de borboletas em fragmentos florestais, Ribeiro et al. *Biodiv Cons* 2012



Exemplo de Seleção de variáveis

Pergunta

Em qual escala a cobertura da matriz da paisagem afeta mais a abundância de borboletas que usam manchas de floresta em uma paisagem altamente fragmentada?

Dados

- 10 fragmentos florestais em São Luis do Paraitinga (SP)
- Borboletas frugívoras capturadas em armadilhas
- Classificação da cobertura vegetal da paisagem circundante com imagem de satélites

Exemplo de Seleção de variáveis

Variáves

Respostas :

- número capturas de borboletas de quatro subfamílias, por fragmento;
- número de espécies de borboletas capturadas, por fragmento.

Preditoras :

- proporção ocupada por pastagens, reflorestamento, floresta e capoeiras em áreas circulares centradas em cada fragmento;
- : raio da área circular em torno de cada fragment(100 a 2.000 m).

Exemplo de Seleção de variáveis

Hipóteses de menos, modelos demais!

- 4 classes de cobertura $\times$ 11 raios = 44
- $2^{44} = 1.8 \times 10^{13}$ modelos possíveis, só de efeitos aditivos !

Exemplo de Seleção de variáveis

Solução: um modelo com efeito de cada cobertura em cada escala, e um de ausência de efeitos

$$
\begin{aligned}
N_{capturas} &\sim \text{floresta}_{100m} \\
N_{capturas} &\sim \text{pasto}_{100m} \\
&\dots \\
N_{capturas} &\sim \text{pasto}_{2000m} \\
N_{capturas} &\sim 1
\end{aligned}
$$

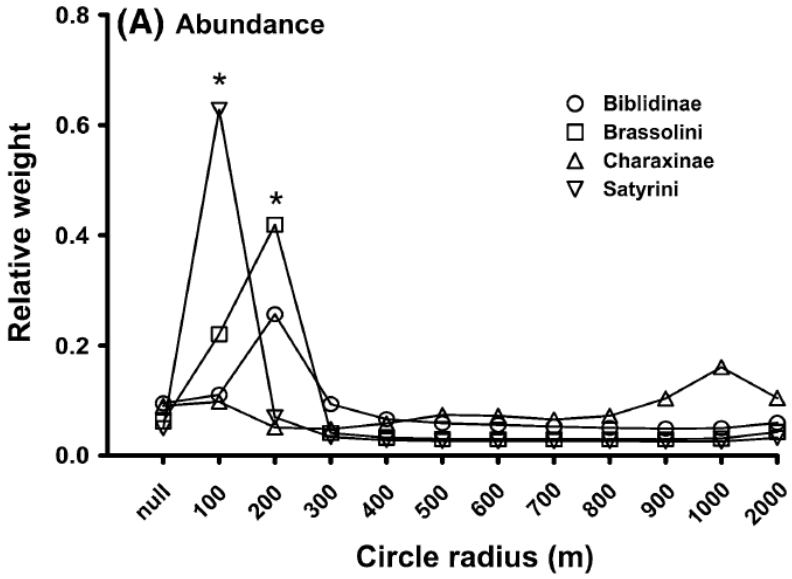
Exemplo de Seleção de variáveis

45 modelos de uma preditora cada

4 classes de cobertura $\times$ 11 escalas + modelo nulo

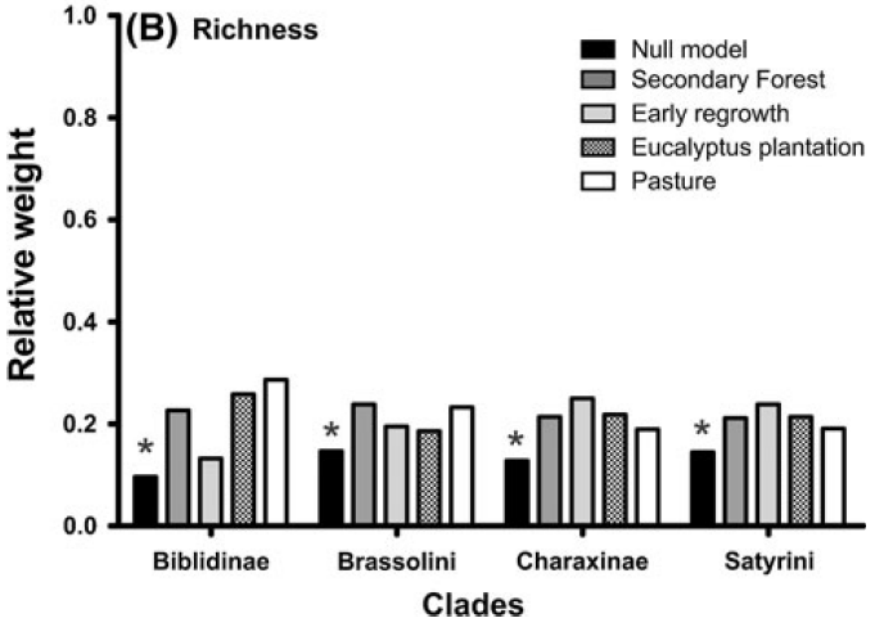
Exemplo de Seleção de variáveis

Raios: soma dos pesos de evidência



Exemplo de Seleção de variáveis

Classes de cobertura: somas dos pesos de evidência



Para saber mais leia Burham & Anderson 2010