Introdução aos Modelos Lineares em Ecologia

Prof. Adriano Sanches Melo - Dep. Ecologia – UFG
asm.adrimelo no gmail.com

Página do curso: www.ecologia.ufrgs.br/~adrimelo/lm/

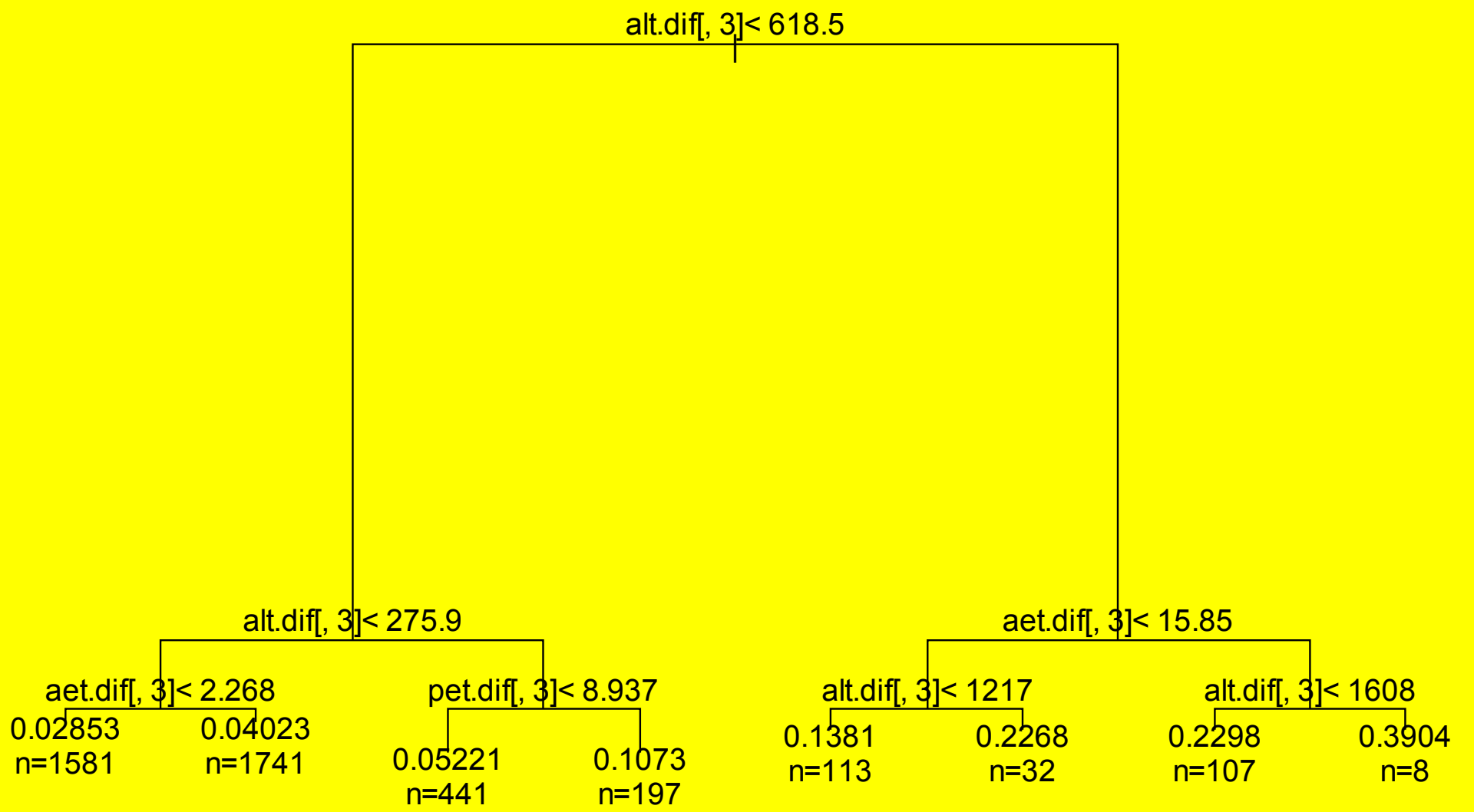
Livro-texto: Crawley, M.J. 2005. Statistics: An Introduction using R. John Wiley & Sons.

Página do livro na internet:
http://www3.imperial.ac.uk/naturalsciences/research/statisticsusingr

Árvore de Regressão

Exemplo Crawley aula anterior

Outro exemplo: Diversidade Beta de Aves nas Américas e 16 preditores



Árvore de Regressão

-- 2 pacotes no R: rpart e tree (em Mac existe um terceiro)

Uso geral:

```
> modelo<-rpart(y~x1+x2+x3)
> modelo
n= 4220
node), split, n, deviance, yval
      * denotes terminal node
 1) root 4220 11.168180000 0.04973464
   2) alt.dif[, 3]< 618.5273 3960  3.427418000 0.04022913
     4) alt.dif[, 3]< 275.9421 3322  1.442206000 0.03466060
       8) aet.dif[, 3]< 2.268313 1581  0.501331700 0.02853162 *
       9) aet.dif[, 3]>=2.268313 1741  0.827553200 0.04022632 *
     5) alt.dif[, 3]>=275.9421 638  1.345839000 0.06922387
      10) pet.dif[, 3]< 8.937062 441  0.392178300 0.05220608 *
      11) pet.dif[, 3]>=8.937062 197  0.540043000 0.10731950 *
   3) alt.dif[, 3]>=618.5273 260  1.933306000 0.19451100
     6) aet.dif[, 3]< 15.85458 145  0.803425300 0.15764720
      12) alt.dif[, 3]< 1216.863 113  0.402907000 0.13805000 *
      13) alt.dif[, 3]>=1216.863 32  0.203873800 0.22684960 *
     7) aet.dif[, 3]>=15.85458 115  0.684384300 0.24099140
      14) alt.dif[, 3]< 1607.963 107  0.487356300 0.22982370 *
      15) alt.dif[, 3]>=1607.963 8  0.005195071 0.39036010 *
```

```
> plot(modelo)
> text(modelo)
> summary(modelo)
Call:
rpart(formula = beta.sor.add.aves[, 10] ~ alt.dif[, 3] + temp.dif[,3] + npp.dif[, 3]
+ aet.dif[, 3] + pet.dif[, 3] + preci.dif[,3] + humi.dif[, 3] + hetero2[, 3] +
hetero2[, 4] + hetero2[,5] + hetero2[,6] + hetero2[, 7] + biomas.jac[, 3] +
realm.jac[,3] + eco.num.jac[,3]+  eco.name.jac[, 3] + poligono.r.jac[,3],cp=0.001)
n= 4220

          CP nsplit rel error    xerror       xstd
1 0.52000017      0 1.0000000 1.0004895 0.05884552
2 0.05724959      1 0.4799998 0.4944784 0.02649525
3 0.03988976      2 0.4227502 0.4375102 0.02330719
4 0.03703538      3 0.3828605 0.4186438 0.02263688
5 0.01760757      4 0.3458251 0.3759179 0.02013223
6 0.01717674      5 0.3282175 0.3671985 0.01935219
7 0.01014675      6 0.3110408 0.3551583 0.01865565
8 0.01000000      7 0.3008940 0.3377010 0.01749871

Node number 1: 4220 observations,    complexity param=0.5200002
  mean=0.04973464, MSE=0.002646488
  left son=2 (3960 obs) right son=3 (260 obs)
  Primary splits:
      alt.dif[, 3]  < 618.5273    to the left,  improve=0.5200002, (0 missing)
      temp.dif[, 3] < 2.98235     to the left,  improve=0.4752946, (0 missing)
      pet.dif[, 3]  < 13.23969    to the left,  improve=0.3856393, (0 missing)
Surrogate splits:
      temp.dif[, 3]  < 3.137775    to the left,  agree=0.985, adj=0.750, (0 split)
      humi.dif[, 3]  < 9.603662    to the left,  agree=0.957, adj=0.300, (0 split)
      pet.dif[, 3]   < 21.02715    to the left,  agree=0.952, adj=0.223, (0 split)
```

GLM

Modelos de regressão com variável resposta binária

Até agora vimos que uma das suposições dos modelos lineares é que o erro deve ser uma variável contínua e seguir uma distribuição normal. Para algumas situações esta suposição simplesmente não se aplica. Nestes casos, usamos uma extensão dos modelos lineares, que chamamos Modelos Lineares Generalizados (GLM). Nestes casos, a distribuição dos erros pode assumir diversas distribuições, por exemplo Binomial e de Poisson. Os modelos lineares vistos até aqui nada mais são do que um caso particular dos Modelos Generalizados, a saber quando usamos a distribuição Gaussiana (Normal) para os erros. Nesta aula veremos como usar GLM para situações onde a variável resposta é binária. A extensão para outros tipos de variável resposta é bastante direta.

Situações em que temos uma variável resposta binária:

1. Existe relação entre probabilidade de ataque de ovelhas por onça e distância do fragmento florestal mais próximo ? Neste caso, nossa variável resposta é categórica com apenas dois níveis: fazendas com e sem ataque. Cada fazenda é uma observação.

2. Existe relação entre presença de câncer de pulmão e quantidade de cigarros consumidos por mês?

3. Quais fatores ambientais estão associados com a morte de animais silvestres em rodovias. Neste caso, poderíamos sortear trechos de tamanhos semelhantes ao longo da rodovia. Para cada trecho (nossa observação) anotaríamos se houve ou não morte de animais no período de estudo (nossa variável resposta) e diversas variáveis ambientais que possam explicar nossa resposta, tais como i) presença ou não de corpos d'água (categórica 2 níveis), densidade de vegetação nas margens da rodovia (contínua), intensidade de tráfego no trecho (contínua), velocidade média dos automóveis no trecho (contínua) etc. Veja que este estudo seria muito semelhante àqueles em que usamos Regressão Múltipla. De fato, a análise é muito semelhante e chamamos Regressão Logística Múltipla. Aqui testamos se cada variável preditora é importante (significante) na determinação da resposta.

Categorização

Nos casos em que temos uma variável categórica com 2 níveis, é costume designá-las como ‘0’ e ‘1’. Embora a designação de um determinado nível seja arbitrária, costumamos designar ‘0’ como a ausência/falha do evento e ‘1’ como ‘presença/sucesso’.

Problemas quando a variável resposta é binária

1. Erros não são normais. De fato, a resposta pode apenas assumir valores ‘0’ e ‘1’.
2. Variância não é constante; depende do valor de X
3. Truncamento dos valores de resposta entre ‘0’ e ‘1’

Regressão Logística Simples

Note que para cada quantidade n de observações, uma parte será de sucessos. Sabendo a proporção de sucessos (=sucessos/total) também sabemos a proporção de falhas. No modelo de regressão logística, a probabilidade de sucessos (ou falhas) é a expectativa (E{Y}) da variável resposta. Se E{Y} = 0.9, entendemos que em 90% das observações a resposta é 1 (sucesso) e em apenas 10% a resposta é 0 (falha). Nosso modelo portanto é:

$$E\{Y_i\} = \pi_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$$

O evento *i* tem probabilidade de assumir valor 1 de acordo com a equação acima. Veja que *i* pode assumir apenas dois estados: ‘0’ e ‘1’.

Podemos também representar nosso modelo com a seguinte equação:

$$\pi_i' = \log_e\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i$$

A transformação da variável resposta acima é dita *Transformação Logit*.

O termo $\pi_i/(1-\pi_i)$ é chamado *odds* = probabilidade de 'sucesso' dividido pela probabilidade de 'falha'. A interpretação de *odds* é importante.

Podemos pensar que se a probabilidade de sucesso é 0.9, a de falha é de 0.1. Neste caso temos 9:1 (9 sucessos para cada 1 falha) chances de sucesso.

Exemplo

Uma pesquisadora (mineira) quis investigar se a utilização ou não de árvores por gaviões era dependente da altura daquelas. Ela designou sucesso (1) quando haviam fezes ao redor da árvore e falha (0) quando não haviam fezes de gaviões. Os resultados seguem abaixo, onde *alt* é a altura da árvore*10 (m) e *resp* é a resposta:

No R:

```
gaviao.logi<-glm(resp~alt,data=gaviao,family=binomial)
summary(gaviao.logi)
```

alt	resp
14	0
29	0
6	0
25	1
18	1
4	0
18	0
12	0
22	1
6	0
30	1
11	0
30	1
5	0
20	1
13	0
9	0
32	1
24	0
13	1
19	0
4	0
28	1
14	1
29	1

Parte dos resultados:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.05970    1.25935  -2.430   0.0151 *
alt          0.16149    0.06498   2.485   0.0129 *
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

Nossa equação ajustada portanto:

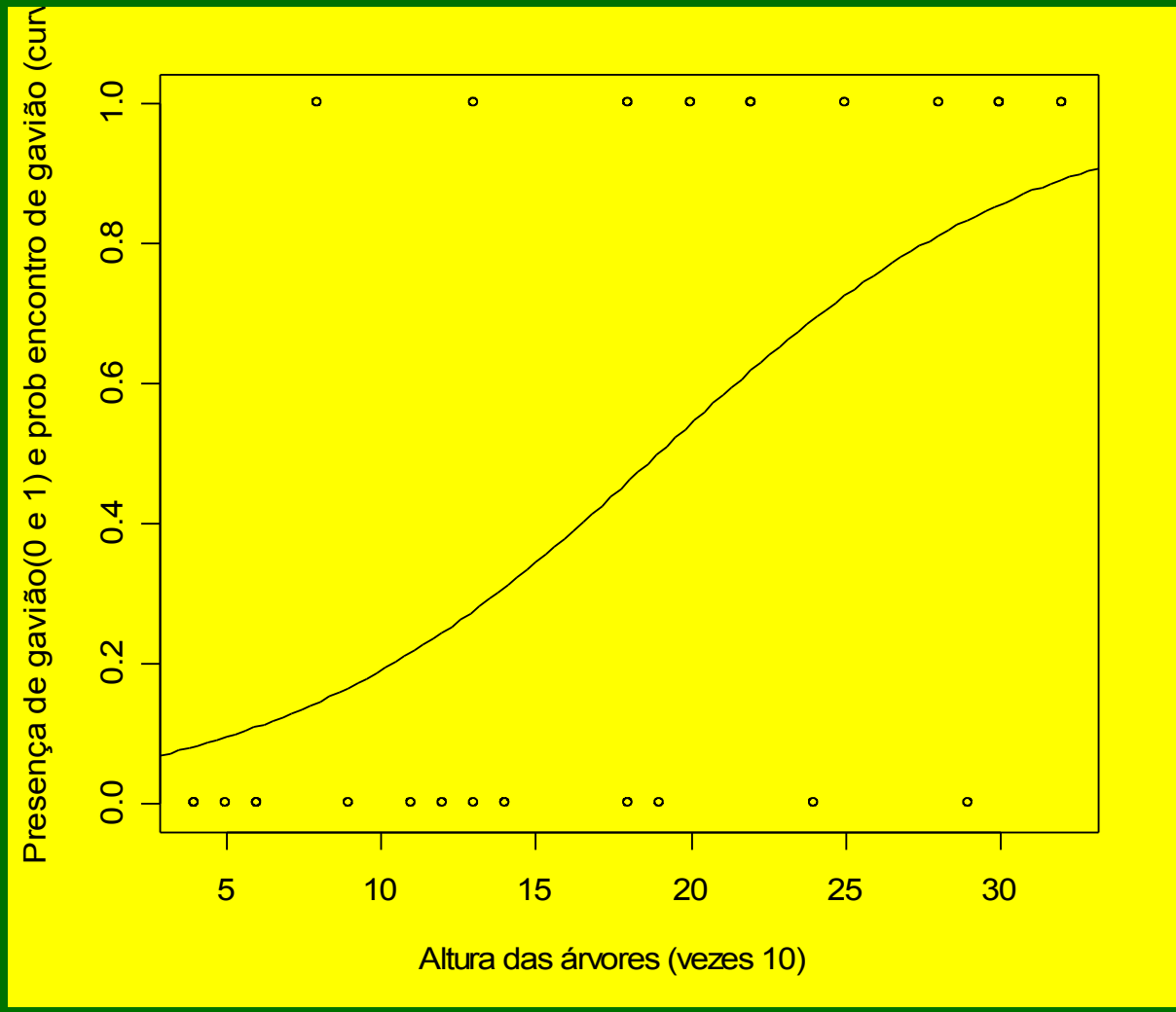
$$\hat{\pi} = \frac{\exp(-3.0597 + 0.16149X)}{1 + \exp(-3.0597 + 0.16149X)}$$

Para árvores com tamanho = 14, temos que a probabilidade de observarmos a presença de gaviões é:

$$\frac{\exp(-3.0597 + 0.16149 * 14)}{1 + \exp(-3.0597 + 0.16149 * 14)} = 0.310$$

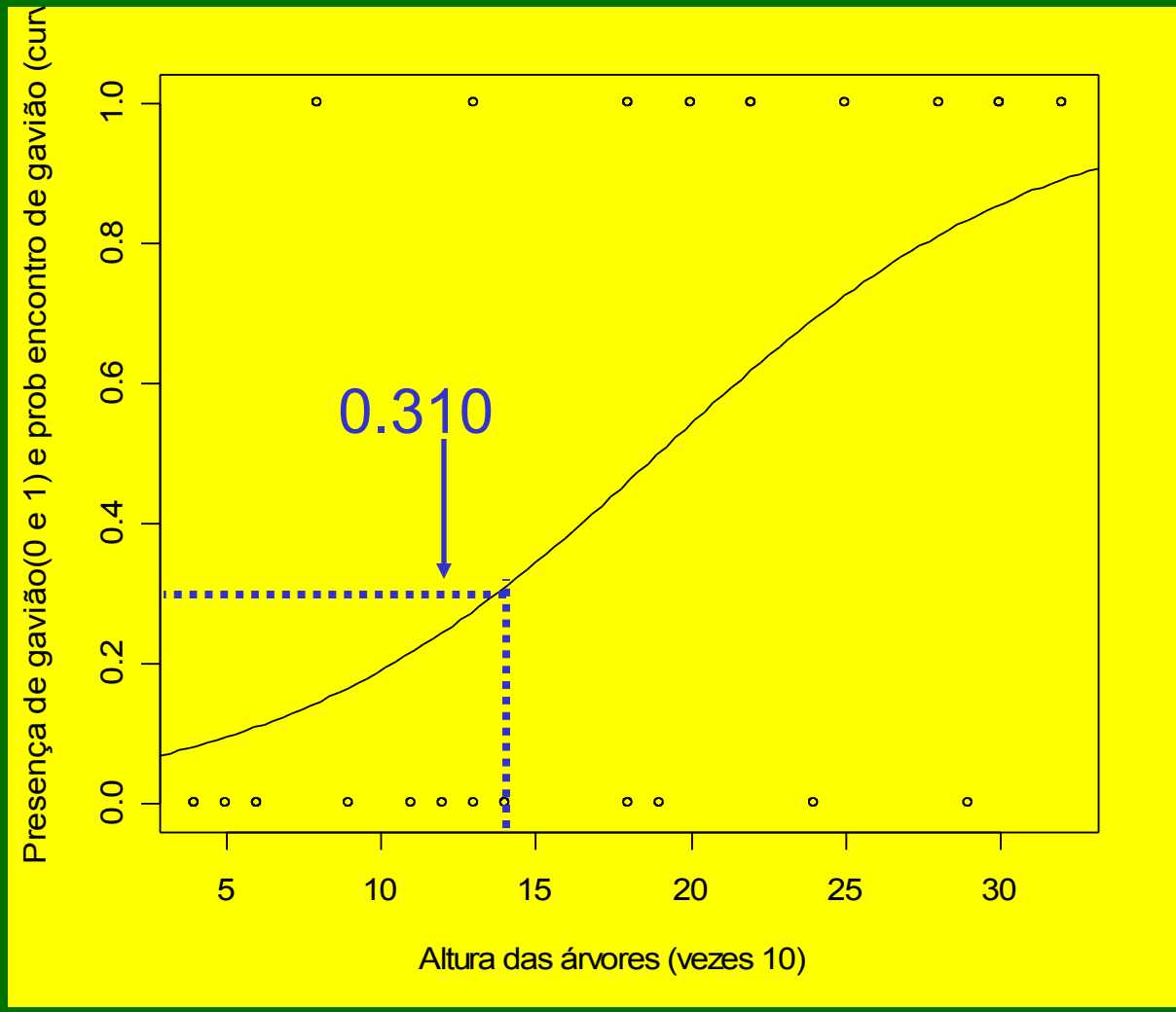
Podemos ajustar nosso modelo aos dados:

```
plot(alt,resp,xlab="Altura das árvores (vezes 10)",
ylab="Presença de gavião(0 e 1) e prob encontro de gavião
(curva)")
curve((exp(-3.0597 + 0.1615*x))  / (1+(exp(-3.0597 +
0.1615*x))),add=TRUE)
```



Podemos ajustar nosso modelo aos dados:

```
plot(alt,resp,xlab="Altura das árvores (vezes 10)",
ylab="Presença de gavião(0 e 1) e prob encontro de gavião
(curva)")
curve((exp(-3.0597 + 0.1615*x))  / (1+(exp(-3.0597 +
0.1615*x))),add=TRUE)
```



Interpretação de b

A interpretação de b não é direta como era na regressão linear, pois o efeito em Y de cada aumento de uma unidade de X, varia conforme o valor de X. Lembre que *odds* é a razão entre a probabilidade de sucesso e a probabilidade de não sucesso:

odds = $\pi_i / (1 - \pi_i)$

$$b = \log_e \left(\frac{odds_2}{odds_1} \right)$$

onde

$odds_1$ = valor de *odds* para X = k

$odds_2$ = valor de *odds* para X = k+1

$$b = \log_e \left(\frac{odds_2}{odds_1} \right)$$

$$OR = \left(\frac{odds_2}{odds_1} \right) = \exp(b)$$

onde OR = *Odds Ratio*.

No nosso exemplo, vimos que existe uma relação positiva entre altura de árvore e probabilidade de encontro de gavião. Especificamente, podemos dizer que cada aumento de unidade em altura de árvore equivale a um aumento de 17.5% no *odds ratio* [exp(0.1615) = 1.175]

Vimos que para árvores com altura 14, a probabilidade de encontramos um gavião era de 0.310. O *odds* era portanto: 0.310/(1-0.310) = 0.449

Para uma árvore com altura 15, a probabilidade é de 0.34588. O *odds* aqui é de 0.34588/(1-0.34588) = 0.5287

OR = 0.5287/0.449 = 1.175 = exp(0.1615)

Uma interpretação talvez mais assimilável do *odds ratio:*

Na mídia ouvimos falar que pessoas que fumam possuem 17.5% a mais de chance de desenvolver câncer de pulmão. Neste exemplo, a variável preditora (fuma ou não-fuma) é categórica com dois níveis.
Veja que os 17.5% aqui não se referem à probabilidade de desenvolver câncer, mas sim ao aumento relativo em chance de desenvolver câncer quando o paciente fuma.

Regressão Logística Polinomial

A idéia também é muito parecida com o modelo polinomial quando a resposta é contínua.

Esta análise não é muito utilizada em ecologia mas pode ser muito útil (na verdade acho que Regressão Logística em geral é pouco utilizada por ecólogos por simples desconhecimento...).

Frequentemente estamos testando se existe uma relação modal entre presença de determinada espécie e uma variável ambiental.

Em geral esperamos que a probabilidade de presença seja alta no meio do gradiente (o ótimo da espécie) e baixa nas pontas (longe do ótimo).

No nosso modelo de regressão logística basta incluir um termo quadrático.

Exercícios e estudo individual:

-- Lista em sala de aula

-- Crawley: Cap. 11, 14 e 16

-- Gotelli e Ellison: 273-275

Bibliografia para Árvores de Regressão:

De'ath, G., and K. E. Fabricius. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology **81**:3178-3192.

Therneau, T. M. and E. J. Atkinson. 1997. An introduction to recursive partitioning using the *rpart* routines. Technical Report 61, Mayo Clinic, Section of Biostatistics. Rochester, Minnesota.

http://mayoresearch.mayo.edu/mayo/research/biostat/upload/61.pdf

http://mayoresearch.mayo.edu/mayo/research/biostat/upload/rpartmini.pdf