Modelos Não-Gaussianos com covariáveis

Paulo Inácio Prado e João L.F. Batista

BIE5781 - Pós-Graduação em Ecologia USP

Novembro de 2020

Objetivo da Aula

Os objetivos dessa aula são:

1. Generalizar os modelos estatísticos com covariáveis para distribuições não Gaussianas;
2. Exemplificar essas generalização com modelos Poisson e binomial;
3. Mostrar os comandos básicos no R para ajustar e avaliar esses modelos;
4. Mostrar que alguns deles pertencem à classe dos modelos lineares generalizados (glms);
5. Apresentar ajuste de glms no R.

Onde estamos

Até agora vimos:

- Modelos de várias distribuições com parâmetros constantes:

$$
\begin{aligned}
Y &\sim N(\mu = a_0,\, \sigma = b_0) \\
Y &\sim P(\lambda = a_0) \\
&\ldots
\end{aligned}
$$

- Modelos Gaussianos com covariáveis:

$$
\begin{aligned}
Y &\sim N(\mu = a_0 + a_1 X_1,\, \sigma = b_0) \\
Y &\sim N\left(\mu = a_0 + a_1 X_1 + a_2 X_2,\, \sigma = e^{b_0 + b_1 X_1}\right) \\
&\ldots
\end{aligned}
$$

Para onde vamos

Hoje veremos modelos de distribuições não Gaussianas com covariáveis, como:

- Modelos Binomiais:

$$Y \sim \mathrm{Bin}(N = n,\, p = g(X_i))$$

- Modelos Poisson:

$$Y \sim P(\lambda = g(X_i))$$

Ou seja

Uma expressão geral para modelos estatísticos

$$Y \sim f(Y \mid \Theta = g(X_i))$$

Em palavras

Y é uma variável aleatória, cujos valores tem probabilidades definidas pela função de densidade $f(Y)$, cujos parâmetros Θ são uma função qualquer g de covariáveis X_i.

Modelos Binomiais

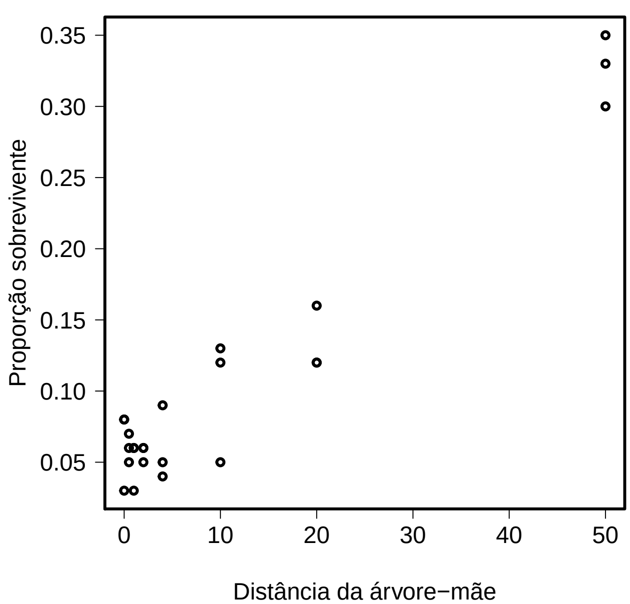
- Descrevem número de ocorrência de uma variável binária em um certo número de tentativas, em função de variáveis preditoras contínuas ou discretas.
- Úteis principalmente quando temos:
  - Respostas binárias (regressão logística!)
  - Proporções
- Exemplos de respostas binomiais:
  - Presença de alguma espécie em manchas, fragmentos, sítios;
  - Ocorrência de morte, doença ou qualquer outro evento por indivíduo
  - Proporção de frutos parasitados
  - Experimentos de dose-resposta

Modelos binomiais: um exemplo

- Sobrevivência de sementes colocadas as diferentes distâncias da árvore-mãe
- 8 distâncias, 3 réplicas de 100 sementes
- Resposta: número de sobreviventes após 60 dias:

```
> head(pred.seed)
  distancia n.sobrev
1       0.0        8
2       0.0        8
3       0.0        3
4       0.5        7
5       0.5        6
6       0.5        5
```

Modelos binomiais: um exemplo



Modelo Binomial: Função Logística

- Nosso modelo é:

$$Y \sim Bin(N = 100, p = g(\text{distância}))$$

- Que função usar para o efeito da distância sobre p?
- Uma função sigmóide limitada entre zero e um, como:

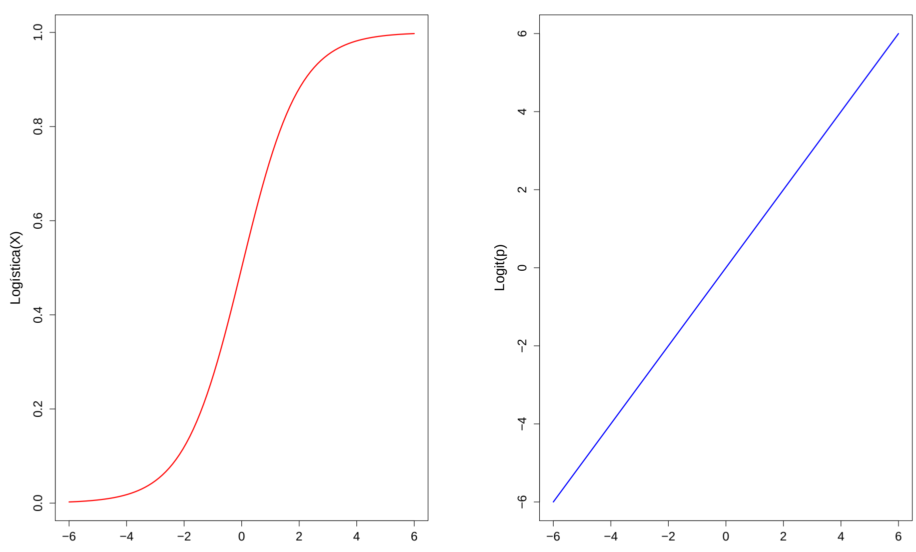
$$p = \frac{e^{a_0 + a_1 \text{dist}}}{1 + e^{a_0 + a_1 \text{dist}}}$$

- o que implica em

$$\ln\left(\frac{p}{1-p}\right) = a_0 + a_1 \text{dist}$$

Modelo Binomial: Funções logística e *logito*

$$p = \frac{e^{a+bX}}{1+e^{a+bX}} \qquad Y = \ln\left(\frac{p}{1-p}\right)$$



Modelo Binomial: Ajuste no R

Função auxiliar: logística

```
> logistica <- function(X, a0, a1){
+     exp(a0 + a1*X) / (1 + exp(a0 + a1*X))
+ }
```

Modelo Binomial: Ajuste no R

Função de log-verossimilhança negativa

```
> seed.LL1 <- function(a0, a1){
+     eta  <- logistica(X=pred.seed$distancia,
+                              a0=a0, a1=a1)
+     -sum(dbinom(pred.seed$n.sobrev, size = 100,
+                  p = eta, log=T))
+ }
```

Modelo Binomial: Ajuste no R

Valores iniciais: regressão linear dos logitos

```
> p1 <- pred.seed$n.sobrev/100
> logito1 <- log( p1 / (1-p1) )
> lm1 <- lm(logito1~pred.seed$distancia)
> (cf1 <- coef(lm1))

        (Intercept) pred.seed$distancia
        -2.86122674          0.04383901
```

Modelo Binomial: Ajuste no R

Ajuste numérico

```
> seed.m1 <- mle2(seed.LL1,
+                 start=list(a0=cf1[1], a1=cf1[2]) )
```

Modelo Binomial: Ajuste no R

Coeficientes

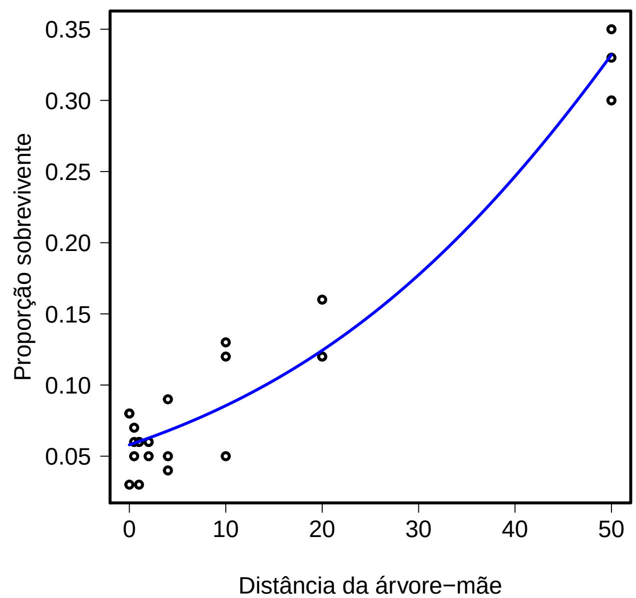
```
> (cf2 <- coef(seed.m1))

        a0         a1
-2.7874067  0.0417725
```

Modelo Binomial: Ajuste no R

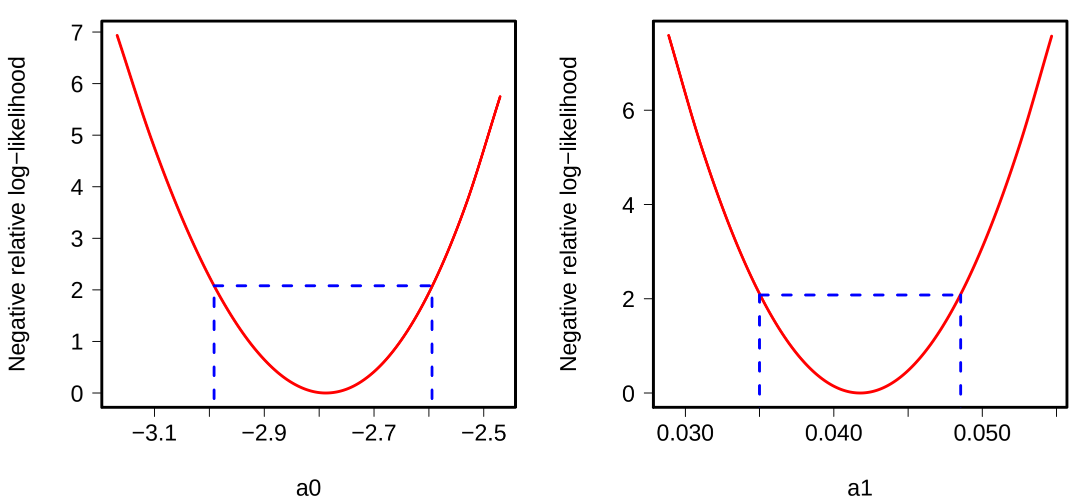
Código do Gráfico observado e estimado

```
> plot(I(n.sobrev/100)~dist, data=pred.seed,
+      xlab="Distância da árvore-mãe",
+      ylab="Proporção sobrevivente")
> curve(logistica(x, a0 = cf2[1], a1 = cf2[2]),
+       add=T, col="blue")
```

Modelo Binomial: Gráfico previsto



Modelo Binomial: Perfis de log-verossimilhança



Modelo Binomial: incerteza das estimativas

Intervalo de plausibilidade

```
> likelregions( profile(seed.m1) )

Likelihood regions for ratio = 2.079442


 a0:
          lower     upper
[1,] -2.991295 -2.593989


 a1:
           lower      upper
[1,] 0.03504051 0.04852435
```

Modelo Binomial: incerteza das estimativas

Intervalo de confiança

```
> confint( seed.m1 )

          2.5 %      97.5 %
a0 -2.98316282 -2.60111255
a1  0.03530014  0.04825401
```

Modelo binomial: seleção de modelos

Modelo sem efeitos: log-verossimilhança negativa

```
> seed.LL0 <- function(a0){
+     eta <- exp(a0)/(1+exp(a0))
+     -sum(dbinom(pred.seed$n.sobrev,
+                 size=100, p=eta, log=T)) }
```

Modelo binomial: seleção de modelos

Modelo sem efeitos: uma LL mais simples

```
> seed.LL0 <- function(p){
+     -sum(dbinom(pred.seed$n.sobrev,size=100,
+                 prob=p, log=T)) }
```

Modelo binomial: seleção de modelos

Ajuste do modelo sem efeito da distância

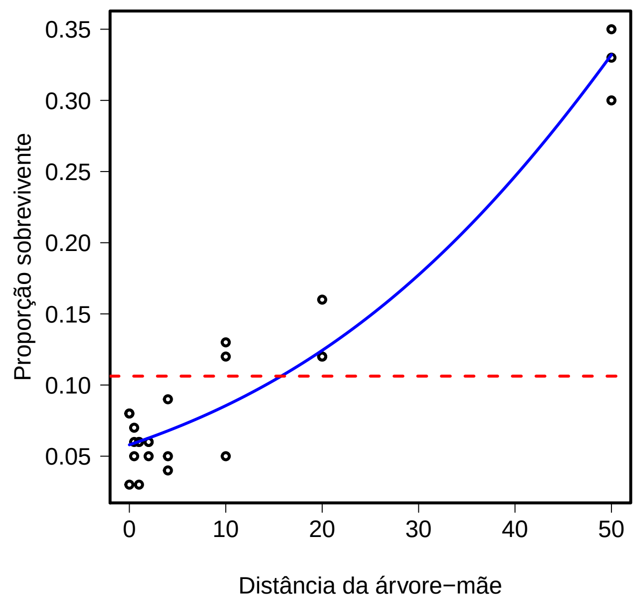
```
> seed.m0 <- mle2(seed.LL0,
+                 start=list(
+                     p=sum(pred.seed$n.sobrev)/2400)
+                 )
```

Modelo binomial: seleção de modelos

Comparação com AIC (pacote bbmle)

```
> AICtab(seed.m0, seed.m1, base=T)

        AIC   dAIC  df
seed.m1 112.2   0.0 2
seed.m0 259.8 147.6 1
```

Modelos Binomiais: Previstos pelos dois modelos



Modelos Poisson

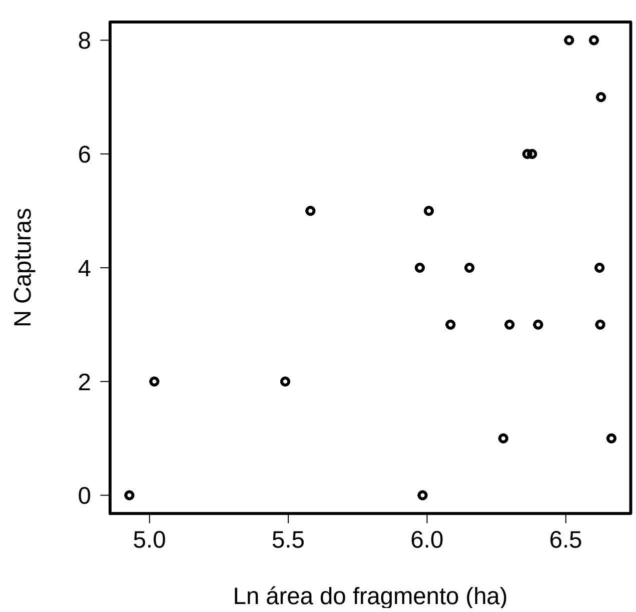
- Descrevem contagens de eventos independentes em função de variáveis preditoras contínuas ou discretas.
- Úteis principalmente quando temos:
  - contagens com médias baixas
  - contagens com variância igual à média
- Exemplos de respostas Poisson:
  - Número de capturas, avistamentos, registros por unidade de tempo ou espaço
  - Taxas de ocorrência de eventos (*e.g.* fotos/hora, células/área)

Modelos Poisson: um exemplo

- Capturas de *Pyriglena leucoptera* em redes de neblina
- 20 fragmentos de diferentes tamanhos, 500 horas/rede
- Resposta: número de capturas

```
> head(aves)
  area ncap
1  736    8
2  753    3
3  265    5
4  673    8
5  531    1
6  439    3
```

Modelo Poisson: Gráfico



Modelos Poisson

- Nosso modelo é:

$$Y \sim P(\lambda = g(\log(\text{área}))$$

- Que função usar para o efeito do log da área sobre λ?
- Uma função monotônica positiva como a exponencial:

$$\lambda = e^{a_0 + a_1 \log(\text{área})}$$

- O que implica em

$$\ln(\lambda) = a_0 + a_1 \log(\text{área})$$

Modelo Poisson: Ajuste no R

Função de log-verossimilhança negativa

```
> aves.LL1 <- function(a0, a1) {
+   eta <-  exp( a0 + a1*log(aves$area) )
+   -sum( dpois(aves$ncap, lambda = eta, log = T) )
+ }
```

Modelo Poisson: Ajuste no R

Valores iniciais: regressão linear dos logarítmos

```
> pm1 <- lm(log(ncap+0.1)~log(aves$area))
> ( cf1 <- coef(pm1) )

   (Intercept) log(aves$area)
     -6.201162       1.165563
```

Modelo Poisson: Ajuste no R

Ajuste numérico

```
> aves.m1 <- mle2(aves.LL1,
+                 start= list(a0 = cf1[1],
+                             a1 = cf1[2]) )
```

Modelo Poisson: Ajuste no R

Coeficientes

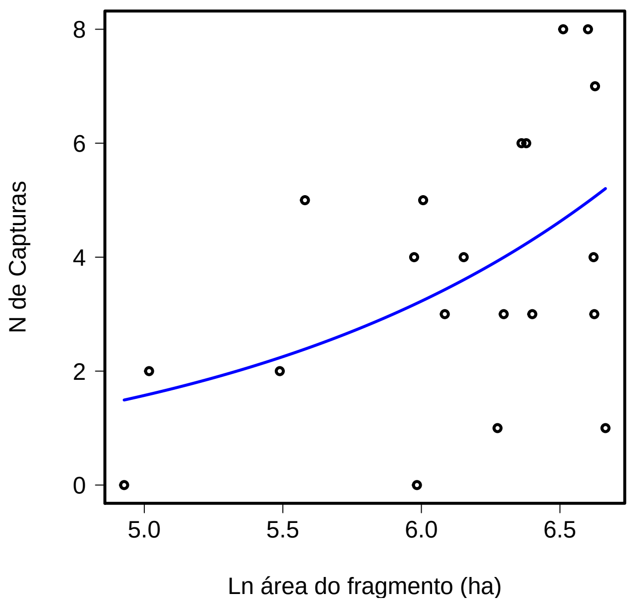
```
> ( cf3 <- coef(aves.m1) )

        a0         a1
-3.1409439  0.7187925
```

Modelo Poisson: Ajuste no R

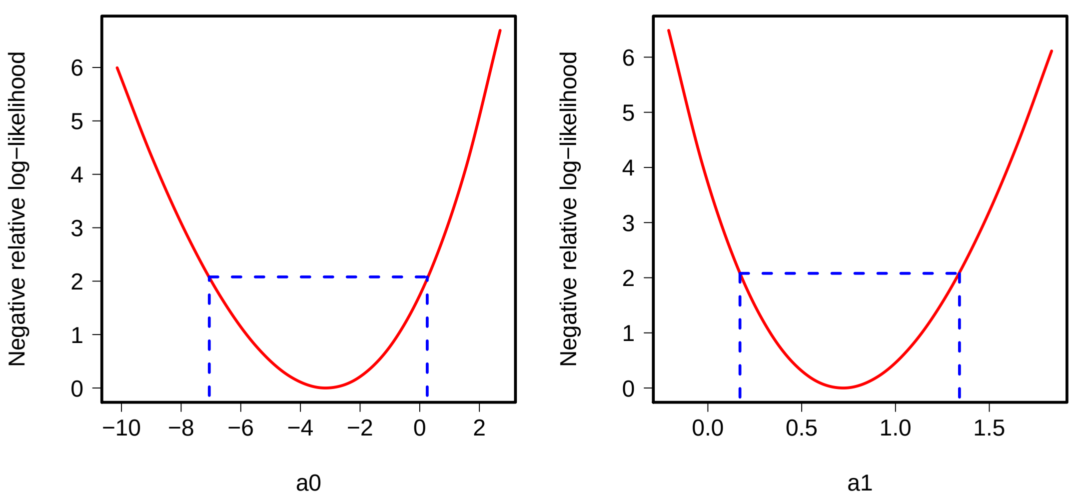
Código do Gráfico observado e estimado

```
> f3 <- function(x) exp(cf3[1]+cf3[2]*x)
>
> plot(ncap~log(area), data = aves,
+      xlab = "Ln área do fragmento (ha)",
+      ylab = "N de Capturas")
> curve(f3(x), add = T, col = "blue")
```

Modelo Poisson: Gráfico observado e previsto



Modelo Poisson: Perfis de log-verossimilhança



Poisson: incerteza das estimativas

Intervalo de plausibilidade

```
> likelregions( profile(aves.m1) )

Likelihood regions for ratio = 2.079442


 a0:
         lower     upper
[1,] -7.065969 0.2807356


 a1:
         lower    upper
[1,] 0.1714548 1.339336
```

Poisson: incerteza das estimativas

Intervalo de confiança

```
> confint( aves.m1 )

        2.5 %    97.5 %
a0 -6.9031009 0.1522048
a1  0.1914572 1.3129705
```

Modelo Poisson: seleção de modelos

Modelo sem efeitos: log-verossimilhança negativa

```
> aves.LL0 <- function(a0){
+     eta <- exp(a0)
+     -sum(dpois(aves$ncap, lambda=eta, log=T))
+ }
```

Modelo Poisson: seleção de modelos

Ajuste do modelo sem efeito da área

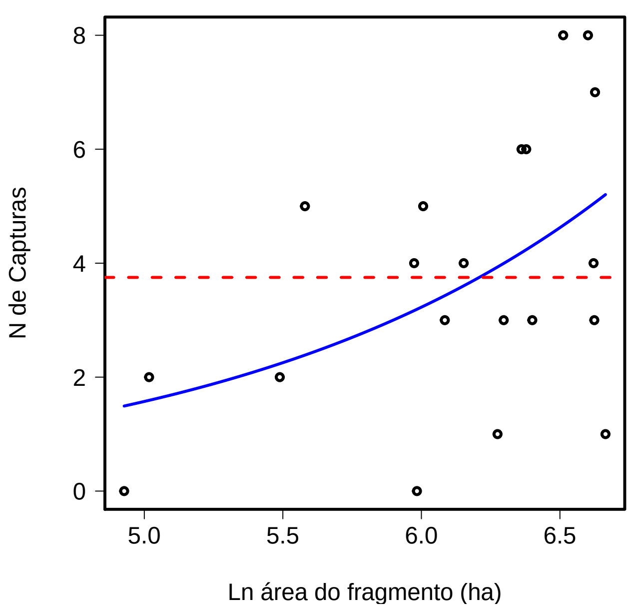
```
> aves.m0 <- mle2(aves.LL0,
+                 start=list(
+                     a0=log(mean(aves$ncap))))
```

Modelo Poisson: seleção de modelos

Comparação com AIC

```
> AICtab(aves.m0,aves.m1, base=T)

        AIC  dAIC df
aves.m1 89.4  0.0 2
aves.m0 94.8  5.4 1
```

Modelos Poisson: previstos pelos dois modelos



Generalized linear models: exemplo Poisson

Generalized linear models: exemplo Poisson

Função glm no R

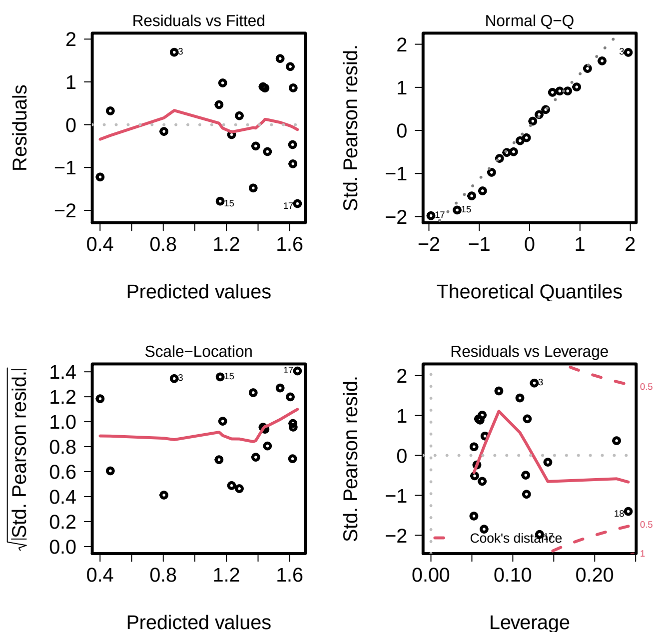
```
> aves.glm1 <- glm(ncap~log(area), family=poisson)
> coef(aves.glm1)

(Intercept)   log(area)
 -3.1425010   0.7190406

> coef(aves.m1)

        a0          a1
-3.1409439  0.7187925
```

Generalized linear models: exemplo Poisson

Função `glm` no R

```
> logLik(aves.glm1)
'log Lik.' -42.70365 (df=2)
> logLik(aves.m1)
'log Lik.' -42.70365 (df=2)
```

GLMs: gráficos diagnósticos



Modelo lineares generalizados (glm)

- Generalizam a lógica da regressão linear gaussiana para algumas outras distribuições;
- Preservam várias propriedades dos modelos lineares gaussianos;
- Abrangem vários modelos sob um conjunto único de soluções analíticas e de propriedades matemáticas.
- a.k.a uma teoria unificada.
- Definiu a sintaxe de modelos do R e de várias outras linguagens computacionais.

Modelos lineares generalizados (glm)

J. R. Statist. Soc. A, (1972), **135**, *Part* 3, p. 370

370

Generalized Linear Models

By J. A. Nelder and R. W. M. Wedderburn

Rothamsted Experimental Station, Harpenden, Herts

Summary

The technique of iterative weighted linear regression can be used to obtain maximum likelihood estimates of the parameters with observations distributed according to some exponential family and systematic effects that can be made linear by a suitable transformation. A generalization of the analysis of variance is given for these models using log-likelihoods. These generalized linear models are illustrated by examples relating to four distributions; the Normal, Binomial (probit analysis, etc.), Poisson (contingency tables) and gamma (variance components).

The implications of the approach in designing statistics courses are discussed.

A família exponencial de distribuições [1]

- Distribuições que podem ser expressas por

$$y \sim \exp\left\{\frac{y\Theta - f(\Theta)}{g(\phi)} + h(y, \phi)\right\}$$

- como:
  - Gaussiana
  - Binomial
  - Poisson
  - Gama
  - Inversa da normal

[1] Não é o mesmo que distribuição exponencial

A família exponencial de distribuições - Exemplo

- A distribuição Poisson

$$f(y) = \frac{e^{-\lambda}\lambda^y}{y!}$$

- Pertence à família exponencial porque pode ser expressa como

$$\exp\left\{\frac{y\Theta - f(\Theta)}{g(\phi)} + h(y, \phi)\right\}$$

- onde

$$\begin{aligned} \Theta &= \ln(\lambda) \\ f(\Theta) &= e^{\lambda} \\ g(\phi) &= 1 \\ h(y, \Theta) &= -\ln(y!) \end{aligned}$$

A função de ligação

- Os glms têm este nome porque generalizam a ideia de estabelecer uma relação do valor esperado com uma combinação linear de covariáveis:

$$\eta(E[Y]) \;=\; a_0 + a_1 X_1 + \ldots + a_i X_i \;=\; \sum_0^i a_i X_i$$

- A função η estabelece esta relação. Ela é chamada **função de ligação**.
- Quando a função de ligação corresponde ao parâmetro Θ da família exponencial ela é chamada **função de ligação canônica**. Funções de ligação canônicas dão várias propriedades estatísticas importantes aos modelos.

Exemplos de funções de ligação canônicas

- Função log para a Poisson:

$$\ln E[Y] \;=\; \ln \lambda \;=\; \sum_0^i a_i X_i$$

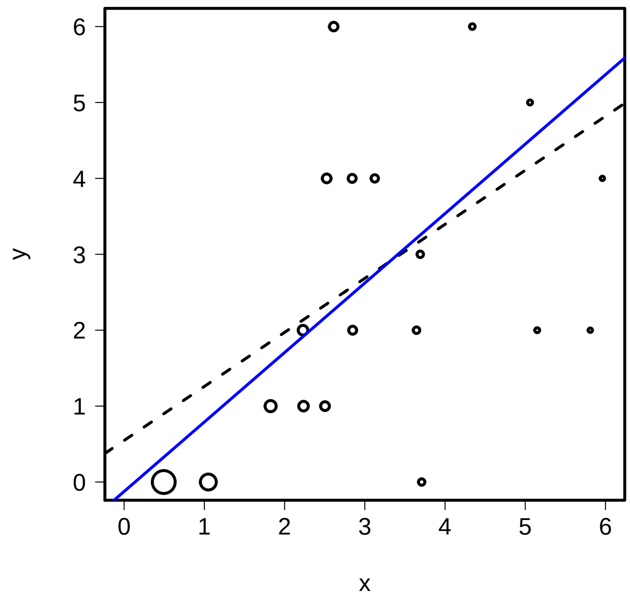
- Função logito para a binomial:

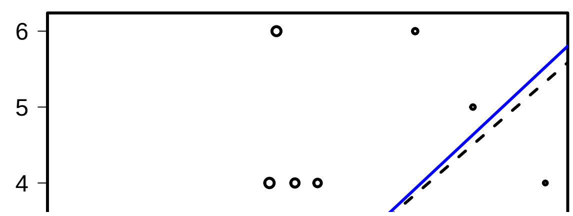
$$\ln\left(\frac{E[Y]}{n - E[Y]}\right) \;=\; \ln\left(\frac{p}{1-p}\right) \;=\; \sum_0^i a_i X_i$$

- Função identidade para a normal

$$E[Y] \;=\; \sum_0^i a_i X_i$$

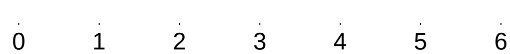
Iteratively reweighted least squares method





Modelo lineares generalizados

- Função `glm` no R
- Generalizações:
  - *Quasi-likelihood* para dados com sobredispersão
  - Modelo similar para binomial negativa (função `glm.nb`)
  - Modelos lineares generalizados de efeitos mistos
  - Modelos lineares generalizados bivariados
  - ...



Para saber mais

- Bolker, B.M. 2008 Ecological Models and Data in R Princeton: Princeton University Press, caps 6 e 9.
- Crawley, M.J. 2007. The R Book. (caps.13,14 e 16)
- Dobson, A.J. 1990. An Introduction to Generalized Linear Models. London: Chapman and Hall.
- McCullagh P. & Nelder, J.A. 1989. Generalized Linear Models. London: Chapman and Hall.
- Rodríguez, G. 2007. Lecture Notes on Generalized Linear Models. `http://data.princeton.edu/wws509/notes/`