Modelos Matemáticos e Aplicações
Modelos Lineares Generalizados

Jorge Cadima

Matemática (DCEB), Instituto Superior de Agronomia (ULisboa)

2023-24

Bibliografia

- Agresti, A. (2015) *Foundations of Linear and Generalized Linear Models*, Wiley.
- Dobson, A.J. & Barnett, A.G. (2008) *An Introduction to Generalized Linear Models*, 3rd ed., CRC Press.
- McCullaugh, P. & Nelder, J. (1989) *Generalized Linear Models*, Chapman & Hall.
- McCulloch, C. & Searle, S. (2001) *Generalized, Linear, and Mixed Models*, John Wiley & Sons. **Mat 600-62**.
- Agresti, A. (1990) *Categorical Data Analysis*, John Wiley & Sons. **Mat 401-62**.
- Hosmer, D.W. & Lemeshow, S. (1989) *Applied Logistic Regression*, John Wiley & Sons. **Mat 258-62**.

GLMs no  :

- Faraway, J.J. (2006) *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapman & Hall/CRC.
- Fox, J. & Weisberg, S. (2011) *An R Companion to Applied Regression, 2d Ed*, Sage Publications (`R` *package*: `car`).
- Venables & Ripley (2002). *Modern Applied Statistics with S* (4a. edição), Springer. (`R` *package*: `MASS`).

Modelos Lineares Generalizados

Modelos Lineares Generalizados (MLGs)
(Generalized Linear Models, GLMs):

- são uma família muito vasta de modelos;
- são extensão do Modelo Linear;
- englobam muitos modelos previamente conhecidos e usados, nalguns casos há largas décadas, entre eles:
  - modelo *probit*
  - modelo *logit* (ou Regressão Logística)
  - modelos log-lineares
  - o próprio modelo linear.
- o "chapéu de chuva comum" dos MLGs foi introduzido e formalizado por McCullagh e Nelder (1989);

Exemplo motivador: variável resposta dicotómica

Exemplo Hosmer & Lemeshow

Hosmer e Lemeshow, em *Applied Logistic Regression* (Wiley, 1989) têm dados sobre $n = 100$ pacientes, com variáveis:

- idade – numérica;
- doença arterial coronária (DAC) – variável dicotómica (sim/não; 1/0).

Eis as primeiras seis linhas da *data frame* HL correspondente:

```
> head(HL)
```

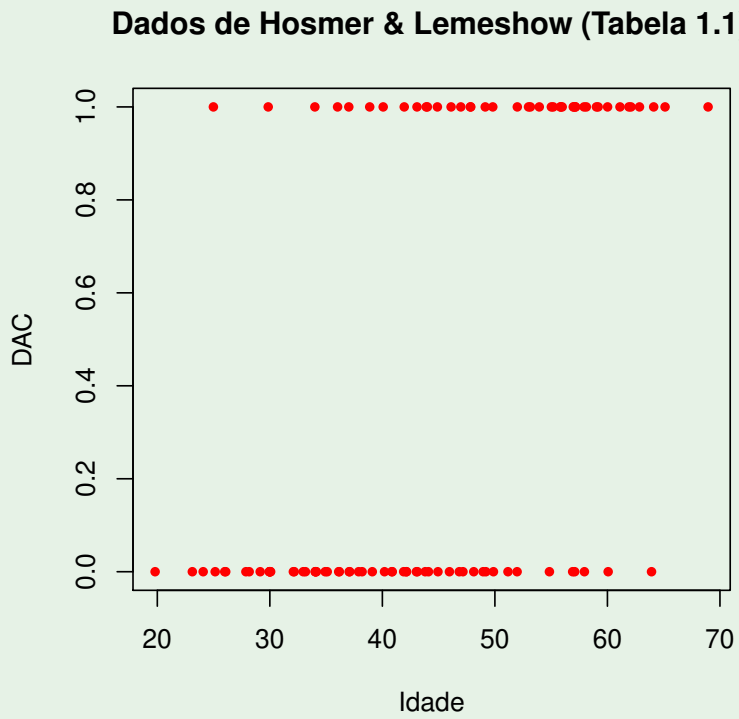
```
  Idade  DAC
1    20    0
2    23    0
3    24    0
4    25    0
5    25    1
6    26    0
```

Quer-se relacionar a existência de DAC (variável resposta Y) com a idade (preditor X). O gráfico Y vs. X é pouco prometedor.

Exemplo 1: DAC *vs.* idade

Exemplo Hosmer & Lemeshow

```
> plot(DAC ~ Idade , data=HL , cex=0.8 , col="red" , pch=16 ,
+ xlab="Idade" , main="Dados de Hosmer & Lemeshow (Tabela 1.1)")
```



O Modelo Linear

Recorde-se que o modelo linear relaciona

- uma variável resposta numérica Y com
- preditores X_1, X_2, ..., X_p,

através da equação, para n observações independentes Y_i:

$$Y_i = \beta_0 + \beta_1 X_{1_{(i)}} + \beta_2 X_{2_{(i)}} + ... \beta_p X_{p_{(i)}} + \varepsilon_i \ ,$$

com $\varepsilon_i \frown \mathscr{N}(0, \sigma^2)$ $(i = 1, 2, ..., n)$.

Isto é, no Modelo Linear:

- $E[Y_i | X_1 = x_{1_{(i)}}, ..., X_p = x_{p_{(i)}}] = \beta_0 + \beta_1 x_{1_{(i)}} + \beta_2 x_{2_{(i)}} + ... \beta_p x_{p_{(i)}}$,
- Y_i independentes, com distribuição Normal (e variâncias iguais).

A generalização do modelo linear

Modelo Linear

- $E[Y_i] = \beta_0 + \beta_1 x_{1_{(i)}} + \beta_2 x_{2_{(i)}} + \ldots \beta_p x_{p_{(i)}}$,
- Y_i com distribuição Normal.

Num Modelo Linear Generalizado há duas extensões:

- $g(E[Y_i]) = \beta_0 + \beta_1 x_{1_{(i)}} + \beta_2 x_{2_{(i)}} + \ldots \beta_p x_{p_{(i)}}$,
  com g uma função invertível chamada função de ligação (link function).
- Y_i com distribuição na família exponencial de distribuições.

Assim, um MLG modela o valor esperado duma variável resposta com distribuição na família exponencial, através da equação:

$$\mu_i = E[Y_i] = g^{-1}(\beta_0 + \beta_1 x_{1_{(i)}} + \beta_2 x_{2_{(i)}} + \ldots \beta_p x_{p_{(i)}}) .$$

Nota: O Modelo Linear é caso particular de MLG: a Normal pertence á família exponencial de distribuições e a função de ligação é a identidade: $g(x) = x, \forall\, x$.

As três componentes dum MLG (cont.)

Ou seja, e nas palavras de Agresti (1990, p.81):

> *um MLG é um modelo linear para uma transformação da esperança duma variável aleatória cuja distribuição pertence à família exponencial.*

Nota: ao contrário do Modelo Linear, nos MLGs não são explicitados erros aleatórios aditivos. A flutuação aleatória da variável resposta é dada directamente pela sua distribuição de probabilidades.

A família exponencial de distribuições

A família exponencial de distribuições inclui, entre outras:

- a Normal
- a Poisson (para variáveis de contagem)
- a Bernoulli (para variáveis dicotómicas – binary)
- a "Binomial/n" (para proporções de êxitos em n provas de Bernoulli)
- a Gama (distribuição contínua assimétrica);
  inclui a Exponencial como caso particular.
- a Gaussiana inversa (distribuição contínua assimétrica).

Nota: Repare-se como a família exponencial inclui distribuições, quer de variáveis aleatórias contínuas, quer de variáveis aleatórias discretas.

As três componentes dum MLG

Na definição de McCullagh e Nelder (1989), um Modelo Linear Generalizado assenta sobre três componentes fundamentais:

1) Componente aleatória (random component):
A variável resposta Y que se quer modelar, tratando-se duma:

- variável aleatória;
- da qual se recolhem n observações independentes; e
- cuja distribuição de probabilidades faz parte da família exponencial de distribuições (definida mais adiante);

Nota: a distribuição de probabilidades da variável resposta aleatória Y já não se restringe à Normal, podendo ser qualquer distribuição numa classe designada família exponencial de distribuições. Algumas generalizações de GLMs admitem distribuições além da família exponencial.

As três componentes dum MLG (cont.)

2) Componente Sistemática:
Consiste numa combinação linear de variáveis preditoras, admitidas não aleatórias.

Havendo p variáveis preditoras e n observações:

$$\beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + \beta_3 x_{3(i)} + ... + \beta_p x_{p(i)} \quad , \qquad \forall\, i \in \{1, ..., n\} \ .$$

Os preditores podem ser variáveis numéricas, factores ou uma mistura de ambos, tal como no Modelo Linear.

Define-se a matriz do modelo $\mathbf{X}_{n\times(p+1)}$ de forma idêntica ao Modelo Linear: uma primeira coluna de uns (associada à constante aditiva) e p colunas adicionais dadas pelas observações de cada variável preditora (variáveis indicatrizes, no caso de factores).

As três componentes dum MLG (cont.)

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1_{(1)}} & x_{2_{(1)}} & \cdots & x_{p_{(1)}} \\ 1 & x_{1_{(2)}} & x_{2_{(2)}} & \cdots & x_{p_{(2)}} \\ 1 & x_{1_{(3)}} & x_{2_{(3)}} & \cdots & x_{p_{(3)}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1_{(n)}} & x_{2_{(n)}} & \cdots & x_{p_{(n)}} \end{bmatrix}$$

O vector da componente sistemática do modelo é dado por:

$$\vec{\eta} = \mathbf{X}\vec{\beta} \ ,$$

sendo $\vec{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_p)$ o vector de coeficientes que define as n combinações lineares (afins) das variáveis preditoras, dado em $\vec{\eta}$.

As três componentes dum MLG (cont.)

3) Função de ligação (link function):
uma função diferenciável e estritamente monótona g que associa as componentes aleatória e sistemática, através duma relação da forma:

$$\begin{aligned} g(\mu_i) &= g(E[Y_i]) = \vec{\mathbf{x}}_{[i]}^t \vec{\boldsymbol{\beta}} \\ &= \beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + \ldots + \beta_p x_{p(i)} \qquad (\forall\, i = 1:n) \end{aligned}$$

sendo $\vec{\mathbf{x}}_{[i]}^t$ a i-ésima linha da matriz $\mathbf{X}$, com os valores dos preditores na i-ésima observação.

O valor esperado de Y, dados os valores dos preditores, é então:

$$\mu_i = g^{-1}\left(\vec{\mathbf{x}}_{[i]}^t \vec{\boldsymbol{\beta}}\right) = g^{-1}\left(\beta_0 + \beta_1 x_{1(i)} + \beta_2 x_{2(i)} + \ldots + \beta_p x_{p(i)}\right)$$

A família exponencial de distribuições

Diz-se que uma variável aleatória Y tem distribuição na família exponencial (bi-paramétrica) usada por McCullagh & Nelder (1989), se a sua função densidade (caso Y contínua) ou de massa probabilística (se Y discreta) se puder escrever na forma:

$$f(\, y \mid \theta, \phi \,) \quad = \quad \mathrm{e}^{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)}$$

onde

- θ e ϕ são parâmetros (escalares reais); e
- $a(\cdot), b(\cdot)$ e $c(\cdot)$ são funções reais conhecidas.

Os parâmetros designam-se:

- θ – parâmetro natural; e
- ϕ – parâmetro de dispersão.

A Normal

A família exponencial inclui a distribuição Normal, cuja f. densidade é:

$$f(y|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} = e^{\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)} e^{-\left(\frac{y^2-2y\mu+\mu^2}{2\sigma^2}\right)}$$

$$= e^{-\ln(\sigma\sqrt{2\pi})-\left(\frac{y^2-2y\mu+\mu^2}{2\sigma^2}\right)} = e^{\frac{y\mu-\frac{\mu^2}{2}}{\sigma^2}-\ln(\sigma\sqrt{2\pi})-\frac{y^2}{2\sigma^2}}$$

é da forma $f(y\,|\,\theta,\phi) = e^{\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)}$, com:

- $\theta = \mu$ (parâmetro natural)
- $\phi = \sigma^2$ (parâmetro de dispersão)
- $b(\theta) = \frac{\theta^2}{2} = \frac{\mu^2}{2}$
- $a(\phi) = \phi = \sigma^2$
- $c(y,\phi) = -\ln(\sqrt{2\pi\phi}) - \frac{y^2}{2\phi} = -\ln(\sigma\sqrt{2\pi}) - \frac{y^2}{2\sigma^2}$

A Bernoulli

A variável aleatória dicotómica (binária) Y diz-se de Bernoulli com parâmetro p, se toma valor 1 com probabilidade p e valor 0 com probabilidade $1-p$.

Para os valores $y=0$ ou $y=1$, a função de massa probabilística duma Bernoulli pode escrever-se como:

$$\begin{aligned} f(y|p) &= p^y(1-p)^{1-y} = e^{\ln p^y}\cdot e^{\ln(1-p)^{(1-y)}} = e^{y\ln p}\cdot e^{(1-y)\ln(1-p)} \\ &= e^{y\ln p+\ln(1-p)-y\ln(1-p)} = e^{y\ln\left(\frac{p}{1-p}\right)+\ln(1-p)} \end{aligned}$$

que é da familia exponencial $f(y\,|\,\theta,\phi) = e^{\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)}$, com:

- $\theta = \ln\left(\frac{p}{1-p}\right)$
- $\phi = 1$
- $b(\theta) = \ln\left(1+e^{\theta}\right) = \ln\left(1+\frac{p}{1-p}\right) = \ln\left(\frac{1}{1-p}\right) = -\ln(1-p)$
- $a(\phi) = 1$
- $c(y,\phi) = 0$

MLGs para variáveis resposta dicotómicas

Considere-se um Modelo com variável resposta dicotómica (binária), i.e., que apenas toma dois possíveis valores: 0 e 1, e cuja distribuição é Bernoulli:

$$Y = \begin{cases} 1 & , \quad p \\ 0 & , \quad 1-p \end{cases}$$

Admite-se que o parâmetro p varia nas n observações de Y.

O valor esperado da i-ésima observação de Y é (também) a probabilidade de êxito:

$$E[Y_i] = 1 \cdot p_i + 0 \cdot (1 - p_i) = p_i$$

Uma função de ligação relaciona a probabilidade de êxito p_i com uma combinação linear dos preditores:

$$g(E[Y_i]) = g(p_i) = \vec{\mathbf{x}}_{[i]}^t \vec{\boldsymbol{\beta}} \quad \Longleftrightarrow \quad E[Y_i] = p_i = g^{-1}\left(\vec{\mathbf{x}}_{[i]}^t \vec{\boldsymbol{\beta}}\right) ,$$

com $\vec{\mathbf{x}}_{[i]}^t \vec{\boldsymbol{\beta}} = \beta_0 + \beta_1 x_{1_i} + \beta_2 x_{2_i} + \ldots + \beta_p x_{p_i}$.

Funções de ligação

A mais simples é a ligação identidade: $g(\mu) = \mu$, que é a função ligação utilizada no Modelo Linear.

Cada distribuição da família exponencial, tem uma função de ligação que torna o valor esperado da variável resposta igual ao parâmetro natural, θ.

Função de ligação canónica

Num Modelo Linear Generalizado, a função $g(\cdot)$ diz-se uma função de ligação canónica para a variável resposta Y, se $g(E[Y]) = \theta$.

As funções de ligação canónica são úteis porque simplificam o estudo do Modelo. A ligação canónica representa de alguma forma uma função de ligação "natural" para o respectivo tipo de distribuição da variável resposta.

O Modelo Linear como um MLG

Eis alguns exemplos de MLGs:

O Modelo Linear

O Modelo Linear é um caso particular de MLG, em que:

- cada uma das n observações da variável resposta Y tem distribuição Normal, com variância constante σ^2;
- a função de ligação é a função identidade, pois:

$$\mu = E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \ .$$

A função de ligação identidade é a ligação canónica para a distribuição Normal, pois $\theta = \mu = E[Y]$.

A Regressão Logística

Regressão Logística

A função de ligação canónica da distribuição Bernoulli é o *logit*, que transforma $p = E[Y]$ no parâmetro natural $\theta = \ln\left(\frac{p}{1-p}\right)$:

$$g(p) = \ln\left(\frac{p}{1-p}\right),$$

Um MLG para variáveis resposta dicotómicas, com a função de ligação *logit* é conhecido por Regressão Logística.

A função de ligação *logit* é o logaritmo do quociente entre a probabilidade de Y tomar o valor 1 ("êxito") e a probabilidade de tomar o valor 0 ("fracasso"). Esse quociente é conhecido na literatura anglo-saxónica por *odds ratio*.

A função de ligação *logit* designa-se o *log-odds ratio*.

A Regressão Logística (cont.)

Dado um conjunto $\vec{\mathbf{x}} = (x_1, x_2, ..., x_p)$ de observações nas variáveis preditoras, tem-se com este modelo:

$$g(p) \;=\; \ln\left(\tfrac{p}{1-p}\right) \;=\; \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + ... + \beta_p\, x_p \;=\; \vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}$$

$$\Leftrightarrow \quad \tfrac{p}{1-p} = \mathrm{e}^{\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}} \quad \Leftrightarrow \quad \tfrac{1-p}{p} = \mathrm{e}^{-\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}} \quad \Leftrightarrow \quad \tfrac{1}{p} = 1 + \mathrm{e}^{-\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}} \quad \Leftrightarrow \quad p = \frac{1}{1 + \mathrm{e}^{-\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}}}.$$

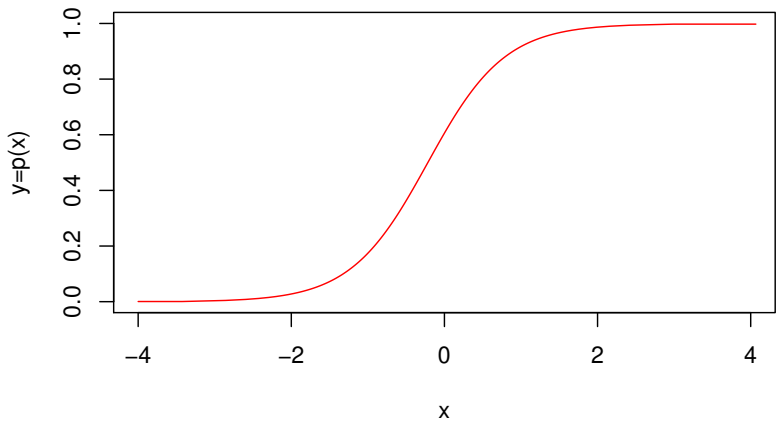
Numa Regressão Logística, a relação entre a probabilidade de êxito $p = E[Y]$ e os valores das variáveis preditoras, $\vec{\mathbf{x}}$, é:

$$p(\vec{\mathbf{x}}) \;=\; g^{-1}\left(\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}\right) \;=\; \frac{1}{1 + \mathrm{e}^{-\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}}} \;=\; \frac{1}{1 + \mathrm{e}^{-(\beta_0 + \beta_1\, x_1 + ... + \beta_p\, x_p)}}$$

A Regressão Logística (cont.)

No caso duma única variável preditora quantitativa, a relação entre Y e X é uma curva logística, daí a designação Regressão Logística.

$$p(x) = g^{-1}(\beta_0 + \beta_1 x) = \frac{1}{1 + \mathrm{e}^{-(\beta_0 + \beta_1 x)}}$$



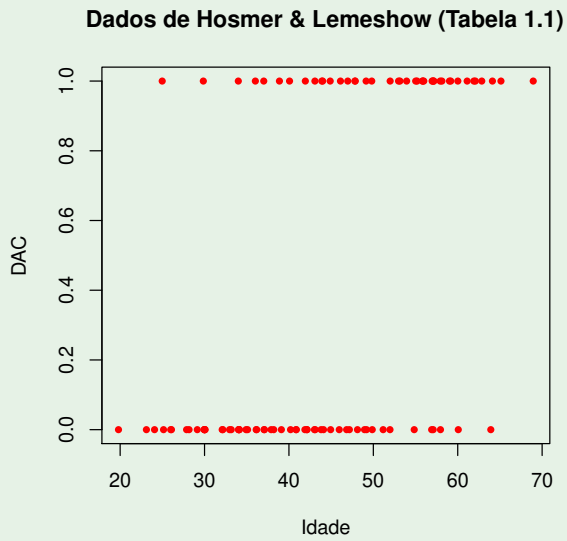
É uma função crescente, caso $\beta_1 > 0$, e decrescente caso $\beta_1 < 0$.

Quando há vários preditores, $p(\vec{\mathbf{x}}) = \frac{1}{1+\mathrm{e}^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}}$ define uma hipersuperfície em $\mathbb{R}^{p+1}$.

Novamente o exemplo DAC

Dados Hosmer & Lemeshow

- idade – numérica;
- doença arterial coronária – variável dicotómica (sim/não; 1/0).



A variável resposta é dicotómica (binária). Será preciso relacionar $p = E[Y]$, a probabilidade de ter doença arterial coronária, com a idade X.

Exemplo: A função de ligação

Dados Hosmer & Lemeshow (cont.)

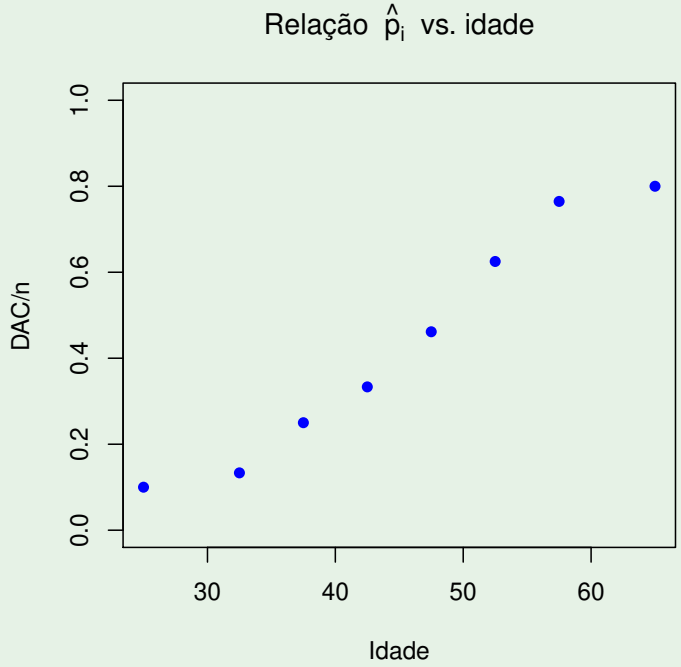
Para procurar uma função de ligação adequada, é necessário outro gráfico, para visualizar a relação entre idade e probabilidade de DAC.

- Havendo repetições para cada idade, pode estimar-se p_i a partir da frequência relativa de DAC na i-ésima idade;
- Havendo poucas repetições em cada idade, pode-se agrupar as observações em classes de idade.

Classe	n_i	DAC	$\hat{p}_i$
20-30-	10	1	0.100
30-35-	15	2	0.133
35-40-	12	3	0.250
40-45-	15	5	0.333
45-50-	13	6	0.462
50-55-	8	5	0.635
55-60-	17	13	0.765
60-70-	10	8	0.800

Exemplo: $\hat{p}_i$ *vs.* idade

Eis o gráfico das probabilidades estimadas vs. idade:



Nota: Aqui "idade" é o ponto médio de cada classe de idade.

Temos uma relação sigmóide. Talvez logística?

Exemplo 1: criando a tabela no 

Para criar a tabela do acetato 24, usou-se a função `hist`, que permite agrupar as idades em classes de idade. Com o argumento `plot=FALSE`, em vez de produzir o histograma, o comando devolve a informação usada.
O argumento `right=FALSE` abre à direita os intervalos das classes.

Vamos começar por definir as fronteiras das classes.

```
> frclass <- c(20,30,35,40,45,50,55,60,70)
> hist(HL$Idade, breaks=frclass, plot=FALSE, right=FALSE)
$breaks     <--  fronteiras das classes
[1] 20 30 35 40 45 50 55 60 70
$counts     <-- frequências absolutas
[1] 10 15 12 15 13  8 17 10
$density    <-- alturas rectângulos histograma
[1] 0.010 0.030 0.024 0.030 0.026 0.016 0.034 0.010
$mids       <-- pontos intermédios de cada classe
[1] 25.0 32.5 37.5 42.5 47.5 52.5 57.5 65.0
[...]
```

Exemplo 1: a tabela (cont.)

Vamos seleccionar as componentes `counts` e `mids` para construir duas colunas da tabela:

```
> info <- hist(HL$Idade, breaks=frclass, plot=FALSE, right=FALSE)
> HL.tab <- data.frame(idade=info$mids, nobs=info$counts)
> HL.tab
  idade nobs
1  25.0   10
2  32.5   15
3  37.5   12
4  42.5   15
5  47.5   13
6  52.5    8
7  57.5   17
8  65.0   10
```

Falta a coluna com o número de pacientes com DAC em cada classe.
Para a obter, repetimos procedimentos, mas seleccionando apenas as linhas de `HL` em que `DAC` tem valor 1.

Exemplo 1: os dados tabelados

Eis o resto da tabela, incluindo os nomes de linhas:

```
> HL.tab$DAC <-        <-- cria uma nova coluna de nome DAC
+     hist(HL$Idade[HL$DAC==1], breaks=frclass, plot=F, right=F)$counts

> rownames(HL.tab) <-   <-- atribui os nomes de linhas da tabela
+           paste("[" , frclass[-9] , "," , frclass[-1] , "[" , sep="")

> HL.tab
```

	idade	nobs	DAC
[20,30[	25.0	10	1
[30,35[	32.5	15	2
[35,40[	37.5	12	3
[40,45[	42.5	15	5
[45,50[	47.5	13	6
[50,55[	52.5	8	5
[55,60[	57.5	17	13
[60,70[	65.0	10	8

O gráfico no acetato 25 foi obtido com o comando:

```
> plot(DAC/nobs ~ idade , ylim=c(0,1) , data=HL.tab ,
+     main=expression(paste("Relação ", hat(p)[i] , "vs. idade")), pch=16, col="blue")
```

Resposta dicotómica e Binomial

Agrupar as idades em classes transforma a variável resposta Bernoulli (1/0) Y_i, numa variável resposta Y_j que conta, em cada classe j, o número de "êxitos" (uns) nas n_j provas de Bernoulli dessa classe.

Para observações independentes, Y_j tem distribuição Binomial: $Y_j \frown B(n_j, p_j)$, onde p_j é a probabilidade de "êxito" na classe j.

A Binomial não é da família exponencial. Mas a proporção de êxitos em n provas de Bernoulli, $W = Y/n$, sim (se n conhecido).

Existe ligação íntima, no contexto de MLGs, entre considerar que temos:

- n observações de variáveis resposta Bernoulli, com parâmetros p_i; ou
- m observações de variáveis resposta 'Binomial/n' $W_j = \frac{Y_j}{n_j}$, com $Y_j \frown B(n_j, p_j)$.

A ligação canónica, quer da Bernoulli, quer da Binomial/n é a função *logit*:

$$g(p) = \ln\left(\frac{p}{1-p}\right)$$

A 'Binomial/n'

Se $X \frown B(n,p)$, tem-se $P[X=x] = \binom{n}{x} p^x (1-p)^{n-x}$.
Então $Y = \frac{1}{n}X$ tem distribuição na família exponencial.

Tem-se $P[Y=y] = P[X=ny]$. A função de massa probabilística de Y pode escrever-se da seguinte forma, para $y \in \{0, \frac{1}{n}, \frac{2}{n}, ..., 1\}$:

$$\begin{aligned} f(y|p) &= \binom{n}{ny} p^{ny}(1-p)^{n-ny} = e^{\ln\left[\binom{n}{ny}\right]} \cdot e^{ny \ln p} \cdot e^{(n-ny)\ln(1-p)} \\ &= e^{\ln\left[\binom{n}{ny}\right] + ny \ln p + n\ln(1-p) - ny\ln(1-p)} = e^{\frac{y\ln\left(\frac{p}{1-p}\right)+\ln(1-p)}{\frac{1}{n}} + \ln\left[\binom{n}{ny}\right]} \end{aligned}$$

que é da familia exponencial $f(y\,|\,\theta,\phi\,) = e^{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)}$, com:

- $\theta = \ln\left(\frac{p}{1-p}\right)$
- $\phi = 1$
- $b(\theta) = \ln\left(1+e^{\theta}\right) = \ln\left(1+\frac{p}{1-p}\right) = \ln\left(\frac{1}{1-p}\right) = -\ln(1-p)$
- $a(\phi) = \frac{\phi}{n} = \frac{1}{n}$
- $c(y,\phi) = \ln\left[\binom{n}{ny}\right]$

GLMs no R

No R, o comando para ajustar Modelos Lineares Generalizados é `glm`.

Há três argumentos fundamentais nesta função:

formula indica a componente aleatória (variável resposta) e a componente sistemática (preditores), de forma análoga à usada no modelo linear:

$$y \quad \sim \quad x1 + x2 + x3 + ... + xp$$

family indica simultaneamente a distribuição de probabilidades da componente aleatória Y e a função de ligação do modelo.

data indica a *data frame* onde se encontram as variáveis.

GLMs no R (cont.)

A indicação da distribuição de probabilidades de Y faz-se através duma palavra-chave, que se segue ao nome do argumento `family`.

Por exemplo, um modelo com componente aleatória Bernoulli ou Binomial/n, indica-se assim:

```
family = binomial
```

Por omissão, é usada a função de ligação canónica dessa distribuição.

Caso se deseje outra função de ligação (implementada) acrescenta-se ao nome da distribuição, entre parenteses, o argumento `link` com a especificação da função de ligação.

Por exemplo, um modelo probit pode ser indicado da seguinte forma:

```
family = binomial(link=probit)
```

Exemplo: o ajustamento do modelo

Assim, ajusta-se um MLG no R invocando o comando `glm` com três argumentos:

`glm`(*formula, family, data*)

Numa Regressão Logística,

- `family=binomial`.

  Não é necessário especificar a função de ligação: por omissão é usada a ligação canónica da distribuição especificada.

- podem usar-se dados numa de 2 formas:
  - observações dicotómicas individuais (como a *data frame* `HL`);
  - observações tabeladas para valores repetidos do(s) preditor(es) (como a *data frame* `HL.tab`).

As formulas para a Regressão Logística

As fórmulas do comando `glm` são semelhantes às do Modelo Linear:

$$y \quad \sim \quad x1 + x2 + ... + xp$$

Mas numa Regressão Logística, aos dois tipos de dados possíveis correspondem objectos y de natureza diferente:

- Se dados contêm observações individuais, y é vector de 0s e 1s:

  > glm(DAC ∼ Idade , family=binomial , data=HL)

- Se os dados estão tabelados, y deve ser uma matriz de duas colunas: uma com o número de "sim"s e outra com os número de "não"s, para cada valor (ou ponto médio da classe) do(s) preditor(es):

  > glm(cbind(DAC,nobs-DAC) ∼ idade , family=binomial, data=HL.tab)

De novo o exemplo de Hosmer & Lemeshow

Dados Hosmer & Lemeshow (cont.)

Ajustar o modelo com base nas observações dicotómicas individuais:

> glm(DAC ~ Idade , family=binomial, data=HL)

```
Call:  glm(formula = DAC ~ Idade , family = binomial , data = HL)
Coefficients:
(Intercept)         Idade
    -5.3095        0.1109          <---- parâmetros estimados
```

A equação da logística ajustada é:

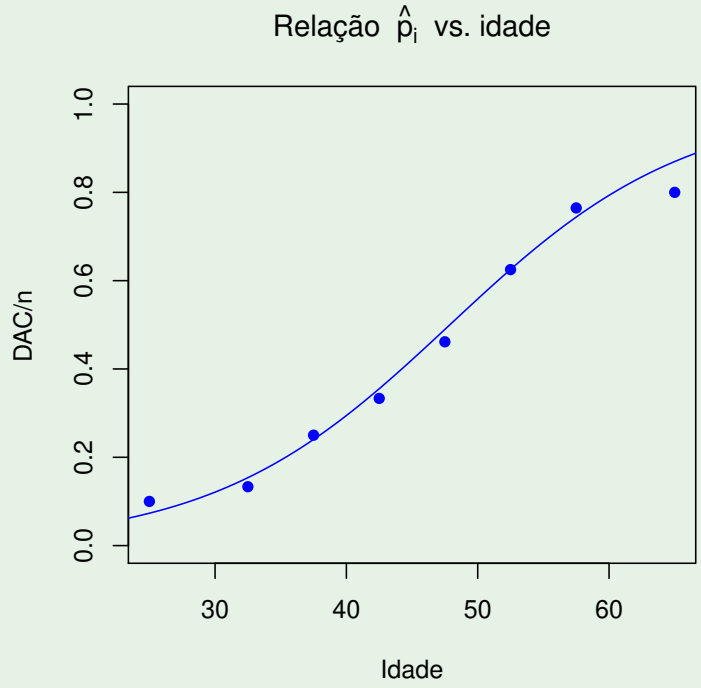
$$y = \frac{1}{1+e^{-(b_0+b_1 x)}} = \frac{1}{1+e^{-(-5.3095+0.1109x)}}$$

Exemplo (cont.)

Dados Hosmer & Lemeshow (cont.)

Sobrepondo a logística ajustada ao gráfico dos $\hat{p}_i$ *vs.* idade:

```
> logistica <- function( b0 , b1 , x ){ 1/(1+exp(-(b0+b1*x))) }
> curve(logistica(b0=-5.3095, b1=0.1109, x), from=20, to=70, col="blue", add=TRUE)
```



Exemplo: ajustamento do modelo (cont.)

Dados Hosmer & Lemeshow (cont.)

Ajustar o modelo com base nos dados tabelados:

> glm(cbind(DAC,nobs-DAC) ~ idade , family=binomial, data=HL.tab)

```
Call: glm(formula=cbind(DAC,nobs-DAC)~idade, family=binomial,data=HL.tab)
Coefficients:
(Intercept)         idade
     -5.091         0.105          <----  parâmetros estimados
```

A equação da logística ajustada é:

$$y = \frac{1}{1+\mathrm{e}^{-(b_0+b_1 x)}} = \frac{1}{1+\mathrm{e}^{-(-5.091+0.105x)}}$$

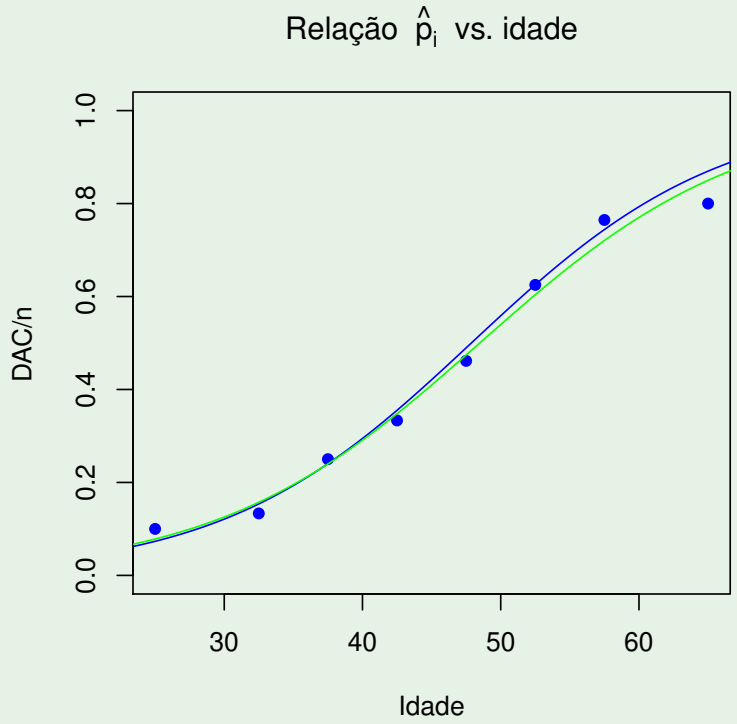
Nota: A pequena discrepância em relação ao ajustamento anterior resulta do agrupamento em classes de idade: os dados são diferentes.

Exemplo: ajustamento do modelo (cont.)

Dados Hosmer & Lemeshow (cont.)

Sobrepondo a logística ajustada ao gráfico dos $\hat{p}_i$ *vs.* idade (e à curva ajustada antes):

```
> curve(logistica(b0=-5.091, b1=0.105, x), from=20, to=70, col="green", add=TRUE)
```



O resultado do comando `glm`

Tal como o comando `lm`, também o comando `glm` produz uma *list*. Nas componentes dessa lista há informação sobre o ajustamento.

```
> HLtab.glm <- glm(cbind(DAC,nobs-DAC)~idade,family=binomial,data=HL.tab)
> names(HLtab.glm)

 [1] "coefficients"      "residuals"         "fitted.values"     "effects"
 [5] "R"                 "rank"              "qr"                "family"
 [9] "linear.predictors" "deviance"          "aic"               "null.deviance"
[13] "iter"              "weights"           "prior.weights"     "df.residual"
[17] "df.null"           "y"                 "converged"         "boundary"
[21] "model"             "call"              "formula"           "terms"
[25] "data"              "offset"            "control"           "method"
[29] "contrasts"         "xlevels"
```

Para aprofundar cada componente consultar: `help(glm)`

Para invocar uma componente usa-se a referenciação usual de listas:

Dados Hosmer & Lemeshow (cont.)

```
> HLtab.glm$coef

(Intercept)       idade
 -5.0907332   0.1050191
```

O comando `coef`

Como para os Modelos Lineares, existem comandos que facilitam a extracção de informação dum ajustamento de MLG. Eis algumas:

`coef` – devolve um vector com os valores estimados dos parâmetros β_0, β_1, ...,β_p, ou seja, com os valores b_0, b_1, ..., b_p:

Dados Hosmer & Lemeshow (cont.)

```
> HL.glm <- glm(DAC ~ Idade, family=binomial, data=HL)
> coef(HL.glm)

 (Intercept)        idade
  -5.3094534    0.1109211
```

O comando `predict`

`predict` – por omissão, devolve os valores da combinação linear estimada dos preditores usados no ajustamento, ou seja, da componente sistemática $b_0 + b_1 x_{1_{(i)}} + \ldots + b_p x_{p_{(i)}}$.

Dados Hosmer & Lemeshow (cont.)

```
> predict(HLtab.glm)

   [20,30[    [30,35[    [35,40[    [40,45[    [45,50[    [50,55[    [55,60[    [60,70[
-2.4652550 -1.6776115 -1.1525158 -0.6274202 -0.1023245  0.4227711  0.9478668  1.7355102
```

```
> predict(HL.glm)

          1           2           3           4           5           6           7
-3.09103053 -2.75826710 -2.64734596 -2.53642482 -2.53642482 -2.42550368 -2.42550368
.....
         99         100
 1.90042087  2.34410544
```

O comando `predict` (cont.)

Pode também usar-se para estimar a combinação linear de valores não usados no ajustamento.

Os novos valores são dados numa *data frame* com nomes iguais aos usados nos dados originais.

Dados Hosmer & Lemeshow (cont.)

```
> predict(HL.glm, newdata=data.frame(Idade=26))

        1
-2.425504

> predict(HLtab.glm, newdata=data.frame(idade=c(26,53,74)))

         1          2          3
-2.3602358  0.4752807  2.6806824
```

O comando `fitted`

`fitted` – devolve os valores ajustados do valor esperado de Y_i, ou seja, de $\hat{p}_i = g^{-1}(b_0 + b_1 x_{1_{(i)}} + \ldots + b_p x_{p_{(i)}})$.

```
> fitted(HLtab.glm)

   [20,30[    [30,35[    [35,40[    [40,45[    [45,50[    [50,55[    [55,60[    [60,70[
0.07833012 0.15741201 0.24002985 0.34809573 0.47444116 0.60414616 0.72068596 0.85011588
```

Resultado análogo é obtido com o comando `predict`, usando o argumento `type="response"`:

```
> predict(HLtab.glm, type="response")

   [20,30[    [30,35[    [35,40[    [40,45[    [45,50[    [50,55[    [55,60[    [60,70[
0.07833012 0.15741201 0.24002985 0.34809573 0.47444116 0.60414616 0.72068596 0.85011588
```

Assim, pode-se estimar os valores de $\hat{p}$ para novos valores dos preditores.

```
> predict(HLtab.glm, newdata=data.frame(idade=c(26,53,74)),type="response")

        1         2         3
0.0862556 0.6166329 0.9358771
```

Notas sobre a Regressão Logística

- A função logística tem boas propriedades para representar uma probabilidade: para *qualquer* valor da componente sistemática, a função logística

$$p(x_1, x_2, ..., x_p) = \frac{1}{1 + \mathrm{e}^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}}$$

  toma valores entre 0 e 1. O mesmo não acontece com uma relação linear $p(x_1, ..., x_p) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$, que toma valores em $\mathbb{R}$.

- No caso de haver uma única variável preditora quantitativa, trocando os acontecimentos associados aos valores 0 e 1, uma função decrescente para $p = P[Y = 1]$ transforma-se numa função crescente.

Mais notas sobre a regressão logística

No caso de haver uma única variável preditora quantitativa, o parâmetro β_1 tem a seguinte interpretação:

- como

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x} = e^{\beta_0} \cdot e^{\beta_1 x} ,$$

cada aumento de uma unidade na variável preditora X traduz-se num efeito multiplicativo sobre o *odds ratio*, de e^{β_1} :

$$\frac{p(x+1)}{1-p(x+1)} = e^{\beta_0} \cdot e^{\beta_1 (x+1)} = e^{\beta_0} \cdot e^{\beta_1 x} \cdot e^{\beta_1} = \frac{p(x)}{1-p(x)} \cdot e^{\beta_1} .$$

- o que é o mesmo que dizer que se traduz num efeito aditivo, de β_1 unidades, sobre o *log-odds ratio*:

$$\log\left[\frac{p(x+1)}{1-p(x+1)}\right] = \log\left[\frac{p(x)}{1-p(x)}\right] + \beta_1 .$$

Mais notas sobre a Regressão Logística

Quando há mais do que uma variável preditora quantitativa:

- a função de ligação *logit* gera uma relação logística para a probabilidade de êxito p, como função dos valores da parte sistemática η (combinação linear das variáveis preditoras).
- a interpretação dos coeficientes β_j generaliza-se: um aumento de uma unidade na variável preditora j (mantendo as restantes constantes) traduz-se numa multiplicação do *odds ratio* por um factor e^{β_j}.

Para preditores categóricos (factores),

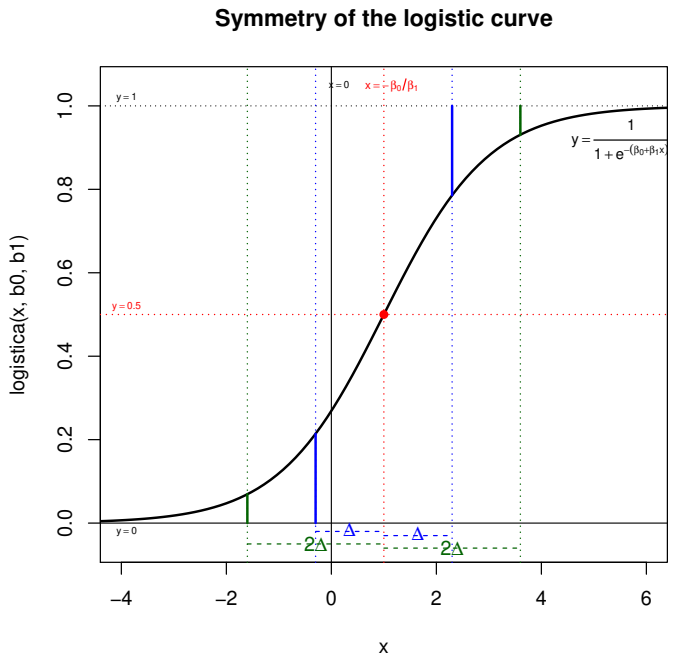
- seja $\vec{\mathscr{I}}_j$ uma variável indicatriz. O correspondente parâmetro β_j indica o incremento no *log-odds ratio* resultante de uma observação passar a pertencer à categoria de que $\vec{\mathscr{I}}_j$ é indicatriz.

Características da curva logística

A relação logística pode ser demasiado rígida. Com um único preditor numérico, a logística $f(x) = \left[1 + \mathrm{e}^{-(\beta_0 + \beta_1 x)}\right]^{-1}$ tem um ponto de inflexão associado à probabilidade $p = 0.5$, em torno do qual há simetria da curva:

$$f''(x) = 0 \quad \Leftrightarrow \quad \mathrm{e}^{-(\beta_0+\beta_1 x)} = 1 \quad \Leftrightarrow \quad x = -\frac{\beta_0}{\beta_1} \quad \Rightarrow \quad f\left(-\frac{\beta_0}{\beta_1}\right) = \frac{1}{2} \, .$$

$$f\left(-\frac{\beta_0}{\beta_1} - \Delta\right) = \frac{1}{1+\mathrm{e}^{\beta_1 \Delta}} = 1 - f\left(-\frac{\beta_0}{\beta_1} + \Delta\right)$$



Estimação de parâmetros em MLGs

A estimação dos parâmetros β_j em Modelos Lineares Generalizados é feita pelo Método da Máxima Verosimilhança.

O facto da estimação se basear na função verosimilhança significa que, ao contrário do que acontece com o Modelo Linear, em MLGs as hipóteses distribucionais são cruciais para a estimação dos parâmetros.

O facto das distribuições consideradas em MLGs pertencerem à família exponencial de distribuições gera algumas particularidades na estimação.

Verosimilhança na família exponencial

A função verosimilhança para n observações independentes $y_1, y_2, ..., y_n$ numa qualquer distribuição da família exponencial é:

$$\mathbf{L}(\vec{\theta}, \vec{\phi}\ ;\ y_1, y_2, ..., y_n) = \prod_{i=1}^{n} f(y_i; \theta_i, \phi_i) = e^{\sum_{i=1}^{n}\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i)\right]}$$

Maximizar a verosimilhança equivale a maximizar a log-verosimilhança:

$$\mathscr{L}(\vec{\theta}, \vec{\phi}\ ;\ y_1, y_2, ..., y_n) = \sum_{i=1}^{n}\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i)\right]$$

Onde estão os parâmetros do modelo, β_j?

Máxima Verosimilhança em MLGs

Num MLG, a componente sistemática e o valor esperado da variável resposta estão relacionados por $g(E[Y]) = \vec{\mathbf{x}}^t \vec{\boldsymbol{\beta}}$.

No caso de uma função de ligação canónica tem-se $\theta = \vec{\mathbf{x}}^t \vec{\boldsymbol{\beta}}$.

Em geral, pode escrever-se a log-verosimilhança como função dos parâmetros do modelo: $\mathscr{L}(\vec{\boldsymbol{\beta}})$.

Estimar os parâmetros pelo método da máxima verosimilhança consiste em escolher o vector $\vec{\boldsymbol{\beta}}$ que torne máxima a função de log-verosimilhança $\mathscr{L}(\vec{\boldsymbol{\beta}})$.

Máxima Verosimilhança em MLGs (cont.)

A maximização da função de $p+1$ variáveis $\mathscr{L}(\vec{\boldsymbol{\beta}})$ tem como condição necessária:

$$\frac{\partial \mathscr{L}(\vec{\boldsymbol{\beta}})}{\partial \beta_j} = 0, \qquad \forall\, j = 0 : p$$

Admite-se que as funções $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$ são suficientemente regulares para que as operações envolvidas estejam bem definidas.

No caso de um Modelo Linear Generalizado genérico, não existe a garantia de que haja máximo desta função log-verosimilhança (para valores admissíveis dos parâmetros $\vec{\boldsymbol{\beta}}$), nem que, existindo máximo, este seja único.

Nos casos concretos abordados nesta disciplina, a situação não cria dificuldades.

Exemplo: o caso da Regressão Logística

No Modelo de Regressão Logística, as n observações independentes referem-se a uma Variável aleatória com distribuição de Bernoulli.
A sua função de verosimilhança é dada por:

$$\mathbf{L}(\vec{\mathbf{p}}\ ;\ \vec{\mathbf{y}}) = \prod_{i=1}^{n} \mathrm{e}^{\ln(1-p_i)+y_i \ln\left(\frac{p_i}{1-p_i}\right)}$$

e a log-verosimilhança por:

$$\mathscr{L}(\vec{\mathbf{p}}\ ;\ \vec{\mathbf{y}}) = \sum_{i=1}^{n} \left[\ln(1-p_i) + y_i \ln\left(\frac{p_i}{1-p_i}\right)\right]$$

Como a função de ligação é $g(p) = \ln\left(\frac{p}{1-p}\right) = \vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}$, a log-verosimilhança é função dos parâmetros $\vec{\boldsymbol{\beta}}$:

$$\mathscr{L}(\vec{\boldsymbol{\beta}}\ ;\ \vec{\mathbf{y}}) \ = \ \sum_{i=1}^{n} \left[-\ln\left(1+\mathrm{e}^{\vec{\mathbf{x}}_i^t\vec{\boldsymbol{\beta}}}\right) + y_i\,\vec{\mathbf{x}}_i^t\vec{\boldsymbol{\beta}}\right]$$

Estimação na Regressão Logística (cont.)

Tem-se:

$$\mathscr{L}(\vec{\boldsymbol{\beta}}) = \sum_{i=1}^{n}\left(\beta_0 y_i + \sum_{k=1}^{p} y_i\, x_{k(i)}\beta_k\right) - \sum_{i=1}^{n} \ln\left(1+\mathrm{e}^{\beta_0+\Sigma_{k=1}^{p} x_{k(i)}\beta_k}\right)$$

Condição necessária para a existência de extremo da log-verosimilhança no ponto $\vec{\boldsymbol{\beta}} = \hat{\vec{\boldsymbol{\beta}}}$ é que:

$$\begin{cases} \dfrac{\partial \mathscr{L}(\hat{\vec{\boldsymbol{\beta}}})}{\partial \beta_0} &= \displaystyle\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{\mathrm{e}^{\hat{\beta}_0+\Sigma_{k=1}^{p} x_{k(i)}\hat{\beta}_k}}{1+\mathrm{e}^{\hat{\beta}_0+\Sigma_{k=1}^{p} x_{k(i)}\hat{\beta}_k}} \cdot 1 = 0 \\ \dfrac{\partial \mathscr{L}(\hat{\vec{\boldsymbol{\beta}}})}{\partial \beta_j} &= \displaystyle\sum_{i=1}^{n} y_i x_{j(i)} - \sum_{i=1}^{n} \frac{\mathrm{e}^{\hat{\beta}_0+\Sigma_{k=1}^{p} x_{k(i)}\hat{\beta}_k}}{1+\mathrm{e}^{\hat{\beta}_0+\Sigma_{k=1}^{p} x_{k(i)}\hat{\beta}_k}} \cdot x_{j(i)} = 0 \qquad \forall j = 1:p \end{cases}$$

Estas $p+1$ equações normais formam um sistema não-linear de equações nas $p+1$ incógnitas $\hat{\beta}_j$ $(j = 0:p)$.

Algoritmos de estimação

Em geral, o sistema de $p+1$ equações normais associado à maximização da função de log-verosimilhança num Modelo Linear generalizado é um sistema não-linear:

$$\frac{\partial \mathscr{L}(\vec{\boldsymbol{\beta}})}{\partial \beta_j} = 0 \qquad , \qquad \forall\, j = 0 : p.$$

Algoritmos numéricos de resolução utilizados no contexto de MLGs são adaptações do algoritmo de Newton-Raphson, conhecidos por vários nomes: Método Iterativo de Mínimos Quadrados Ponderados (Iterative Weighted Least Squares, IWLS) ou Re-ponderados (or Reweighted, IRLS), ou ainda Método de Fisher (*Fisher Scoring Method*).

O Método de Newton-Raphson trabalha com uma aproximação de segunda ordem (fórmula de Taylor) da função log-verosimilhança , com desenvolvimento em torno duma estimativa inicial do vector $\vec{\boldsymbol{\beta}}$.

Algoritmos de estimação (cont.)

Método de Newton-Raphson

Sejam:

- $\vec{\boldsymbol{\beta}}^{[0]}$, uma solução inicial para $\vec{\boldsymbol{\beta}}$;
- $\vec{\nabla}\mathscr{L}_{\vec{\boldsymbol{\beta}}}$ o vector gradiente de $\mathscr{L}$ (vector das derivadas parciais $\frac{\partial \mathscr{L}}{\partial \beta_j}$) em $\vec{\boldsymbol{\beta}}$;
- $\mathscr{H}_{\vec{\boldsymbol{\beta}}}$ a matriz Hessiana das segundas derivadas parciais da função $\mathscr{L}(\cdot)$

Tem-se a aproximação de 2a. ordem dada pela fórmula de Taylor:

$$\mathscr{L}(\vec{\boldsymbol{\beta}}) \approx \mathscr{L}_*(\vec{\boldsymbol{\beta}}) = \mathscr{L}(\vec{\boldsymbol{\beta}}^{[0]}) + (\vec{\nabla}\mathscr{L}_{\vec{\boldsymbol{\beta}}^{[0]}})^t(\vec{\boldsymbol{\beta}} - \vec{\boldsymbol{\beta}}^{[0]}) + \frac{1}{2}(\vec{\boldsymbol{\beta}} - \vec{\boldsymbol{\beta}}^{[0]})^t \mathscr{H}_{\vec{\boldsymbol{\beta}}^{[0]}}(\vec{\boldsymbol{\beta}} - \vec{\boldsymbol{\beta}}^{[0]})$$

Em vez de maximizar $\mathscr{L}(\vec{\boldsymbol{\beta}})$, maximiza-se a aproximação $\mathscr{L}_*(\vec{\boldsymbol{\beta}})$.

Algoritmos de estimação (cont.)

O cálculo do vector gradiente é simples para produtos internos ou formas quadráticas:

Se $h(\vec{\mathbf{x}}) = \vec{\mathbf{a}}^t\vec{\mathbf{x}}$, tem-se $\frac{\partial h(\vec{\mathbf{x}})}{\partial \vec{\mathbf{x}}} = \frac{\partial(\vec{\mathbf{a}}^t\vec{\mathbf{x}})}{\partial \vec{\mathbf{x}}} = \vec{\mathbf{a}}$.

Se $h(\vec{\mathbf{x}}) = \vec{\mathbf{x}}^t\mathbf{A}\vec{\mathbf{x}}$, tem-se $\frac{\partial h(\vec{\mathbf{x}})}{\partial \vec{\mathbf{x}}} = \frac{\partial(\vec{\mathbf{x}}^t\mathbf{A}\vec{\mathbf{x}})}{\partial \vec{\mathbf{x}}} = 2\mathbf{A}\vec{\mathbf{x}}$.

Assim,

$$\nabla\vec{\mathscr{L}}_{*_{\vec{\boldsymbol{\beta}}}} = \nabla\vec{\mathscr{L}}_{\vec{\boldsymbol{\beta}}^{[0]}} + \mathscr{H}_{\vec{\boldsymbol{\beta}}^{[0]}}(\vec{\boldsymbol{\beta}} - \vec{\boldsymbol{\beta}}^{[0]}).$$

Admitindo a invertibilidade de $\mathscr{H}_{\vec{\boldsymbol{\beta}}^{[0]}}$, tem-se:

$$\nabla\vec{\mathscr{L}}_{*_{\vec{\boldsymbol{\beta}}}} = \vec{\mathbf{0}} \quad \Leftrightarrow \quad \vec{\boldsymbol{\beta}} = \vec{\boldsymbol{\beta}}^{[0]} - \mathscr{H}^{-1}_{\vec{\boldsymbol{\beta}}^{[0]}} \cdot \nabla\vec{\mathscr{L}}_{\vec{\boldsymbol{\beta}}^{[0]}}.$$

O algoritmo Newton-Raphson itera esta relação.

Algoritmos de estimação (cont.)

Tome-se:

$$\vec{\boldsymbol{\beta}}^{[i+1]} = \vec{\boldsymbol{\beta}}^{[i]} - \mathscr{H}^{-1}_{\vec{\boldsymbol{\beta}}^{[i]}} \cdot \vec{\nabla}\mathscr{L}_{\vec{\boldsymbol{\beta}}^{[i]}}$$

Notas:

- A possibilidade de aplicar com êxito este algoritmo exige a existência e invertibilidade das matrizes Hessianas de $\mathscr{L}$ nos sucessivos pontos $\vec{\boldsymbol{\beta}}^{[i]}$;
- Não está garantida a convergência do algoritmo, mesmo quando existe e é único o máximo da função log-verosimilhança;
- Caso exista um único máximo, a convergência é tanto melhor quanto mais próximo $\vec{\boldsymbol{\beta}}^{[0]}$ estiver do máximo.
- Podem existir vários máximos locais, e uma má escolha inicial $\vec{\boldsymbol{\beta}}^{[0]}$ levar à convergência para uma solução sub-óptima.

Algoritmos de estimação (cont.)

Método de Fisher

O cálculo da matriz Hessiana da log-verosimilhança nos pontos $\vec{\boldsymbol{\beta}}^{[i]}$ é computacionalmente exigente.

O algoritmo de Fisher (Fisher Scoring Method) é uma modificação do algoritmo de Newton-Raphson, que substitui a matriz Hessiana pela matriz de informação de Fisher, definida como o simétrico da esperança da matriz Hessiana:

$$\mathbf{I}_{\vec{\boldsymbol{\beta}}^{[i]}} = -E\left[\mathscr{H}_{\vec{\boldsymbol{\beta}}^{[i]}}\right]$$

Assim, a iteração que está na base do Algoritmo de Fisher é:

$$\vec{\boldsymbol{\beta}}^{[i+1]} = \vec{\boldsymbol{\beta}}^{[i]} + \mathbf{I}^{-1}_{\vec{\boldsymbol{\beta}}^{[i]}} \cdot \vec{\nabla}\mathscr{L}_{\vec{\boldsymbol{\beta}}^{[i]}}$$

Algoritmos (cont.)

Quando se considera uma MLG com a função de ligação canónica, a matriz Hessiana da log-verosimilhança não depende da variável resposta Y, pelo que a Hessiana e o seu valor esperado coincidem.

Logo, neste caso os métodos de Fisher e Newton-Raphson coincidem.

Esta é uma das razões que confere às ligações canónicas a sua importância.

A Regressão Probit

Outro exemplo de MLG é o modelo probit de Bliss (1935), muito frequente em Toxicologia.

Modelo *Probit*

Tal como na Regressão Logística, tem-se:

- variável resposta dicotómica (com distribuição Bernoulli).
- componente sistemática, dada por uma combinação linear de variáveis preditoras.

Diferente da Regressão Logística é a função de ligação.

A Regressão Probit (cont.)

Na Regressão Logística, a função de ligação exprime $p = E[Y]$ como uma função logística da componente sistemática $\eta = \vec{\mathbf{x}}^t \vec{\boldsymbol{\beta}}$.

No Modelo *Probit*, a probabilidade de êxito e valor esperado, $p = E[Y]$, é dada por outra curva sigmóide, a função de distribuição cumulativa (f.d.c.), Φ, duma Normal Reduzida:

$$p(\vec{\mathbf{x}}) \quad = \quad g^{-1}\left(\vec{\mathbf{x}}^t \vec{\boldsymbol{\beta}}\right) \quad = \quad \Phi(\vec{\mathbf{x}}^t \vec{\boldsymbol{\beta}})$$

onde Φ indica a f.d.c. duma $\mathcal{N}(0,1)$.

Esta opção significa considerar como função de ligação a inversa da f.d.c. duma Normal reduzida, ou seja, $g = \Phi^{-1}$:

$$\vec{\mathbf{x}}^t \vec{\boldsymbol{\beta}} \; = \; \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \; = \; g(E[Y]) \; = \; g\left(p(\vec{\mathbf{x}})\right) \; = \; \Phi^{-1}\left(p(\vec{\mathbf{x}})\right) \; .$$

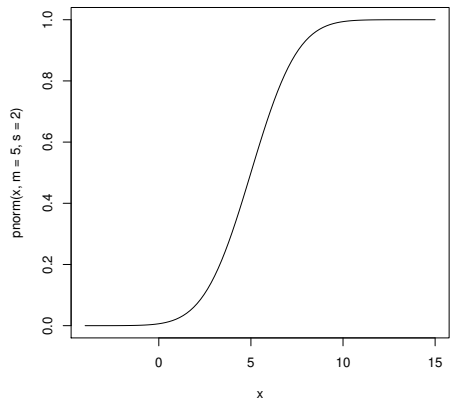
A Regressão Probit (cont.)

No caso de haver uma única variável preditora numérica, tem-se:

$$p(x;\beta_0,\beta_1) \;=\; g^{-1}\left(\beta_0+\beta_1\, x\right) \;=\; \Phi(\beta_0+\beta_1\, x) \;=\; \Phi\left(\frac{x-\mu}{\sigma}\right)\;,$$

com $\beta_0 = -\frac{\mu}{\sigma}$ e $\beta_1 = \frac{1}{\sigma}$.

Assim, a probabilidade de êxito p relaciona-se com a variável preditora X através da f.d.c. duma $\mathcal{N}(\mu,\sigma^2)$, com $\sigma = \frac{1}{\beta_1}$ e $\mu = -\frac{\beta_0}{\beta_1}$.



A Regressão Probit em toxicologia

No contexto toxicológico, é frequente:

- existir uma variável preditora X que indica a dosagem (ou log-dosagem) dum determinado produto tóxico;
- para cada individuo há um nível de tolerância t: a dosagem acima do qual o produto tóxico provoca a morte do indivíduo;
- esse nível de tolerância varia entre indivíduos e pode ser representado por uma variável aleatória T.

Definindo a variável aleatória binária Y:

$$Y = \begin{cases} 1 & , \quad \text{individuo morre} \\ 0 & , \quad \text{individuo sobrevive} \end{cases}$$

Tem-se:

$$P[\ Y = 1 \mid x\ ] = P[T \leq x] = p(x)$$

A Regressão Probit em toxicologia (cont.)

$$P[\ Y = 1 \mid x\ ] \quad = \quad P[T \leq x] \quad = \quad p(x)$$

Admitindo que a tolerância T segue uma distribuição $\mathcal{N}(\mu, \sigma^2)$,

$$p(x) \quad = \quad \Phi\left(\frac{x-\mu}{\sigma}\right) .$$

Tem-se o Modelo Probit com X como única variável preditora.

Como $\beta_0 = -\frac{\mu}{\sigma}$ e $\beta_1 = \frac{1}{\sigma}$ (slide 62), os parâmetros da distribuição da tolerância T são dados por:

$$\mu = -\frac{\beta_0}{\beta_1} \qquad \text{e} \qquad \sigma = \frac{1}{\beta_1} .$$

Regressão Probit no 

Ilustremos a aplicação duma Regressão Probit, no R, aos dados do exemplo DAC, já considerado antes.

Dados Hosmer & Lemeshow

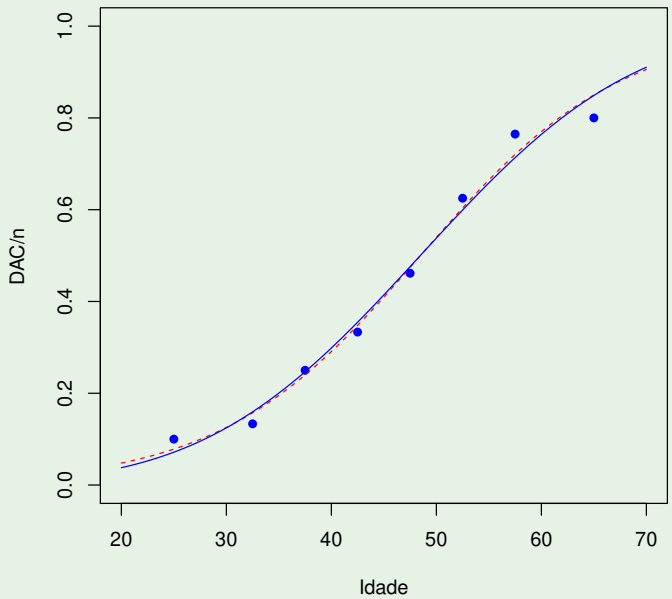
Numa regressão probit, há que especificar a respectiva função de ligação, como opção do argumento `family`, da seguinte forma:

```
> glm(cbind(DAC,nobs-DAC)~idade,family=binomial(link=probit),data=HL.tab)

Call:  glm(formula = cbind(DAC, nobs - DAC) ~ idade,
                family = binomial(link = probit), data = HL.tab)
Coefficients:
(Intercept)         idade
    -3.0245        0.0624
[...]
```

Tal como na Regressão Logística, a variável resposta pode ser uma matriz de duas colunas, indicando o número de "êxitos" e o número de "fracassos" (como acima) ou, alternativamente, como um vector de zeros e uns.

Regressão Probit no R (cont.)

A curva ajustada de probabilidade de DAC sobre idade (x), tem equação: $p(x) = \Phi(-3.0245 + 0.0624\,x)$. Eis a curva, sobreposta à nuvem de pontos (a tracejado tem-se a curva logística ajustada):



A curva foi traçada com o seguinte comando:

```
> curve(pnorm(-3.0245+0.0624*x), add=TRUE, col="blue")
```

O comando `update`

O modelo *probit* agora ajustado apenas difere, em relação ao modelo de regressão logística (`HLtab.glm`), no argumento `family`, onde se especifica a função de ligação.

O comando do `R` `update` é útil nestes casos, pois permite re-ajustar um modelo alterando argumentos.

```
> update(HLtab.glm, family=binomial(link=probit))

Call: glm(formula=cbind(DAC,nobs-DAC)~idade, family=binomial(link=probit), data=HL.tab)
Coefficients:
 (Intercept)        idade
     -3.0245       0.0624
Degrees of Freedom: 7 Total (i.e. Null);  6 Residual
Null Deviance:    28.7
Residual Deviance: 0.6529   AIC: 25.79
```

NOTA: O objecto `HLtab.glm` não foi alterado. E o novo modelo não foi guardado.

A Regressão Probit (cont.)

Para qualquer número de variáveis preditoras, a probabilidade de êxito $p = P[Y = 1]$ é dada, no Modelo Probit, por uma função cujo comportamento é muito semelhante ao do Modelo Logit:

- função estritamente crescente,
- um único ponto de inflexão, quando o preditor linear $\mathbf{x}^t\vec{\boldsymbol{\beta}} = 0$,
- a que corresponde uma probabilidade de êxito $p(0) = 0.5$.
- com simetria em torno do ponto de inflexão, isto é, $p(-\eta) = 1 - p(\eta)$, para qualquer η.

Inconvenientes:

- não há interpretação fácil do significado dos parâmetros β_j;
- a função de ligação é não-canónica.

O modelo log-log do complementar

Modelo log-log do complementar

No mesmo contexto de variável resposta dicotómica Y, outra escolha frequente de função de ligação , com tradição histórica desde 1922 no estudo de organismos infecciosos consiste em tomar para probabilidade de êxito ($Y = 1$):

$$p(\vec{\mathbf{x}}) \quad = \quad g^{-1}\left(\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}\right) \quad = \quad 1 - \mathrm{e}^{-\mathrm{e}^{\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}}}$$

O contradomínio da função agora definida é o intervalo $]0,1[$.

A função de ligação será, neste caso, da forma:

$$\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}} \;=\; g\left(p(\vec{\mathbf{x}})\right) \;=\; \ln\left[-\ln(1-p(\vec{\mathbf{x}}))\right]$$

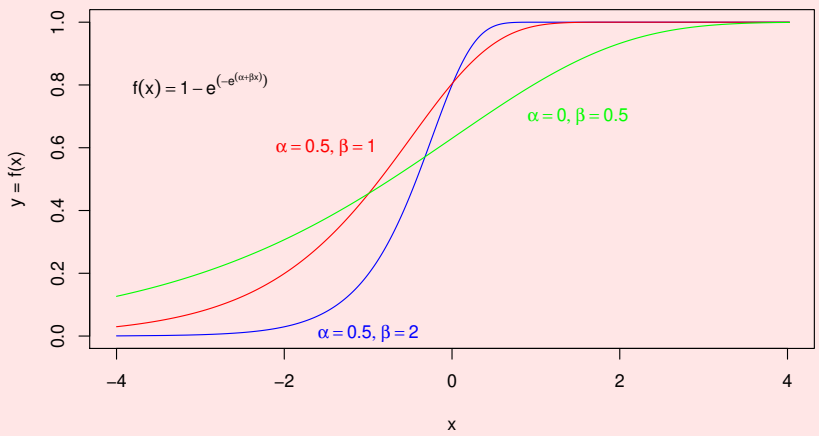
donde a designação log-log do complementar.

O modelo log-log do complementar (cont.)

No caso de um único preditor numérico, a função p é a diferença para 1 duma curva de Gompertz ($y = \alpha e^{-\beta e^{-\gamma x}}$) com valor assintótico $\alpha = 1$. Ou seja,

$$p(x) = 1 - e^{-\beta e^{-\gamma x}} \ .$$

Então, a função $p(x)$ é a função distribuição cumulativa da distribuição de Gumbel:



O modelo log-log do complementar (cont.)

Esta função para p tem analogias e diferenças de comportamento em relação aos Modelos Logit e Probit:

- é igualmente estritamente monótona;
- tem igualmente um único ponto de inflexão, quando $\eta = 0$;
- mas o valor de probabilidade associado já não se encontra a meio caminho na escala de probabilidades, sendo $p(0) = 1 - \frac{1}{e}$;
- isso significa que a "fase inicial" da curva de probabilidades decorre até um valor superior da probabilidade ($1 - \frac{1}{e} \approx 0.632$) do que nas Regressões *Logit* e *Probit*.

Tal como no caso do Modelo Probit, os coeficientes β_j da componente sistemática não têm um significado tão facilmente interpretável como numa Regressão Logística.

Log-log do complementar no 

Ajustar o modelo com função de ligação log-log do complementar faz-se especificando o valor `cloglog` no argumento `link`.

```
> update(HLtab.glm, family=binomial(link=cloglog))


Call: glm(formula=cbind(DAC, nobs-DAC)~idade,
                  family=binomial(link=cloglog), data=HL.tab)

Coefficients:
(Intercept)         idade
   -4.00470       0.07311
[...]
```

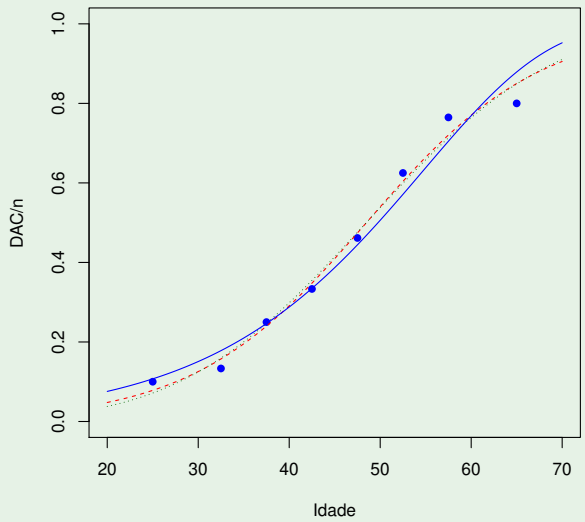
A curva ajustada é:

$$p(x) = 1 - \mathrm{e}^{-\mathrm{e}^{-4.00470+0.07311\,x}}.$$

Log-log do complementar no R (cont.)

A curva ajustada, sobreposta à nuvem de pontos do exemplo DAC, é

$$p(x) = 1 - \mathrm{e}^{-\mathrm{e}^{-4.00470+0.07311\,x}}$$



A tracejado, no gráfico, a curva do modelo `probit` e a ponteado a da regressão logística. A nova curva foi traçada com os seguintes comandos:

```
> cloglog <- function( b0,b1,x ){1-exp(-exp(b0+b1*x))}
> curve(cloglog(b0=-4.0047,b1=0.07311,x), add=TRUE, col="blue")
```

Outras funções de ligação para respostas binárias

Foram consideradas três funções de ligação em modelos de resposta Bernoulli (ou Binomial/n), cujas inversas são sigmóides. Em dois casos, usaram-se inversas de funções de distribuição cumulativas:

- f.d.c. duma Normal reduzida, no Modelo Probit;
- f.d.c. duma Gumbel, no Modelo log-log do Complementar

Uma generalização óbvia consiste em utilizar outra f.d.c. duma variável aleatória contínua, gerando novos MLGs de resposta dicotómica.

No `R`, além das opções acima referidas, pode usar-se uma f.d.c. da distribuição de Cauchy (`link=cauchit`).

Inferência: propriedades dos estimadores MV

Propriedades de estimadores de Máxima Verosimilhança

Quaisquer estimadores $\hat{\vec{\beta}}$ de máxima verosimilhança são:

- assintoticamente multinormais
- assintoticamente centrados ($E[\hat{\vec{\beta}}] \to \vec{\beta}$).
- assintoticamente de matriz de variâncias-covariâncias $\mathbb{I}_{\vec{\beta}}^{-1}$, onde

$$\mathbb{I}_{\vec{\beta}} = -E[\mathscr{H}_{\vec{\beta}}]$$

é a matriz de Informação de Fisher, sendo $\mathscr{H}_{\vec{\beta}}$ a matriz Hessiana da log-verosimilhança $\mathscr{L}$, no ponto $\vec{\beta}$, cujo elemento (j,m) é:

$$\left(\mathscr{H}_{\vec{\beta}}\right)_{(j,m)} = \frac{\partial^2 \mathscr{L}}{\partial \beta_j \, \partial \beta_m}$$

Conclusão: Pode fazer-se inferência (assintótica) em MLGs!

Inferência em MLGs

Distribuição assintótica de $\vec{\hat{\beta}}$ num MLG

Num Modelo Linear Generalizado, o vector de estimadores de Máxima Verosimilhança, $\vec{\hat{\beta}}$, verifica assintoticamente:

$$\vec{\hat{\beta}} \sim \mathcal{N}_{p+1}\left(\vec{\beta}, \mathbb{I}_{\vec{\beta}}^{-1}\right)$$

onde $\mathbb{I}_{\vec{\beta}}$ é a matriz de informação de Fisher da log-verosimilhança da amostra, calculada no ponto $\vec{\beta}$.

A dimensão da amostra tem uma importância grande para garantir a fiabilidade da aproximação assintótica.

Repare-se na semelhança com o resultado distribucional que serve de base à inferência num modelo linear. As mesmas propriedades da Multinormal podem ser usadas para obter resultados análogos.

Inferência em MLGs (cont.)

Distribuição para combinações lineares dos parâmetros

Dado um MLG (admitindo certas condições de regularidade) e um vector não-aleatório $\mathbf{a}_{p+1}$, os estimadores de Máxima Verosimilhança $\vec{\hat{\boldsymbol{\beta}}}$ verificam, assintoticamente:

$$\frac{\vec{\mathbf{a}}^t \vec{\hat{\boldsymbol{\beta}}} - \vec{\mathbf{a}}^t \vec{\boldsymbol{\beta}}}{\sqrt{\vec{\mathbf{a}}^t \mathbb{I}_{\beta}^{-1} \vec{\mathbf{a}}}} \sim \mathcal{N}(0,1) \ .$$

O Teorema permite obter intervalos de confiança e testes de hipóteses (aproximados) para combinações lineares dos parâmetros $\vec{\boldsymbol{\beta}}$.

Inferência em MLGs (cont.)

Na expressão que serve de base aos ICs e Testes de Hipóteses surge a inversa da matriz de informação no ponto desconhecido $\vec{\boldsymbol{\beta}}$. Essa matriz desconhecida é substituída por outra, conhecida: a matriz de informação calculada para a estimativa $\hat{\vec{\boldsymbol{\beta}}}$.

Para distribuições com parâmetro de dispersão ϕ desconhecido, existe ainda o problema (ainda não considerado) da estimação de ϕ.

Tudo isto reforça a necessidade de grandes amostras para que se possa confiar nos resultados.

Inferência em MLGs (cont.)

Intervalos de Confiança (assintóticos)

Um intervalo assintótico a $(1-\alpha)\times 100\%$ de confiança para a combinação linear $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$ é dado por:

$$\left]\ \vec{\mathbf{a}}^t\vec{\mathbf{b}} - z_{\frac{\alpha}{2}}\cdot\sqrt{\vec{\mathbf{a}}^t\mathbb{I}_{\hat{\beta}}^{-1}\vec{\mathbf{a}}}\quad,\quad \vec{\mathbf{a}}^t\vec{\mathbf{b}} + z_{\frac{\alpha}{2}}\cdot\sqrt{\vec{\mathbf{a}}^t\mathbb{I}_{\hat{\beta}}^{-1}\vec{\mathbf{a}}}\ \right[$$

sendo $\mathbb{I}_{\hat{\beta}}^{-1}$ a inversa da matriz de informação de Fisher da log-verosimilhança, calculada no ponto $\vec{\hat{\boldsymbol{\beta}}}$.

Inferência em MLGs (cont.)

Teste de Hipóteses (assintótico)

Num MLG, um teste de hipóteses (assintótico) bilateral a uma combinação linear dos β_j é:

- Hipóteses:

$$H_0 : \vec{\mathbf{a}}^t \vec{\boldsymbol{\beta}} = c \qquad vs. \qquad H_1 : \vec{\mathbf{a}}^t \vec{\boldsymbol{\beta}} \neq c$$

- Estatística do Teste:

$$Z = \frac{\vec{\mathbf{a}}^t \hat{\vec{\boldsymbol{\beta}}} - \vec{\mathbf{a}}^t \vec{\boldsymbol{\beta}}_{|H_0}}{\sqrt{\vec{\mathbf{a}}^t \mathbb{I}^{-1}_{\hat{\vec{\boldsymbol{\beta}}}} \vec{\mathbf{a}}}} \quad \sim \quad \mathcal{N}(0,1) \ ,$$

- Região Crítica: Bilateral. Rejeitar H_0 se $|Z_{calc}| > z_{\frac{\alpha}{2}}$.

Definem-se testes unilaterais, com hipóteses e RCs análogas às do modelo linear.

A função `summary`

A função `summary` tem método para MLGs, gerando resultados análogos aos de modelos lineares.

A tabela `Coefficients` tem colunas análogas:

- `Estimate` – valores estimados dos parâmetros β_j;
- `Std.Error` – os respectivos desvios padrão estimados, $\hat{\sigma}_{\hat{\beta}_j}$, i.e., as raízes quadradas dos elementos diagonais da matriz $\mathscr{I}^{-1}_{\vec{\hat{\boldsymbol{\beta}}}}$;
- `z value` – o valor calculado da estatística $Z = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$, para um teste às hipóteses $H_0 : \beta_j = 0$ *vs.* $H_1 : \beta_j \neq 0$;
- `Pr(>|z|)` – o *p-value* (bilateral) da estatística da coluna anterior (calculado numa $\mathscr{N}(0,1)$).

O teste referido pode servir para determinar a dispensabilidade de algum preditor.

Inferência no R

Na listagem do comando `summary` tem-se a informação fundamental para construir ICs ou Testes a parâmetros, num MLG.

```
> summary(HLtab.glm)

Call:
glm(formula=cbind(DAC, n-DAC) ~ idade, family=binomial, data=HL.tab)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.09073    1.09753  -4.638 3.51e-06 ***
Idade        0.10502    0.02308   4.551 5.35e-06 ***
[...]
Number of Fisher Scoring iterations: 4      <-- no. passos do algoritmo
```

A matriz de covariâncias dos estimadores no R

O comando `vcov` devolve a matriz de (co-)variâncias dos estimadores $\vec{\hat{\boldsymbol{\beta}}}$, ou seja, a inversa da matriz de informação de Fisher, $\mathbb{I}_{\hat{\beta}}^{-1}$.

```
> vcov(HLtab.glm)
             (Intercept)          idade
(Intercept)  1.20457613 -0.0247424726
idade       -0.02474247  0.0005325726
```

Esta é a matriz usada para construir um intervalo assintótico a $(1-\alpha)\times 100\%$ de confiança para a combinação linear $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$:

$$\left]\ \vec{\mathbf{a}}^t\vec{\mathbf{b}} - z_{\frac{\alpha}{2}}\cdot\sqrt{\vec{\mathbf{a}}^t\mathbb{I}_{\hat{\beta}}^{-1}\vec{\mathbf{a}}}\ ,\ \vec{\mathbf{a}}^t\vec{\mathbf{b}} + z_{\frac{\alpha}{2}}\cdot\sqrt{\vec{\mathbf{a}}^t\mathbb{I}_{\hat{\beta}}^{-1}\vec{\mathbf{a}}}\ \right[$$

Intervalos de confiança para β_j no R

Os intervalos de confiança para os parâmetros individuais β_j são dados pela função `confint.default`.

```
> confint.default(HLtab.glm)
                  2.5 %     97.5 %
(Intercept) -7.24185609 -2.9396103
idade        0.05978799  0.1502503
```

Venables & Ripley, no módulo `MASS`, disponibilizam um método alternativo (computacionalmente mais exigente) de construir intervalos de confiança em MLGs, denominado *profiling*. É automaticamente invocado, pela função `confint`:

```
> confint(HLtab.glm)
Waiting for profiling to be done...
                  2.5 %     97.5 %
(Intercept) -7.42548805 -3.0887956
idade        0.06276942  0.1539715
```

MLGs para variáveis resposta de Poisson

Consideremos agora modelos em que a componente aleatória Y tem distribuição de Poisson.

A distribuição de Poisson é frequente na contagem de acontecimentos aleatórios raros (quando se pode admitir que não há acontecimentos simultâneos).

Se Y tem distribuição de Poisson, toma valores em $\mathbb{N}_0$ com probabilidades $P[Y = y] = \frac{\lambda^y}{y!} \mathrm{e}^{-\lambda}$, para $\lambda > 0$.

Esta distribuição não é indicada para situações em que seja fixado à partida o número máximo de observações ou realizações do fenómeno, como sucede com uma Binomial.

A Poisson na família exponencial

Uma variável aleatória discreta tem distribuição de Poisson se tem função de massa probabilística $P[Y = y] = \frac{\lambda^y}{y!} e^{-\lambda}$, para $y \in \mathbb{N}_0$.

Pode re-escrever-se a função de massa probabilística duma Poisson como:

$$f(y|\lambda) = e^{-\lambda} \frac{\lambda^y}{y!} = e^{-\lambda} \cdot e^{\ln(\lambda^y/y!)} = e^{-\lambda + \ln(\lambda^y) - \ln(y!)} = e^{-\lambda + y\ln(\lambda) - \ln(y!)}$$

que é da família exponencial $f(y\,|\,\theta,\phi) = e^{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)}$, com:

- $\theta = \ln(\lambda)$
- $\phi = 1$
- $b(\theta) = e^{\theta} = \lambda$
- $a(\phi) = 1$
- $c(y,\phi) = -\ln(y!)$

Funções de ligação e ligação canónica

O valor esperado de $Y \frown Po(\lambda)$ é o parâmetro $\lambda = E[Y]$.

Uma função de ligação será uma função $g(\cdot)$ tal que:

$$g(E[Y]) = g(\lambda) = \vec{\mathbf{x}}^t \vec{\boldsymbol{\beta}} \ ,$$

onde $\vec{\mathbf{x}}^t \vec{\boldsymbol{\beta}}$ é a componente sistemática do Modelo.

O parâmetro natural da distribuição de Poisson é $\theta = \ln(\lambda)$.

Assim, a função de ligação canónica para uma componente aleatória com distribuição de Poisson é a função de ligação logarítmica:

$$g(\lambda) = \ln(\lambda) = \vec{\mathbf{x}}^t \vec{\boldsymbol{\beta}} \quad \Leftrightarrow \quad \lambda = g^{-1}\left(\vec{\mathbf{x}}^t \vec{\boldsymbol{\beta}}\right) = \mathrm{e}^{\vec{\mathbf{x}}^t \vec{\boldsymbol{\beta}}}$$

Um Modelo assim definido designa-se um Modelo Log-Linear.

Modelos log-lineares

Modelos Log-lineares

São modelos com:

- componente aleatória de Poisson;
- função de ligação logaritmo natural, que é a ligação canónica para as Poisson.

Nestes modelos, o valor esperado da variável resposta Poisson é dado por:

$$\lambda = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p}$$

Nota: a ligação apenas permite valores positivos do parâmetro λ, o que está estruturalmente de acordo com as características do parâmetro λ duma distribuição Poisson.

Interpretação dos parâmetros β_j

No caso de haver uma única variável preditora X, a relação entre o parâmetro $\lambda = E[Y]$ da distribuição Poisson e o preditor fica:

$$\lambda(x) = e^{\beta_0} \cdot e^{\beta_1 x}$$

O aumento de uma unidade no valor do preditor multiplica o valor esperado da variável resposta por e^{β_1}:

$$\lambda(x+1) = e^{\beta_0} \cdot e^{\beta_1 (x+1)} = e^{\beta_0} \cdot e^{\beta_1 x} \cdot e^{\beta_1} = \lambda(x) \cdot e^{\beta_1} \ .$$

A interpretação generaliza-se para mais do que uma variável preditora. Com p variáveis preditoras tem-se:

$$\lambda(x) = e^{\beta_0} \, e^{\beta_1 x_1} \, e^{\beta_2 x_2} \, \ldots \, e^{\beta_p x_p}.$$

Um aumento de uma unidade no valor do preditor X_j, mantendo as restantes variáveis preditoras constantes, multiplica o valor esperado de Y por e^{β_j}.

Factores preditores e tabelas de contingência

No caso de uma variável indicatriz X_j, tem-se que a pertença à categoria assinalada pela indicatriz X_j multiplica o parâmetro λ da distribuição de Poisson por e^{β_j}.

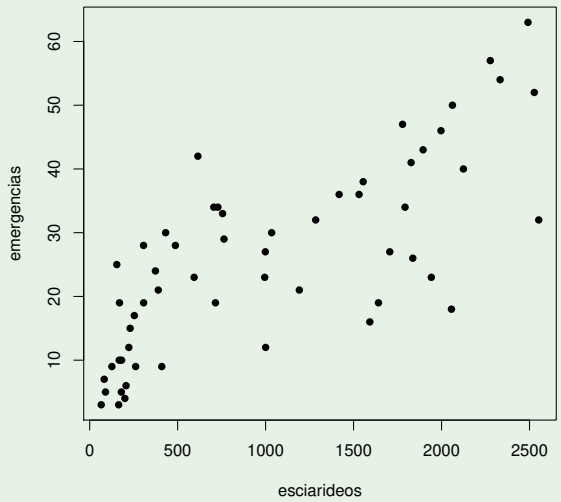
Os modelos log-lineares têm grande importância no estudo de tabelas de contingência, cujos margens correspondem a diferentes factores e cujo recheio corresponde a contagens de observações nos cruzamentos de níveis correspondentes.

Tal como nos casos anteriores, outras funções de ligação são concebíveis para variáveis-resposta com distribuição de Poisson.

Exemplo: Exercício 5 (Modelo Log-linear)

Dados Elisa1 (emergencias de predador)

Quer-se modelar o número de `emergencias` (Y) de adultos dum predador, em função do número de mosquitos (`esciarideos`, x) no subestrato de que se alimentam as larvas do predador. Dados: *data frame* `Elisa1`.



Há crescimento curvilíneo do número médio de emergencias.
Se fôr crescimento exponencial, tem-se $E[Y] = \gamma e^{\beta_1 x} = e^{\beta_0 + \beta_1 x}$.

Exemplo: Exercício 5 (cont.)

Dados Elisa1 (emergencias)

Admitindo Y com distribuição Poisson, $E[Y] = \lambda$. O crescimento exponencial de $E[Y]$ em função de x é um modelo log-linear (canónico para Poisson):

$$\lambda = E[Y] = e^{\beta_0 + \beta_1 x} \quad \Leftrightarrow \quad \ln(\lambda) = \beta_0 + \beta_1 x.$$

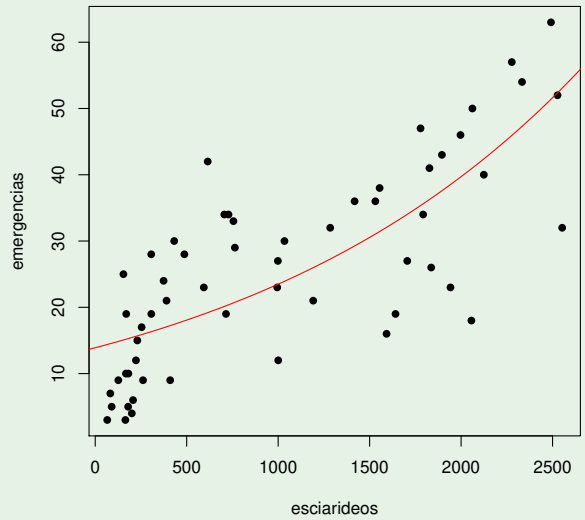
```
> Elisa1.glm <- glm(emergencias~esciarideos,family=poisson,data=Elisa1)

> summary(Elisa1.glm)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.632e+00  5.076e-02   51.85   <2e-16 ***
esciarideos 5.248e-04  3.209e-05   16.36   <2e-16 ***
--
(Dispersion parameter for poisson family taken to be 1)  <---
[...]
```

A curva ajustada é: $\lambda = E[Y] = e^{b_0 + b_1 x} = e^{2.632 + 0.0005248 x}$

Exemplo: Exercício 5 (cont.)

Dados Elisa1 (emergencias de predador)

Eis a curva ajustada $\lambda = E[Y] = e^{b_0 + b_1 x} = e^{2.632 + 0.0005248\, x}$:



Por cada 100 esciarídeos a mais, o número médio de emergências é dado por:

$$\lambda_{x+100} = e^{100\, b1} \times \lambda_x = e^{0.05248} \times \lambda_x = 1.053881 \times \lambda_x \, .$$

Por cada 100 esciarídeos a mais, o número médio de emergências aumenta $\approx 5,4\%$.

Avaliação da qualidade dum MLG

Conceito importante na avaliação da qualidade de um MLG é o conceito de Desvio de um Modelo (*deviance* em inglês).

O desvio desempenha nos GLMs um papel análogo ao da Soma de Quadrados Residual nos Modelos Lineares.

No estudo do Modelo Linear foi introduzida a noção de Modelo Nulo: um Modelo em que o preditor linear é constituído apenas por uma constante e toda a variação nos valores observados é variação residual, não explicada pelo Modelo.

No estudo de Modelos Lineares Generalizados é de utilidade um Modelo que ocupa o extremo oposto na gama de possíveis modelos: o Modelo saturado, que tem tantos parâmetros quantas as observações de Y disponíveis.

Modelo Nulo e modelo saturado (cont.)

Um modelo saturado ocupa o polo oposto em relação ao Modelo Nulo: enquanto que neste último tudo é variação residual, não explicada pelo modelo, num modelo saturado tudo é "explicado" pelo modelo, não havendo lugar a variação residual.

Num modelo saturado, o ajustamento é "perfeito", mas inútil: a estimativa de cada valor esperado de Y coincide totalmente com o valor observado de Y correspondente, isto é, $\hat{\mu}_i = \widehat{E[Y_i]} = Y_i$.

Um tal ajustamento "perfeito" dos dados ao modelo saturado é ilusório. Mas é de utilidade como termo de comparação para medir o grau de ajustamento dum MLG a um conjunto de dados, medindo-se o afastamento em relação a este ajustamento "ideal".

É nessa ideia que se baseia a definição do conceito de Desvio ou *Deviance*.

Desvio

Desvio (Deviance)

Considere-se um Modelo Linear Generalizado baseado em n observações independentes da variável resposta Y. Sejam:

- $\mathscr{L}_M$ a log-verosimilhança correspondente ao vector estimado $\hat{\vec{\boldsymbol{\beta}}}_M$ dos seus parâmetros (máxima com os dados observados);
- $\mathscr{L}_T$ a log-verosimilhança correspondente ao modelo saturado, isto é, a log-verosimilhança obtida substituindo cada valor esperado μ_i pela observação correspondente y_i.

Define-se o desvio como sendo:

$$D^* = -2(\mathscr{L}_M - \mathscr{L}_T)$$

Desvio (cont.)

Para uma distribuição da família exponencial de distribuições, tem-se:

$$\mathscr{L}(\vec{\boldsymbol{\theta}}, \vec{\boldsymbol{\phi}}) = \sum_{i=1}^{n} \left[\frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i) \right]$$

O desvio correspondente, indicando pelas letras M e T os estimadores associados ao parâmetro natural θ, e admitindo conhecido o parâmetro de dispersão ϕ, vem:

$$D^* = -2[\mathscr{L}(\hat{\theta}^M) - \mathscr{L}(\hat{\theta}^T)] = 2 \sum_{i=1}^{n} \left[\frac{y_i(\hat{\theta}_i^T - \hat{\theta}_i^M) - [b(\hat{\theta}_i^T) - b(\hat{\theta}_i^M)]}{a(\phi_i)} \right]$$

Desvio (cont.)

As expressões para os desvios são mais simples caso o parâmetro de dispersão seja uma constante conhecida, que não exige estimação. É o caso das distribuições de Poisson e Bernoulli ou Binomial/n:

- $\phi = 1$ na Poisson;
- $\phi = 1$ na Bernoulli ou Binomial/n.

Para distribuições bi-paramétricas da família exponencial em que o parâmetro ϕ não é conhecido, ϕ tem de ser estimado a partir dos dados para se poder calcular o desvio. Isso complica a situação.

Desvio numa Poisson

Vimos (slide 86) que se Y tem distribuição de Poisson:

$$\theta = \ln(\lambda) \quad ; \quad b(\theta) = e^{\theta} = \lambda \quad ; \quad \phi = a(\phi) = 1 \ .$$

No modelo saturado tem-se: $\hat{\lambda}_i^T = y_i$. No modelo ajustado, fica $\hat{\lambda}_i^M = \hat{\lambda}_i$.

Então a expressão geral do desvio (slide 97) será:

$$\begin{aligned} D^* &= -2[\mathscr{L}(\hat{\theta}^M) - \mathscr{L}(\hat{\theta}^T)] = 2\sum_{i=1}^{n}\left[y_i(\hat{\theta}_i^T - \hat{\theta}_i^M) - [b(\hat{\theta}_i^T) - b(\hat{\theta}_i^M)]\right] \\ &= 2\sum_{i=1}^{n}\left\{y_i\left[\ln(y_i) - \ln(\hat{\lambda}_i)\right] - (y_i - \hat{\lambda}_i)\right\} \\ \Leftrightarrow \quad D^* &= 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i)\right] \end{aligned}$$

Nota: A expressão para $\hat{\lambda}$ (logo, do Desvio) depende também da função de ligação usada.

Exemplo: Exercício 5 (cont.)

Dados Elisa1 (emergencias)

```
> Elisa1.glm <- glm(emergencias~esciarideos,family=poisson,data=Elisa1)
> summary(Elisa1.glm)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.632e+00  5.076e-02   51.85   <2e-16 ***
esciarideos 5.248e-04  3.209e-05   16.36   <2e-16 ***
--
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 513.00  on 56  degrees of freedom   <---
Residual deviance: 244.22  on 55  degrees of freedom   <---
AIC: 526.32
```

O Desvio do modelo é a `Residual Deviance` (244.22, no nosso caso).

O Desvio do Modelo Nulo é a `Null Deviance` (513.00, no nosso caso).

Comparação de modelos: Razão de verosimilhanças

Um teste à admissibilidade dum Submodelo é obtido com um resultado muito geral: o teste à razão de Verosimilhanças (Likelihood Ratio test).

Seja $(Y_1, Y_2, ..., Y_n)$ uma amostra aleatória. Seja $L(\vec{\boldsymbol{\theta}}|\vec{\mathbf{Y}})$ a sua verosimilhança, onde $\vec{\boldsymbol{\theta}}$ designa um vector de parâmetros. Sejam Θ_0 e Θ_1 dois conjuntos alternativos de condições sobre os valores dos parâmetros $\boldsymbol{\theta}$.

No contexto dum Modelo Linear Generalizado com ϕ conhecido, os parâmetros $\boldsymbol{\theta}$ são os $p+1$ coeficientes β_j da combinação linear que constitui a componente sistemática do Modelo.

Sejam:

- Θ_0 valores admissíveis com a restrição do submodelo: $H_0 : \vec{\boldsymbol{\beta}}_{\overline{S}} = \vec{\mathbf{0}}$.
- Θ_1 indica-se a condição complementar: $H_1 : \vec{\boldsymbol{\beta}}_{\overline{S}} \neq \vec{\mathbf{0}}$.
- $\Theta_0 \cup \Theta_1$ indica-se qualquer vector $\vec{\boldsymbol{\beta}}_{\overline{S}}$, sem restrições.

Teorema de Wilks

Designa-se razão de verosimilhanças (likelihood ratio) a:

$$R_n(\mathbf{x}) = \frac{\max\limits_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta}|\mathbf{x})}{\max\limits_{\boldsymbol{\theta} \in (\Theta_0 \cup \Theta_1)} L(\boldsymbol{\theta}|\mathbf{x})}$$

O Teorema de Wilks garante que, sob H_0 (e com certas condições de regularidade da função de verosimilhança) $\boldsymbol{\Lambda} = -2\ln(R_n)$ tem distribuição assintótica χ^2_q, onde q indica o número de restrições impostas aos parâmetros em H_0:

$$\boldsymbol{\Lambda} \quad = \quad -2\left(\max_{\boldsymbol{\theta} \in \Theta_0} \mathscr{L}(\boldsymbol{\theta};\mathbf{x}) - \max_{\boldsymbol{\theta} \in (\Theta_0 \cup \Theta_1)} \mathscr{L}(\boldsymbol{\theta};\mathbf{x})\right) \quad \sim \quad \chi^2_q \, .$$

Assim, $\boldsymbol{\Lambda}$ pode ser utilizada como estatística do teste às hipóteses:

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \qquad \text{vs.} \qquad H_1 : \boldsymbol{\theta} \in \Theta_1 \, .$$

com região crítica unilateral direita.

Teste de Wilks a Submodelos

No contexto de comparação de modelos e submodelos num MLG,

- q é a diferença entre o número de parâmetros do modelo completo ($\Theta_0 \cup \Theta_1$) e do submodelo (Θ_0): $q = p - k$;
- o Desvio do modelo completo, $D_M^* = -2(\mathscr{L}(\hat{\theta}^M) - \mathscr{L}(\hat{\theta}^T))$ é calculado com base na log-verosimilhança $\mathscr{L}(\hat{\theta}^M)$ correspondente às estimativas MV do Modelo Completo;
- o Desvio do submodelo, $D_S^* = -2(\mathscr{L}(\hat{\theta}^S) - \mathscr{L}(\hat{\theta}^T))$ é calculado com base na log-verosimilhança $\mathscr{L}(\hat{\theta}^S)$ correspondente às estimativas MV do Submodelo;
- A log-verosimilhança $\mathscr{L}(\hat{\theta}^T))$ do modelo saturado é igual nos dois casos (os valores esperados de Y são sempre estimados pelos valores observados);
- a estatística do teste é apenas a diferença dos desvios:

$$\Lambda = D_S^* - D_M^*$$

Teste de Wilks a Submodelos

A estatística do Teste de Wilks a modelos encaixados é a diferença dos Desvios de Modelo e Submodelo.

Teste de Wilk a Submodelos Encaixados

Hipóteses:

$$\begin{array}{lll} & H_0 : \beta_j = 0, \quad \forall j \notin S & \textit{vs.} \quad H_1 : \exists j \notin S, \text{ t.q. } \beta_j \neq 0 \\ \Leftrightarrow & H_0 : \vec{\boldsymbol{\beta}}_{\overline{S}} = \vec{\mathbf{0}} & \textit{vs.} \quad H_1 : \vec{\boldsymbol{\beta}}_{\overline{S}} \neq \vec{\mathbf{0}} \\ & [\text{ Submodelo OK}] & \textit{vs.} \quad [\text{ Modelo melhor}] \end{array}$$

Estatística do Teste: $\boldsymbol{\Lambda} = D^*_S - D^*_M \sim \chi^2_{p-k}$,

Região Crítica: Unilateral direita Rejeitar H_0 se $\boldsymbol{\Lambda}_{calc} > \chi^2_{\alpha;(p-k)}$.

Nota: No caso do parâmetro de dispersão ϕ não ser conhecido, o cálculo de D^* (que envolve ϕ) fica condicionado. São necessários testes alternativos, ou trabalhar apenas com resultados aproximados, usando uma estimativa de ϕ. O problema não existe para respostas Binomiais ou Poisson.

Teste de Wilks ao Ajustamento Global

Para MLGs cuja componente sistemática inclui uma parcela aditiva constante, o conceito de ajustamento global do Modelo pode ser semelhante ao usado no estudo do Modelo Linear: compare-se o ajustamento do Modelo e do Submodelo Nulo, que se obtém sem qualquer variável preditora (apenas com a constante).

No Submodelo Nulo tem-se:

$$g(E[Y_i]) = \beta_0 \qquad \Longleftrightarrow \qquad E[Y_i] = g^{-1}(\beta_0), \qquad \forall i = 1 : n.$$

Ou seja, $E[Y]$ é constante.

Se o Modelo sob estudo não se ajustar de forma significativamente melhor que esse Submodelo Nulo, conclui-se pela inutilidade do Modelo.

Teste de Wilks ao Ajustamento Global (cont.)

Para modelos em que não seja necessário estimar o parâmetro de dispersão ϕ, tem-se:

Teste de Wilk ao Ajustamento de um MLG

Hipóteses:

$$H_0 : \beta_j = 0, \quad \forall j = 1 : p \quad vs. \quad H_1 : \exists j = 1 : p, \text{ t.q. } \beta_j \neq 0$$
$$[\text{Modelo inutil}] \quad vs. \quad [\text{Melhor que Modelo Nulo}]$$

Estatística do Teste: $\Lambda = D_N^* - D_M^* \sim \chi^2_p$,

Região Crítica: Unilateral direito. Rejeitar H_0 se $\Lambda_{calc} > \chi^2_{\alpha;p}$.

D_N^* indica o Desvio do Modelo Nulo.

Exemplo: Exercício 5 (cont.)

Dados Elisa1 (emergencias)

Para testar a significância do ganho no desvio (face ao Modelo Nulo), recorre-se ao teste de Wilks:

```
> anova(Elisa1.glm, test="Chisq")

Analysis of Deviance Table
Model: poisson, link: log
Response: emergencias

           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                         56     513.00
esciarideos  1   268.78        55     244.22 < 2.2e-16 ***
```

O modelo ajusta-se significativamente melhor do que o Modelo Nulo.

Desvio na Bernoulli

Vimos (slide 16) que se Y tem distribuição Bernoulli:

$$\theta = \ln\left(\frac{p}{1-p}\right) \quad ; \quad b(\theta) = \ln(1+e^{\theta}) = -\ln(1-p) \quad ; \quad \phi = a(\phi) = 1 \ .$$

No modelo saturado tem-se: $\hat{p}_i^T = y_i$. No modelo ajustado, fica $\hat{p}_i^M = \hat{p}_i$.

Substituindo a expressão geral do Desvio (acetato 97), fica:

$$\begin{aligned} D^* &= -2[\mathscr{L}(\hat{\theta}^M) - \mathscr{L}(\hat{\theta}^T)] = 2\sum_{i=1}^{n}\left[y_i(\hat{\theta}_i^T - \hat{\theta}_i^M) - [b(\hat{\theta}_i^T) - b(\hat{\theta}_i^M)]\right] \\ &= 2\sum_{i=1}^{n}\left\{y_i\left[\ln\left(\frac{y_i}{1-y_i}\right) - \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)\right] + \left[\ln(1-y_i) - \ln(1-\hat{p}_i)\right]\right\} \\ \Leftrightarrow \quad D^* &= 2\sum_{i=1}^{n}\left\{y_i \ln\left(\frac{y_i}{\hat{p}_i}\right) + (1-y_i)\ln\left(\frac{1-y_i}{1-\hat{p}_i}\right)\right\} \end{aligned}$$

Desvio na Binomial/n

Vimos (slide 30) que se Y tem distribuição 'Binomial/n':

$$\theta = \ln\left(\frac{p}{1-p}\right) \quad ; \quad b(\theta) = -\ln(1-p) \quad ; \quad \phi = 1 \quad ; \quad a(\phi) = \frac{1}{n} \, .$$

No modelo saturado tem-se: $\hat{p}_i^T = y_i$. No modelo ajustado, fica $\hat{p}_i^M = \hat{p}_i$.

Substituindo a expressão geral do Desvio (acetato 97), fica:

$$D^* = -2[\mathscr{L}(\hat{\theta}^M) - \mathscr{L}(\hat{\theta}^T)] = 2\sum_{i=1}^{n} \frac{\left[y_i(\hat{\theta}_i^T - \hat{\theta}_i^M) - [b(\hat{\theta}_i^T) - b(\hat{\theta}_i^M)]\right]}{a(\phi_i)}$$

$$\Leftrightarrow \quad D^* = 2\sum_{i=1}^{n} n_i \left\{ y_i \ln\left(\frac{y_i}{\hat{p}_i}\right) + (1-y_i)\ln\left(\frac{1-y_i}{1-\hat{p}_i}\right) \right\}$$

A expressão de $\hat{p}$ (logo, do Desvio) depende também da função de ligação usada.

A expressão do Desvio difere na Bernoulli e Binomial/n.

Exemplo: Exercícios 1 e 10

Consideremos MLGs que misturam preditores numéricos e factores.

Exemplo: larva do tabaco (Venables & Ripley)

Um estudo da resistência da larva do tabaco (tobacco budworm) *heliothis virescens* a doses duma substância tóxica.

Lotes de 20 traças de cada sexo foram expostas a doses da referida substância (em μg). Ao fim de 3 dias registou-se o número de individuos mortos em cada lote. Os resultados estão na seguinte tabela.

	Dose					
Sexo	1	2	4	8	16	32
Machos	1	4	9	13	18	20
Fêmeas	0	2	6	10	12	16

Trata-se de dados com variável resposta Binomial (número de mortes em $n = 20 \times 12 = 240$ larvas expostas ao tóxico).

Um exemplo de MLG (cont.)

Exemplo: larva do tabaco

Criação de *data frame* `tabaco` com os dados:

```
> morte <- c(1,4,9,13,18,20,0,2,6,10,12,16)
> sexo <- factor(rep(c("macho","femea"),c(6,6)))
> dose <- rep(2^(0:5),2)
> tabaco <- data.frame(morte,sexo,dose)
> tabaco
```

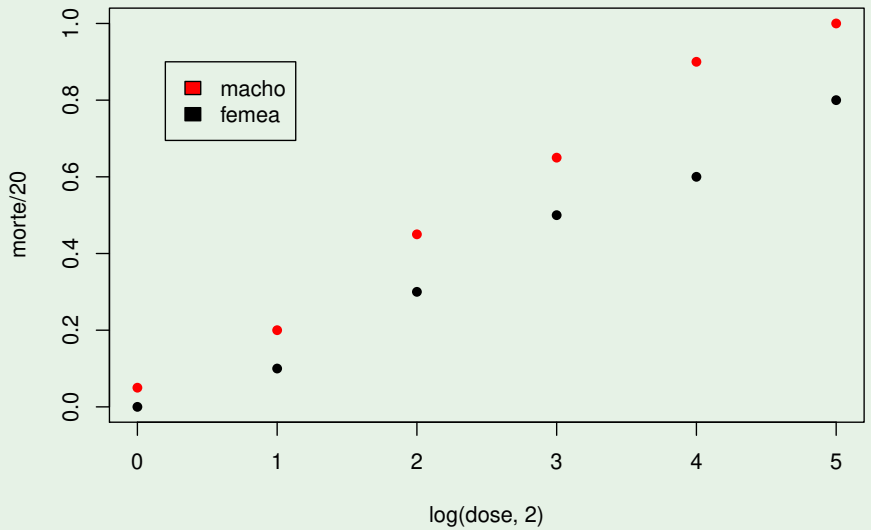
```
   morte  sexo dose
1      1 macho    1
2      4 macho    2
3      9 macho    4
4     13 macho    8
5     18 macho   16
6     20 macho   32
7      0 femea    1
8      2 femea    2
9      6 femea    4
10    10 femea    8
11    12 femea   16
12    16 femea   32
```

É usual em toxicologia usar uma transformação logaritmica (na base 2) de doses que vão sendo duplicadas.

Exercício 1

Exemplo: larva do tabaco

```
> plot(morte/20 ~ log(dose,2),data=tabaco,col=as.numeric(sexo),pch=16)
> legend(0.2,0.9,legend=c("macho","femea"), fill=c("red","black"))
```



Embora uma relação linear pareça adequada, uma relação sigmóide é estruturalmente mais adequada, por apenas tomar valores em $]0,1[$.

Exercício 1 no R (cont.)

Para ajustar uma Regressão Probit, utiliza-se a opção `link=probit` na definição do argumento `family`.

Exemplo: larva do tabaco

```
> glm(cbind(morte,20-morte) ~ log(dose,2),
+       family=binomial(link=probit), data=tabaco)

Call:  glm(formula = cbind(morte, 20 - morte) ~ log(dose, 2),
+         family = binomial(link = probit), data = tabaco)

Coefficients:
 (Intercept)  log(dose, 2)
     -1.6431        0.5966

Degrees of Freedom: 11 Total (i.e. Null);  10 Residual
Null Deviance:      124.9
Residual Deviance: 16.41   AIC: 50.52
```

A relação estimada é: $p(x) = \Phi(-1.6431 + 0.5966 \log_2(x))$, sendo x a dose.

O Desvio do modelo é indicado por `Residual Deviance` (16.41, no nosso caso).

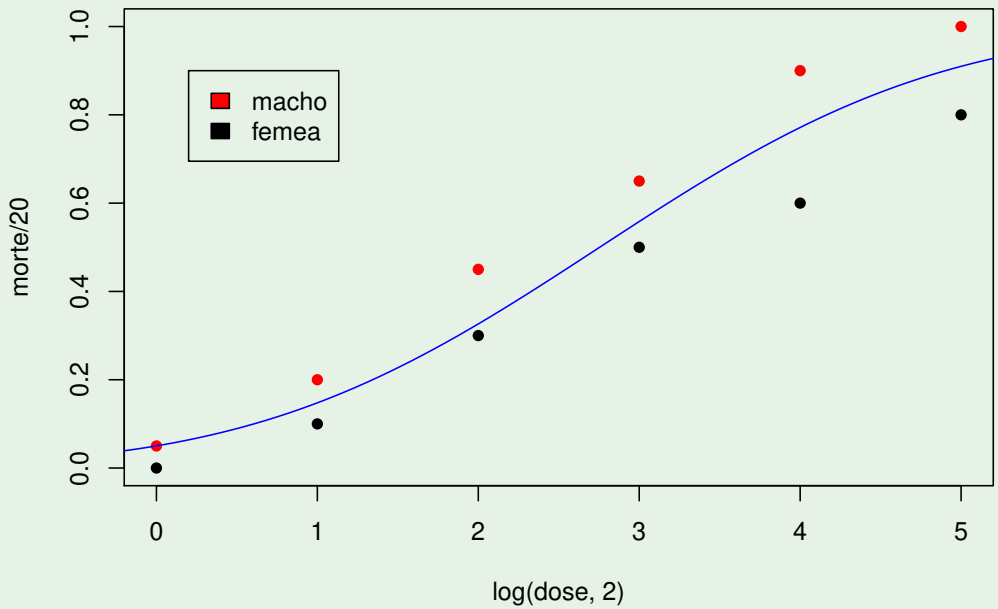
O Desvio do Modelo Nulo é indicado por `Null Deviance` (124.9, no nosso caso).

Exercício 1 no R (cont.)

Exemplo: larva do tabaco

Sobrepõe-se a curva ajustada à nuvem de pontos, com o comando:

```
> curve(pnorm(-1.6431+0.5966*x), from=-1, to=6, col="blue", add=TRUE)
```



Interpretação toxicológica

Larva do tabaco

Em logaritmos de base 2, as dosagens ensaiadas foram $0, 1, 2, 3, 4, 5$.

No contexto toxicológico, podemos afirmar (slide 64) que a tolerância face à dosagem segue uma distribuição $T \frown \mathcal{N}\left(\mu = -\frac{\beta_0}{\beta_1}, \sigma^2 = \frac{1}{\beta_1^2}\right)$.

Com os valores estimados, temos $\hat{\mu} = \frac{1.6431}{0.5966} = 2.754107$ e $\hat{\sigma}^2 = \frac{1}{0.5966^2} = 2.8096$.

Logo, a tolerância à dosagem (em logaritmos de base 2) tem distribuição:

$$T \frown \mathcal{N}(\, 2.7541\,, \underbrace{2.8096}_{=\hat{\sigma}^2}\,)\,.$$

Exercício 1: teste de ajustamento global no R

No R, um teste de Wilks comparando um modelo GLM com o modelo nulo correspondente, pode ser feito utilizando o comando `anova`, com o argumento `test="Chisq"`.

Exemplo: larva do tabaco

```
> tabaco.glm <- glm(cbind(morte,20-morte) ~ log(dose,2),
+                   family=binomial(link=probit), data=tabaco)
> anova(tabaco.glm, test="Chisq")

Analysis of Deviance Table
Model: binomial, link: probit
Response: cbind(morte, 20 - morte)
Terms added sequentially (first to last)
             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                            11    124.876
log(dose, 2)  1   108.46         10     16.414 < 2.2e-16 ***
```

Como previsível, o modelo ajusta-se significativamente melhor do que um modelo nulo, que estima uma probabilidade constante p, para qualquer dose.

Exercício 10

Exemplo: larva do tabaco - modelo tipo ANCOVA

Também é possível conjugar preditores numéricos e factores, tipo ANCOVA.

```
> tabaco.glmSx <- glm(cbind(morte,20-morte) ~ log(dose,2) * sexo ,
+          family=binomial(link=probit), data=tabaco)
> summary(tabaco.glmSx)

(...)
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.80072    0.29832  -6.036 1.58e-09 ***
log(dose, 2)           0.54523    0.09138   5.966 2.43e-09 ***
sexomacho              0.15479    0.41635   0.372    0.710
log(dose, 2):sexomacho 0.19165    0.14259   1.344    0.179
(...)
    Null deviance: 124.876  on 11  degrees of freedom
Residual deviance:   3.768  on  8  degrees of freedom  <-- o desvio baixou de 16.41 para 3.768
AIC: 41.878
```

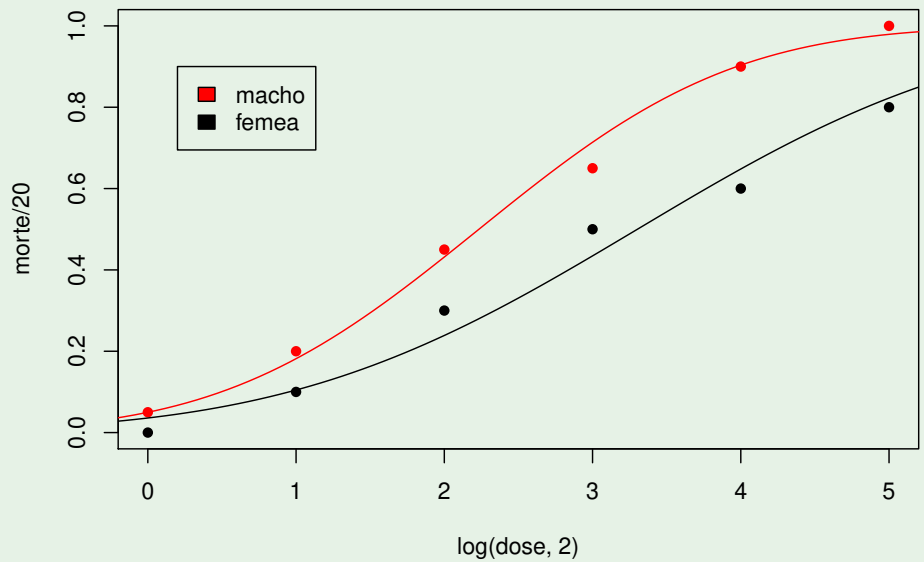
O nível de referência é fêmeas (ordem alfabética). As relações estimadas são:

- $p(x) = \Phi(-1.80072 + 0.54532 \log_2(x))$ nas fêmeas; e
- $p(x) = \Phi((-1.80072 + 0.15479) + (0.54532 + 0.19165) \log_2(x))$ nos machos.

Exercício 10 no R (cont.)

Exemplo: larva do tabaco

```
> plot(morte/20 ~ log(dose,2), col=sexo, data=tabaco, pch=16)
> curve(pnorm(-1.80072+0.54523*x), from=-1, to=6, add=TRUE)
> curve(pnorm((-1.80072+0.15479)+(0.54523+0.19165)*x), from=-1, to=6,
+               col="red", add=TRUE)
> legend(0.2,0.9,legend=c("macho","femea"), fill=c("red","black"))
```

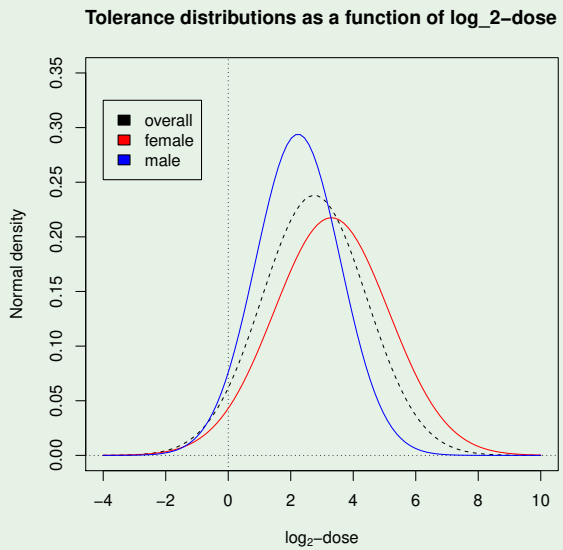


Interpretação toxicológica

Larva do tabaco

A distribuição da tolerância face à dosagem (slide 64), em logaritmos de base 2, pode agora ser calculada em separado para fêmeas e machos:

$$T_F \frown \mathcal{N}(3.30268, 3.36388) \quad ; \quad T_M \frown \mathcal{N}(2.233647, 1.84164).$$



A log-dosagem que mata metade das fêmeas mata quase 80% dos machos:

```
> pnorm(3.30268 , m=2.233647 , sd=sqrt(1.84164))
[1] 0.7845787
```

Exercício 10: teste de Wilks no 

Para saber se há vantagem em considerar modelos diferentes para cada sexo, comparam-se os modelos, como pedido na alínea 10b), usando o teste de Wilks.

Exemplo: larva do tabaco

```
> anova(tabaco.glm, tabaco.glmSx, test="Chisq")

Analysis of Deviance Table
Model 1: cbind(morte, 20 - morte) ~ log(dose, 2)
Model 2: cbind(morte, 20 - morte) ~ log(dose, 2) * sexo
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        10     16.414
2         8      3.768  2   12.646  0.001795 **
```

O modelo com distinção por sexo é significativamente melhor (previsível pelo ajustamento gráfico).

O AIC em GLMs

Critério de Informação de Akaike (AIC)

Define-se, num MLG com m parâmetros, estimados por $\vec{\hat{\boldsymbol{\beta}}}$, como:

$$AIC = -2 \cdot \mathscr{L}(\vec{\hat{\boldsymbol{\beta}}}; \vec{\mathbf{Y}}) + 2m \, .$$

sendo $\mathscr{L}(\vec{\hat{\boldsymbol{\beta}}}; \vec{\mathbf{Y}})$ a log-verosimilhança da amostra.

- Quanto menor o valor do AIC (para igual variável resposta $\vec{\mathbf{Y}}$), melhor o ajustamento do modelo.
- O AIC pode ser usado como critério de comparação de modelos com a mesma amostra e componente aleatória.
- Note-se que num GLM, quer o desvio $D^* = -2(\mathscr{L}_M - \mathscr{L}_T)$, quer o AIC, são definidos à custa da log-verosimilhança.

Selecção de Submodelos

Tal como no Modelo Linear, a escolha dum submodelo adequado pode ser determinado por considerações de diversa ordem.

Caso não haja um submodelo proposto, a pesquisa completa dos $2^p - 2$ possíveis submodelos coloca as mesmas dificuldades computacionais já consideradas no estudo do Modelo Linear.

A função `eleaps` do módulo `R subselect` permite efectuar pesquisas completas para submodelos MLG óptimos duma dada cardinalidade (desde que o número de preditores não seja muito grande).

Alternativamente, é possível usar algoritmos de exclusão ou inclusão sequenciais, semelhantes aos usados no Modelo Linear, mas adoptando como critério para a inclusão/exclusão de variáveis a maior/menor redução (significativa) que geram no Desvio ou AIC.

Algoritmos de selecção de preditores no R

Exclusão sequencial no R: Exercício 10

```
> step(tabaco.glmSx)
Start:  AIC=41.88
cbind(morte, 20 - morte) ~ log(dose, 2) * sexo
                    Df Deviance    AIC
- log(dose, 2):sexo  1    5.566 41.676       <- Repare-se na hierarquia de tipos de efeitos. Num
<none>               3.768 41.878        primeiro passo só considera efeitos de interacção.


Step:  AIC=41.68
cbind(morte, 20 - morte) ~ log(dose, 2) + sexo
               Df Deviance     AIC
<none>               5.566  41.676     <- Tendo excluído a interacção, vai avaliar outros efeitos.
- sexo          1   16.414  50.524
- log(dose, 2)  1  118.799 152.909


Call:  glm(formula = cbind(morte, 20 - morte) ~ log(dose, 2) + sexo,
    family = binomial(link = "probit"), data = tabaco)
Coefficients:
 (Intercept)  log(dose, 2)     sexomacho
     -2.0603        0.6324        0.6536
Degrees of Freedom: 11 Total (i.e. Null);  9 Residual
Null Deviance:      124.9
Residual Deviance: 5.566  AIC: 41.68
```

Opção final: modelo com β_1 igual nos dois sexos, mas β_0 diferente.

Modelos com variável resposta Gama

A distribuição Gama

A Gama é uma distribuição de variáveis aleatórias continuas tomando valores em $\mathbb{R}^+$. Uma parametrização usual em MLGs (ver Exercício 7) é:

$$f(y \mid \mu, \nu) = \frac{\nu^\nu}{\mu^\nu \Gamma(\nu)} y^{\nu-1} \, \mathrm{e}^{-\frac{\nu y}{\mu}}$$

Casos particulares: a distribuição Qui-quadrado (χ^2_n se $\nu = \frac{n}{2}$ e $\mu = n$) e a distribuição Exponencial ($\nu = 1$).

Se $Y \frown G(\mu, \nu)$, a variância é proporcional ao quadrado da média:

$$E[Y] = \mu \qquad \mathrm{e} \qquad V[Y] = \frac{\mu^2}{\nu}$$

MLGs com componente aleatória Gama podem ser úteis em situações onde a variância dos dados não seja constante, mas proporcional ao quadrado da média.

A distribuição Gama na família exponencial

Uma variável aleatória Y tem distribuição Gama com parâmetros μ e ν se toma valores em $\mathbb{R}^+$, com função densidade da forma:

$$f(y \mid \mu, \nu) = \frac{\nu^\nu}{\mu^\nu \Gamma(\nu)} y^{\nu-1} \, \mathrm{e}^{-\frac{\nu y}{\mu}} = \mathrm{e}^{\frac{\left(-\frac{1}{\mu}\right) y + \ln\left(\frac{1}{\mu}\right)}{\frac{1}{\nu}} + \nu \ln \nu - \ln \Gamma(\nu) + (\nu - 1) \ln y}$$

A Gama é da familia exponencial, $f(y|\theta, \phi) = \mathrm{e}^{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)}$, com:

- $\theta = -\frac{1}{\mu}$
- $\phi = \frac{1}{\nu}$
- $b(\theta) = -\ln\left(\frac{1}{\mu}\right) = -\ln(-\theta)$
- $a(\phi) = \phi = \frac{1}{\nu}$
- $c(y, \phi) = \nu \ln \nu - \ln \Gamma(\nu) + (\nu - 1) \ln y$

Funções de ligação e ligação canónica na Gama

Uma vez que para $Y \frown G(\mu, \nu)$ se verifica $E[Y] = \mu$, as funções de ligação g num MLG com variável resposta Gama relacionam a média μ com as combinações lineares das variáveis preditoras:

$$g(\mu) = \vec{\mathbf{x}}^t \vec{\boldsymbol{\beta}} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

A função de ligação canónica para modelos com distribuição Gama transforma o valor esperado de Y no parâmetro natural $\theta = -\frac{1}{\mu}$.

Como o sinal negativo apenas afecta o sinal dos β_js, é hábito definir a função de ligação canónica para modelos com variável resposta Gama apenas como a função recíproco:

$$g(\mu) = \frac{1}{\mu}$$

Modelo Gama com ligação canónica

O modelo fica completo equacionando a parte sistemática a esta transformação canónica do valor esperado de Y:

$$g(\mu) = \frac{1}{\mu} = \vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}} \quad \Leftrightarrow \quad \mu_{\vec{\mathbf{x}}} = g^{-1}\left(\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}\right) = \frac{1}{\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}}$$

Nota: embora o valor esperado da variável resposta Y tenha de ser positivo (uma variável Y com distribuição Gama só toma valores positivos), na relação acima o valor esperado μ pode ser negativo.

Assim, e ao contrário de modelos anteriores, não existe uma "garantia estrutural" de que os valores de μ estimados façam sentido.

Curvas conhecidas em contexto MLG

No caso particular de haver uma única variável preditora numérica, a relação obtida diz que o valor médio de Y é dado por uma curva de tipo hiperbólico,

$$E[Y] = \frac{1}{\beta_0 + \beta_1 x} \ .$$

Esta função tem sido usada em Agronomia para modelar curvas de rendimento por planta (Y), em função da densidade da cultura (X).

Caso se opte por trabalhar com os recíprocos dum único preditor, ou seja com a transformação $X^* = \frac{1}{X}$, o valor esperado fica

$$E[Y] = \frac{1}{\beta_0 + \beta_1 / x} \ = \ \frac{x}{x\beta_0 + \beta_1} \ ,$$

pelo que o valor esperado de Y será dado pela curva de Michaelis-Menten (com a parametrização de Shinozaki-Kira).

Quadro-resumo da família exponencial

Distribuição	$E[Y]$	$V[Y]$	θ	$b(\theta)$	ϕ	$a(\phi)$
Normal	μ	σ^2	μ	$\frac{\theta^2}{2} = \frac{\mu^2}{2}$	σ^2	σ^2
Poisson	λ	λ	$ln(\lambda)$	$e^{\theta} = \lambda$	1	1
Bernoulli	p	$p(1-p)$	$\ln\left(\frac{p}{1-p}\right)$	$ln(1+e^{\theta}) = -ln(1-p)$	1	1
Binomial/n	p	$\frac{p(1-p)}{n}$	$\ln\left(\frac{p}{1-p}\right)$	$e^{\theta} = \lambda$	1	$\frac{1}{n}$(*)
Gama	μ	$\frac{\mu^2}{\nu}$	$-\frac{1}{\mu}$	$-ln(-\theta) = -ln\left(\frac{1}{\mu}\right)$	$\frac{1}{\nu}$	$\frac{1}{\nu}$

(*) Tirando este caso, tem-se sempre $a(\phi) = \phi$.

O parâmetro de dispersão ϕ desconhecido

Em GLMs com variável resposta Poisson ou Bernoulli/Binomial, o parâmetro de dispersão é conhecido: $\phi = 1$.

Mas em GLMs com componente aleatória de distribuiçao Gama, ou Normal, o parâmetro de dispersão é, em geral, desconhecido:

- Numa Normal, $\phi = \sigma^2$ (a variância);
- Numa Gama com a parametrização $Y \frown G(\mu, \nu)$, $\phi = \frac{1}{\nu}$ (com variância $V[Y] = \mu^2 \phi$).

O desconhecimento de ϕ exige estimação e cria problemas.

É frequente admitir que ϕ é comum a todas as observações, ou que varia entre observações apenas devido a constantes conhecidas. Admitir ϕ_i's arbitrariamente diferentes impede a sua estimação.

Desvio e desvio reduzido

Sendo necessário estimar ϕ, define-se o desvio reduzido (scaled deviance).

Desvio e desvio reduzido

Admitindo que $a(\phi_i) = \frac{\phi}{w_i}$, para ϕ comum a todas as observações e pesos w_i conhecidos, o desvio fica:

$$D^* = -2[\mathscr{L}(\hat{\theta}^M) - \mathscr{L}(\hat{\theta}^T)] = 2\sum_{i=1}^{n} \frac{w_i}{\phi}\left\{y_i(\hat{\theta}_i^T - \hat{\theta}_i^M) - [b(\hat{\theta}_i^T) - b(\hat{\theta}_i^M)]\right\}$$

É usual chamar a D^* o desvio reduzido (scaled deviance) e reservar a expressão desvio (deviance) para D, definido tal que:

$$D^* = \frac{D}{\phi}, \quad \Leftrightarrow \quad D = 2\sum_{i=1}^{n} w_i\left\{y_i(\hat{\theta}_i^T - \hat{\theta}_i^M) - [b(\hat{\theta}_i^T) - b(\hat{\theta}_i^M)]\right\}$$

NOTA: Na Poisson e Bernoulli-Binomial/n, desvio e desvio reduzido coincidem, pois $\phi = 1$.

Desvio e desvio reduzido na Normal

Vimos (slide 15) que se Y tem distribuição Normal:

$$\theta = \mu \quad ; \quad b(\theta) = \frac{\theta^2}{2} = \frac{\mu^2}{2} \quad ; \quad \phi = \sigma^2 \quad ; \quad a(\phi) = \phi = \sigma^2 \ .$$

No modelo saturado tem-se: $\hat{\mu}_i^T = y_i$. No modelo ajustado, fica $\hat{\mu}_i^M = \hat{\mu}_i$.

Substituindo na expressão geral do Desvio (slide 97):

$$\begin{aligned} D^* &= 2 \sum_{i=1}^{n} \frac{\left[y_i(\hat{\theta}_i^T - \hat{\theta}_i^M) - [b(\hat{\theta}_i^T) - b(\hat{\theta}_i^M)] \right]}{a(\phi_i)} = 2 \sum_{i=1}^{n} \frac{\left[y_i(y_i - \hat{\mu}_i) - \frac{y_i^2}{2} + \frac{\hat{\mu}_i^2}{2} \right]}{\sigma_i^2} \\ &= 2 \sum_{i=1}^{n} \frac{\left[\frac{y_i^2}{2} - y_i \hat{\mu}_i + \frac{\hat{\mu}_i^2}{2} \right]}{\sigma_i^2} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\sigma_i^2} \ . \end{aligned}$$

Com a hipótese de variâncias homogéneas do Modelo Linear, $\sigma_i^2 = \sigma^2 = \phi$ para todo o i, o desvio da Normal é a Soma de Quadrados Residual:

$$D = \phi \cdot D^* = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 = SQRE \ ,$$

Desvio e desvio reduzido na Gama

Tem-se, a partir do slide 125:

$$\theta = -\frac{1}{\mu} \quad ; \quad b(\theta) = -\ln(-\theta) = \ln(\mu) \quad ; \quad \phi = \frac{1}{\nu} \quad ; \quad a(\phi) = \phi = \frac{1}{\nu} \, .$$

Logo, (slide 97) o desvio reduzido D^* vem:

$$\begin{aligned} D^* &= 2 \sum_{i=1}^{n} \frac{\left[y_i(\hat{\theta}_i^T - \hat{\theta}_i^M) - [b(\hat{\theta}_i^T) - b(\hat{\theta}_i^M)] \right]}{a(\phi_i)} \\ &= 2 \sum_{i=1}^{n} \frac{\left[y_i(-\frac{1}{y_i} + \frac{1}{\hat{\mu}_i}) - [\ln(y_i) - \ln(\hat{\mu}_i)] \right]}{\frac{1}{\nu}} = 2 \sum_{i=1}^{n} \nu_i \left[\left(\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right) - \ln \left(\frac{y_i}{\hat{\mu}_i} \right) \right] \end{aligned}$$

Admitindo que $a(\phi_i) = \phi = \frac{1}{\nu}$ (constante para todas as observações), o desvio D (slide 131) vem:

$$D = \phi \cdot D^* = 2 \sum_{i=1}^{n} \left[\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} - \ln \left(\frac{y_i}{\hat{\mu}_i} \right) \right]$$

Estimação do parâmetro de dispersão ϕ

A estimação do parâmetro de dispersão ϕ exige que se admita que é comum a todas as observações ou, quanto muito, que varia apenas devido a constantes conhecidas.

Uma forma de estimar ϕ envolve o seu estimador de máxima verosimilhança. Mas a forma mais usual de estimar ϕ envolve um tipo de resíduos, chamados resíduos de Pearson, que veremos adiante.

Mesmo em modelos de resposta Binomial ou Poisson, onde $\phi = 1$, esta estimativa da dispersão pode ser útil: um valor de $\hat{\phi}$ muito superior a 1 sugere a existência de sobredispersão, aconselhando modificações ao modelo. No `R`, essa estimativa pode ser obtida usando a opção `quasi` na definição da `family` associada ao MLG.

Resíduos e Validação do Modelo

O conceito usual de resíduos no Modelo Linear, $e_i = y_i - \hat{y}_i = y_i - \hat{\mu}_i$, tem diferentes adaptações nos MLGs, onde, diversamente do que acontecia nos Modelos Lineares, não se contempla a existência de erros aleatórios aditivos.

Em Modelos Lineares Generalizados utilizam-se diversos conceitos de resíduos, sendo os principais os

- resíduos de Pearson; e os
- resíduos do desvio.

Nos resíduos de Pearson, a diferença entre valores observados de Y_i e correspondentes estimativas dos seus valores esperados, $\widehat{E[Y_i]} = \hat{\mu}_i$, é dividida pela raíz quadrada da chamada função de variância do modelo.

Função de Variância

Função de Variância

Dado um MLG com componente aleatória Y de média $E[Y]$, variância $V[Y]$ e parâmetro de dispersão ϕ. A função $f_V(E[Y]) = \frac{V[Y]}{\phi}$ designa-se a função de variância do modelo. Tem-se: $V[Y] = \phi \cdot f_V(E[Y])$.

A função de variância é diferente para cada distribuição de Y:

- Normal: Tem-se $f_V(\mu) = \frac{V[Y]}{\phi} = \frac{\sigma^2}{\sigma^2} = 1$.
- Bernoulli: $f_V(p) = \frac{V[Y]}{1} = V[Y] = p(1-p)$.
- Binomial/n: $f_V(p) = \frac{V[Y]}{1} = V[Y] = \frac{p(1-p)}{n}$.
- Poisson: Tem-se $f_V(\lambda) = \frac{V[Y]}{1} = V[Y] = \lambda$.
- Gama: Tem-se $f_V(\mu) = \frac{V[Y]}{\phi} = \frac{\frac{\mu^2}{\nu}}{\frac{1}{\nu}} = \mu^2$.

Resíduos de Pearson

Resíduos de Pearson

Seja $Y_1, Y_2, ..., Y_n$ uma amostra aleatória da componente aleatória dum Modelo Linear Generalizado. Designa-se resíduo de Pearson de cada observação a:

$$r_i^P = \frac{Y_i - \hat{\mu}_i}{\sqrt{f_v(\hat{\mu}_i)}} .$$

- Normal: Tem-se $f_v(\mu_i) = \frac{V[Y_i]}{\sigma_i^2} = 1$. O resíduo de Pearson é o habitual resíduo do Modelo Linear:

$$r_i^P = Y_i - \hat{\mu}_i$$

- Poisson: Tem-se $f_v(\lambda_i) = \frac{V[Y_i]}{1} = \lambda_i$. O resíduo de Pearson é:

$$r_i^P = \frac{Y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$$

Resíduos de Pearson (cont.)

- Bernoulli: $f_v(p_i) = \frac{V[Y_i]}{1} = p_i(1-p_i)$. O resíduo de Pearson é:

$$r_i^P = \frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} \tag{1}$$

- Binomial/n: $f_v(p_i) = \frac{V[Y_i]}{1} = \frac{p_i(1-p_i)}{n_i}$. O resíduo de Pearson é:

$$r_i^P = \frac{Y_i - \hat{p}_i}{\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n_i}}} \tag{2}$$

- Gama: Tem-se $f_v(\mu_i) = \frac{V[Y_i]}{\phi_i} = \frac{\frac{\mu_i^2}{\nu_i}}{\frac{1}{\nu_i}} = \mu_i^2$. O resíduo de Pearson é:

$$r_i^P = \frac{Y_i - \hat{\mu}_i}{\hat{\mu}_i}$$

Resíduos de Pearson (cont.)

As expressões dos resíduos de Pearson dependem também das funções de ligação. Por exemplo, em modelos de resposta dicotómica, nas fórmulas (1) e (2) do acetato 138 tem-se,

- numa Regressão Logística:

$$\hat{p}_i = \frac{1}{1 + \mathrm{e}^{-(\hat{\beta}_0 + \hat{\beta}_1 x_{1(i)} + \ldots + \hat{\beta}_p x_{p(i)})}}$$

- Numa Regressão Probit:

$$\hat{p}_i = \Phi(\hat{\beta}_0 + \hat{\beta}_1 x_{1(i)} + \ldots + \hat{\beta}_p x_{p(i)})$$

- Num modelo Log-log do complementar:

$$\hat{p}_i = 1 - \mathrm{e}^{-\mathrm{e}^{\hat{\beta}_0 + \hat{\beta}_1 x_{1(i)} + \ldots + \hat{\beta}_p x_{p(i)}}}$$

Estatística de Pearson generalizada

No slide 137 viu-se que, para MLGs com componente aleatória Poisson, o resíduo de Pearson é dado por $r_i^P = \frac{Y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$. Como nesses modelos $\hat{\lambda}_i = E[Y_i]$, a soma de quadrados desses resíduos é a habitual estatística de Pearson dos testes Qui-quadrado, $\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$. Essa relação dá origem ao seguinte conceito.

Estatística de Pearson generalizada, X^2

Dado um MLG com resíduos de Pearson r_i^P chama-se Estatística de Pearson generalizada à soma de quadrados desses resíduos:

$$X^2 = \sum_{i=1}^{n} (r_i^P)^2 .$$

Por vezes, este valor é usado em alternativa ao desvio, como indicação da qualidade de ajustamento do modelo.

Estimação do parâmetro de dispersão ϕ

Em MLGs com parâmetro de dispersão ϕ desconhecido, uma das formas usada para estimar ϕ baseia-se nos resíduos de Pearson.

Estimador de ϕ

Num MLG com m parâmetros na componente sistemática e admitindo parâmetro de dispersão ϕ comum a todas as observações Y_i, um estimador de ϕ é dado por:

$$\hat{\phi} = \frac{X^2}{n-m} = \frac{\sum_{i=1}^{n} (r_i^P)^2}{n-m} .$$

Nota: No caso particular do Modelo Linear este estimador é o *QMRE*.

Resíduos do Desvio

Um conceito alternativo de resíduo baseia-se nas parcelas da definição do Desvio dum MLG (por analogia com a Soma de Quadrados dos Resíduos no Modelo Linear).

Resíduos do Desvio

Seja $Y_1, Y_2, ..., Y_n$ uma amostra aleatória da Componente Aleatória dum Modelo Linear Generalizado. Seja

$$D = \sum_{i=1}^{n} d_i$$

o seu Desvio. Designa-se resíduo do Desvio da observação i a:

$$r_i^D \quad = \quad sinal(y_i - \hat{\mu}_i) \cdot \sqrt{d_i}$$

Resíduos do desvio (cont.)

Concretizando:

- Normal: Tem-se $d_i = (y_i - \hat{\mu}_i)^2$. O resíduo do Desvio vem:

$$r_i^D = y_i - \hat{\mu}_i$$

Os resíduos do Desvio são os resíduos usuais do Modelo Linear.

- Bernoulli: tem-se

$$d_i = -2 \cdot [y_i \ln(\hat{p}_i) + (1 - y_i) \ln(1 - \hat{p}_i)] = \begin{cases} -2\ln(1-\hat{p}_i) & \text{se} \quad y_i = 0 \\ -2\ln(\hat{p}_i) & \text{se} \quad y_i = 1 \end{cases}$$

Os resíduos do Desvio para Y Bernoulli são:

$$r_i^D = sinal(y_i - \hat{p}_i) \cdot \sqrt{d_i} = \begin{cases} -\sqrt{-2\ln(1-\hat{p}_i)} & \text{se} \quad y_i = 0 \\ \sqrt{-2\ln(\hat{p}_i)} & \text{se} \quad y_i = 1 \end{cases}$$

Resíduos do Desvio (cont.)

- Binomial/n: tem-se

$$d_i = \begin{cases} -2n_i \left[y_i \ln\left(\frac{y_i}{\hat{p}_i}\right) + (1-y_i)\ln\left(\frac{1-y_i}{1-\hat{p}_i}\right)\right] & \text{se} \quad y_i \neq 0,1 \\ -2n_i \left[-y_i \ln(\hat{p}_i) - (1-y_i)\ln(1-\hat{p}_i)\right] & \text{se} \quad y_i \in \{0,1\}\ . \end{cases}$$

Os resíduos do Desvio para Y Binomial/n são:

$$r_i^D = \begin{cases} \sqrt{-2n_i \left[y_i \ln\left(\frac{y_i}{\hat{p}_i}\right) + (1-y_i)\ln\left(\frac{1-y_i}{1-\hat{p}_i}\right)\right]} & \text{se} \quad y_i \neq 0,1 \\ \sqrt{2n_i \left[y_i \ln(\hat{p}_i) + (1-y_i)\ln(1-\hat{p}_i)\right]} & \text{se} \quad y_i \in \{0,1\}\ . \end{cases}$$

- Poisson: Neste caso $d_i = 2 \cdot \left[y_i \ln\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i)\right]$. Os resíduos do Desvio para Y Poisson são:

$$r_i^D = sinal(y_i - \hat{\lambda}_i) \cdot \sqrt{2\left[y_i \ln\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i)\right]}$$

Resíduos do Desvio (cont.)

- Gama: neste caso

$$d_i = 2 \cdot \left[\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} - \ln\left(\frac{y_i}{\hat{\mu}_i} \right) \right]$$

Os resíduos do Desvio para Y Gama são:

$$r_i^D = sinal(y_i - \hat{\mu}_i) \cdot \sqrt{2 \cdot \left[\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} - \ln\left(\frac{y_i}{\hat{\mu}_i} \right) \right]}$$

Como em casos anteriores, a cada diferente função de ligação g corresponde uma diferente forma de obter as médias ajustadas $\hat{\mu}$, logo uma diferente expressão concreta para os resíduos do desvio.

Os Resíduos no R

Tal como no modelo linear, o R disponibiliza funções para o cálculo dos resíduos e dos resíduos normalizados.

- `residuals` calcula os resíduos (não estandardizados). Por omissão, trata-se dos resíduos do desvio.

```
> residuals(tabaco.glm)
           1            2            3            4            5            6
-0.003720807  0.631866326  1.149311810  0.834445925  1.498556259  1.944020824
           7            8            9           10           11           12
-1.435052677 -0.632286846 -0.253437946 -0.523178851 -1.708321246 -1.502372783
```

- Podem obter-se os resíduos de Pearson explicitando a opção `type=''pearson''`.

```
> residuals(tabaco.glm, type="pearson")
          1           2           3           4           5           6
-0.00371868  0.65977673  1.17907319  0.82587899  1.37017323  1.40774747
          7           8           9          10          11          12
-1.02793742 -0.60081763 -0.25159052 -0.52497400 -1.82437610 -1.71522073
```

Os Resíduos na Validação de um MLG

Os resíduos podem ser utilizados para:

- estudar a validade da hipótese distribucional associada à sua componente aleatória;
- estudar a adequabilidade da componente sistemática como preditor linear;
- estudar a adequabilidade da função de ligação escolhida;
- como diagnósticos na procura de observações com particularidades especiais.

A utilização dos resíduos tem muitas especificidades, para cada MLG concreto, sendo difícil uma discussão conjunta.

Para uma discussão mais aprofundada, sugere-se a consulta de McCullagh & Nelder (1989) ou outra bibliografia.

MLGs no estudo de tabelas de contingência

MLGs admitem variáveis preditoras quantitativas, qualitativas, ou de ambos os tipos.

Modelos Log-lineares são particularmente importantes no estudo de tabelas de contingência.

Trata-se de um contexto onde a componente aleatória corresponde a contagens (variável discreta), que se pretendem relacionar com os níveis de um ou mais factores.

São frequentes os casos onde a variável resposta se pode considerar como tendo uma distribuição de Poisson (ou ainda binomial ou a sua generalização multinomial).

(Não) Tabelas de contingência para 2 factores

Consideremos o caso frequente de tabelas de contingência com dois factores de classificação.

Exemplo: uma tabela de contagens de observações de espécies (primeiro factor) em vários locais (segundo factor).

Níveis do Factor A	Níveis do Factor B					Marginal de A
	1	2	$\cdots$	$b-1$	b	
1	n_{11}	n_{12}	$\cdots$	$n_{1,(b-1)}$	$n_{1,b}$	$n_{1\cdot}$
2	n_{21}	n_{22}	$\cdots$	$n_{2,(b-1)}$	$n_{2,b}$	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
$a-1$	$n_{(a-1),1}$	$n_{(a-1),2}$	$\cdots$	$n_{(a-1),(b-1)}$	$n_{(a-1),b}$	$n_{(a-1)\cdot}$
a	n_{a1}	n_{a2}	$\cdots$	$n_{a,(b-1)}$	$n_{a,b}$	$n_{a\cdot}$
Marginal de B	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot(b-1)}$	$n_{\cdot b}$	$n = n_{\cdot\cdot}$

(Não) Tabelas de contingência para 2 factores (cont.)

Quando não há restrições sobre o número total de observações, ou sobre qualquer das margens (como será o caso nas tabelas de locais $\times$ espécies), as contagens podem ser consideradas como observações independentes de distribuições de Poisson.

Numa situação dessas, será de considerar um modelo com algumas semelhanças aos modelos ANOVA, mas em que a variável resposta $Y_{ij} = n_{ij}$, tenha distribuição Poisson.

Neste contexto, um modelo tipo ANOVA factorial em que, além de efeitos principais de cada factor, se prevejam efeitos de interacção entre os dois factores, é um modelo saturado, uma vez que:

- há apenas 1 observação em cada uma das ab células (a contagem n_{ij});
- há ab parâmetros num modelo factorial com interacção.

(Não) A hipótese de independência

Mais útil serão modelos associados a hipóteses mais restritivas sobre a natureza da relação entre os factores associados à tabela. Em particular a hipótese de independência entre os factores pode ser interessante.

Existindo independência entre os factores, os valores esperados de $Y_{ij} = n_{ij}$ serão dados (para qualquer i e j) por:

$$E[Y_{ij}] = \lambda_{ij} = n\, p_{ij} = n\, p_{i.}\, p_{.j}$$

onde:

- n é o número total de observações;
- p_{ij} é a probabilidade duma observação recair na célula (i,j);
- $p_{i.}$ é a probabilidade marginal associada ao nível i do Factor A;
- $p_{.j}$ é a probabilidade marginal associada ao nível j do Factor B.

(Não) A hipótese de independência (cont.)

Uma vez que a distribuição Poisson é adequada à variável resposta, surge de forma natural a ideia de usar a função de ligação canónica para essa distribuição, ou seja, de logaritmizar $E[Y_{ij}]$:

$$\ln\left(E[Y_{ij}]\right) \quad = \quad \ln(n\, p_{i.}\, p_{.j}) \quad = \quad \ln(n) + \ln(p_{i.}) + \ln(p_{.j})$$

Trata-se duma relação do tipo ANOVA a dois factores, sem interacção:

$$\ln\left(E[Y_{ij}]\right) \quad = \quad \mu + \alpha_i + \beta_j$$

onde se pode considerar (embora mais tarde se modifique):

- $\mu = \ln(n)$ é uma constante comum a todas as observações;
- $\alpha_i = \ln(p_{i.})$ é um efeito associado ao nível i do factor A;
- $\beta_j = \ln(p_{.j})$ é um efeito associado ao nível j do factor B.

(Não) A hipótese de independência (cont.)

Estamos perante um Modelo Log-linear com:

- componente aleatória Poisson;
- função de ligação logarítmica (ligação canónica da Poisson);
- componente sistemática dada por variáveis indicatrizes de níveis de cada factor.

Tal como nas ANOVAs clássicas, há que impor restrições aos parâmetros e considerar a célula associada ao primeiro nível de cada factor como uma célula de referência, sendo a situação nas restantes células comparada com essa célula de referência.

(Não) As restrições aos parâmetros

Consideramos

$$\begin{aligned} \lambda_{11} &= E[Y_{11}] &&= n \cdot p_{1.} \cdot p_{.1} \\ \lambda_{ij} &= E[Y_{ij}] &&= n \cdot p_{i.} \cdot p_{.j} = \lambda_{11} \cdot \frac{p_{i.}}{p_{1.}} \cdot \frac{p_{.j}}{p_{.1}} \quad , \forall\, i = 1:a \,,\, j = 1:b \end{aligned}$$

Logaritmizando, temos as relações:

$$\ln(\lambda_{ij}) = \ln(E[Y_{ij}]) = \underbrace{\ln(\lambda_{11})}_{=\mu} + \underbrace{\ln\left(\frac{p_{i.}}{p_{1.}}\right)}_{=\alpha_i} + \underbrace{\ln\left(\frac{p_{.j}}{p_{.1}}\right)}_{=\beta_j} \quad , \forall\, i \,,\, j$$

Assim surgem de forma natural as restrições $\alpha_1 = 0$ e $\beta_1 = 0$.

(Não) Um modelo log-linear a dois factores

O valor de n, o número total de observações, é conhecido.

Os estimadores de máxima verosimilhança dos parâmetros μ, α_i e β_j são dados pelas frequências relativas marginais:

$$\hat{p}_{i.} = \frac{n_{i.}}{n} \quad \text{e} \quad \hat{p}_{.j} = \frac{n_{.j}}{n} \,,$$

pelo que

$$\begin{aligned}
\hat{\mu} &= \ln\left(n \cdot \hat{p}_{i.} \cdot \hat{p}_{.j}\right) = \ln\left(n \cdot \frac{n_{1.}}{n} \cdot \frac{n_{.1}}{n}\right) = \ln\left(\frac{n_{1.} \cdot n_{.1}}{n}\right) \\
\hat{\alpha}_i &= \ln\left(\frac{\hat{p}_{i.}}{\hat{p}_{1.}}\right) = \ln\left(\frac{n_{i.}}{n_{1.}}\right) \\
\hat{\beta}_j &= \ln\left(\frac{\hat{p}_{.j}}{\hat{p}_{.1}}\right) = \ln\left(\frac{n_{.j}}{n_{.1}}\right)
\end{aligned}$$

(Não) O Desvio mede afastamento da independência

Já se viu que saturar este modelo log-linear a dois factores corresponde a prever efeitos de interacção. Nesse modelo, cada célula é livre de ter o seu valor, sem qualquer estrutura especial associada à tabela.

O Desvio do modelo sem interacção

$$D^* = -2\left(\mathscr{L}_M(\vec{\hat{\boldsymbol{\beta}}}_M) - \mathscr{L}_T(\vec{\hat{\boldsymbol{\beta}}}_T)\right)$$

corresponde ao valor da estatística de Wilks para uma comparação do submodelo (M) sem interacção (isto é, a hipótese de independência) face ao modelo saturado (T), com interacção (sem qualquer relação especial). Quanto menor D^*, mais os dados se comportam de acordo com a hipótese de independência. Pelo contrário, quanto maior D^*, menos plausível a hipótese de independência.

(Não) Exemplo: modelo para tabela de contingência

Dados `HairEyeColor` (para ambos os sexos)

Na *data frame* `cabelo.olho` há $n = 16$ contagens numa tabela cruzando 4 côres de cabelo e 4 côres de olhos, num grupo de $N = 592$ estudantes.

```
> cabelo.olho
   contagens    cabelo     olhos
1         68     preto castanhos
2        119 castanho castanhos
3         26    ruivo castanhos
4          7    louro castanhos
5         20     preto      azuis
6         84 castanho      azuis
7         17    ruivo      azuis
8         94    louro      azuis
9         15     preto cinzentos
10        54 castanho cinzentos
11        14    ruivo cinzentos
12        10    louro cinzentos
13         5     preto     verdes
14        29 castanho     verdes
15        14    ruivo     verdes
16        16    louro     verdes
```

```
> cabeloOlho
               Cabelo
Olhos         preto  castanho  ruivo  louro
   castanhos     68       119     26      7
   azuis         20        84     17     94
   cinzentos     15        54     14     10
   verdes         5        29     14     16
```

Nota: Estes dados encontram-se na *data frame* `HairEyeColor` da distribuição base do `R`, e resultam de somar os valores relativos a ambos os sexos.

(Não) Exemplo (cont.)

```
> cabelo.glm <- glm(contagens ~ cabelo + olhos, family=poisson, data=cabelo.olho)
> summary(cabelo.glm)

Call: glm(formula = contagens ~ cabelo + olhos, family = poisson, data = cabelo.olho)
(...)
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     4.64312    0.08036  57.776  < 2e-16 ***
cabelolouro    -0.81180    0.10663  -7.613 2.68e-14 ***
cabelopreto    -0.97386    0.11294  -8.623  < 2e-16 ***
cabeloruivo    -1.39331    0.13259 -10.508  < 2e-16 ***
olhoscastanhos  0.02299    0.09590   0.240    0.811
olhoscinzentos -0.83804    0.12411  -6.752 1.46e-11 ***
olhosverdes    -1.21175    0.14239  -8.510  < 2e-16 ***
(...)
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 453.31  on 15  degrees of freedom
Residual deviance: 146.44  on  9  degrees of freedom
AIC: 241.04

Number of Fisher Scoring iterations: 5
```

Nota: Neste contexto, o modelo ajustado corresponde à hipótese de independência.

O modelo Nulo corresponde a admitir que as contagens esperadas de todas as células são iguais, sendo estimadas por $\frac{N}{n} = \frac{592}{16} = 37$.

(Não) Exemplo (cont.)

O modelo log-linear de tipo ANOVA a 2 factores, mas com efeitos de interacção corresponde, como se viu, a um modelo saturado:

```
> cabelo.glmT <- glm(contagens ~ cabelo * olhos, family=poisson, data=cabelo.olho)
> summary(cabelo.glmT)
[...]
Deviance Residuals:
[1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
[...]
Null deviance: 4.5331e+02  on 15  degrees of freedom
Residual deviance: 5.9952e-15  on  0  degrees of freedom
AIC: 112.6
```

O teste de Wilks comparando o modelo saturado e o modelo de independência avalia (e rejeita) a hipótese de independência:

```
> anova(cabelo.glm, cabelo.glmT, test="Chisq")
Analysis of Deviance Table

Model 1: contagens ~ cabelo + olhos
Model 2: contagens ~ cabelo * olhos
      Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1         9    146.44
2         0      0.00       9   146.44 < 2.2e-16 ***
```

(Não) Exemplo (cont.)

Por definição, o desvio é a soma dos quadrados dos resíduos do desvio.

```
> sum(residuals(cabelo.glm)^2)
[1] 146.4436
```

A soma dos quadrados dos resíduos de Pearson tem um valor próximo.

```
> sum(residuals(cabelo.glm, type="pearson")^2)
[1] 138.2898
```

Esta última soma de quadrados é também o valor da usual estatística do teste χ^2 de independência:

```
> chisq.test(cabeloOlho)
Pearson's Chi-squared test
data:  cabeloOlho
X-squared = 138.29, df = 9, p-value < 2.2e-16
```

(Não) Tabelas de contingência (cont.)

O exemplo de uma tabela de dupla entrada foi sobretudo ilustrativo. O interesse maior de modelos log-lineares corresponde ao estudo de tabelas definidas por três ou mais factores.

A diferentes conceitos de independência envolvendo três ou mais factores (independência, independência mútua, independência conjunta, independência condicional, etc.) correspondem diferentes modelos log-lineares.

A validade de um ou outro conceito de independência pode ser estudada através da qualidade do ajustamento do correspondente modelo.

(Não) Tabela de independências

A tabela indica as designações mnemónicas para os vários tipos de modelos considerados até aqui.

Notação	Tipo de Modelo	Equação do Modelo para $\ln(\lambda_{ijk})$	Relação-base
(A,B,C)	Independência Mútua	$\mu+\alpha_i+\beta_j+\gamma_k$	$p_{ijk}=p_{i..}\cdot p_{.j.}\cdot p_{..k}$
(B:C)	Ind. conjunta (B,C) com A	$\mu+\alpha_i+\beta_j+\gamma_k+(\beta\gamma)_{jk}$	$p_{ijk}=p_{i..}\cdot p_{.jk}$
(A:B)	Ind. conjunta (A,B) com C	$\mu+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}$	$p_{ijk}=p_{ij.}\cdot p_{..k}$
(A:C)	Ind. conjunta (A,C) com B	$\mu+\alpha_i+\beta_j+\gamma_k+(\alpha\gamma)_{ik}$	$p_{ijk}=p_{i.k}\cdot p_{.j.}$
(A:C,B:C)	Ind. (A,B) condicional a C	$\mu+\alpha_i+\beta_j+\gamma_k+(\alpha\gamma)_{ik}+(\beta\gamma)_{jk}$	$p_{ijk}=\frac{p_{i.k}\cdot p_{.jk}}{p_{..k}}$
(A:B,B:C)	Ind. (A,C) condicional a B	$\mu+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\beta\gamma)_{jk}$	$p_{ijk}=\frac{p_{ij.}\cdot p_{.jk}}{p_{.j.}}$
(A:B,A:C)	Ind. (B,C) condicional a A	$\mu+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\alpha\gamma)_{ik}$	$p_{ijk}=\frac{p_{ij.}\cdot p_{i.k}}{p_{i..}}$
(A:B:C)	Modelo Saturado	$\mu+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\alpha\gamma)_{ik}+(\beta\gamma)_{jk}+(\alpha\beta\gamma)_{ijk}$	