Modelos Matemáticos e Aplicações
Modelos Lineares Generalizados

Jorge Cadima

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

2016-17

Bibliografia

- McCullaugh, P. & Nelder, J. (1989) *Generalized Linear Models*, Chapman and Hall.
- McCulloch, C. & Searle, S. (2001) *Generalized, Linear, and Mixed Models*, John Wiley & Sons. **Mat 600-62**.
- Hosmer, D.W. & Lemeshow, S. (1989) *Applied Logistic Regression*, John Wiley & Sons. **Mat 258-62**.
- Agresti, A. (1990) *Categorical Data Analysis*, John Wiley & Sons. **Mat 401-62**.
- Agresti, A. (2005) *Foundations of Linear and Generalized Linear Models*, Wiley.
- Turkman, M.A.A. & Silva, G.L. (2000) *Modelos Lineares Generalizados* (Mini-curso no VIII Congresso da Soc. Port. Estatística), Ed. SPE

GLMs no  :

- Venables & Ripley (2002). *Modern Applied Statistics with S* (4a. edição), Springer. ISBN 0-387-95457-0. (módulo `R`: `MASS`).
- John Fox and Sanford Weisberg (2011). *An R Companion to Applied Regression, 2d Ed*. Sage Publications. (módulo `R`: `car`).

Modelos Lineares Generalizados

Os Modelos Lineares Generalizados (MLGs ou GLMs pela ordem inglesa)

- são uma família muito vasta de modelos;
- generalizam o Modelo Linear;
- o "chapéu de chuva comum" dos MLGs foi introduzido e formalizado por McCullagh e Nelder (1989);
- mas englobando muitos modelos já conhecidos e que, nalguns casos, eram utilizados há largas décadas, entre eles:
  - modelo *probit*
  - modelo *logit*
  - modelos log-lineares
  - o próprio modelo linear.

Exemplo motivador: variável resposta dicotómica

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

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

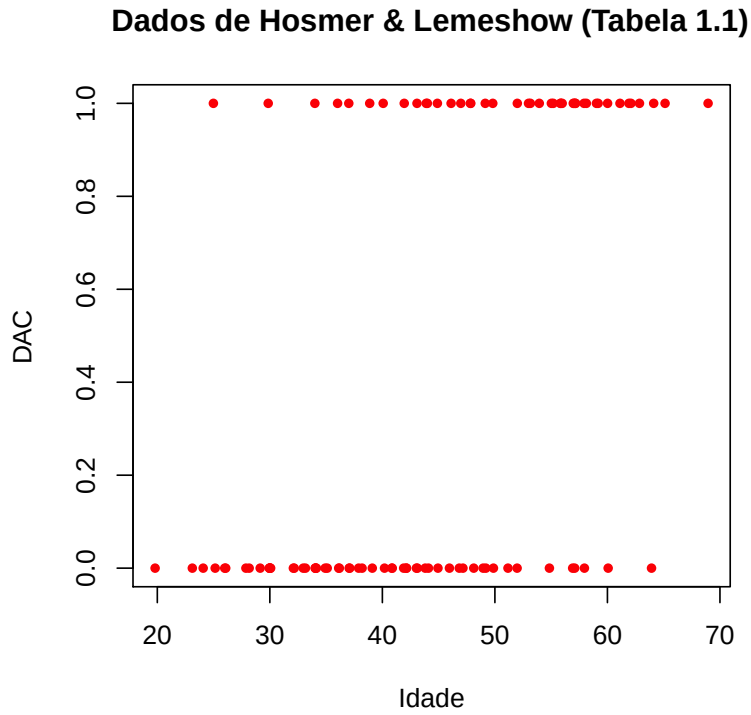
Eis os primeiros seis valores da *data frame* correspondente:

```
> head(HosLem.tudo)
```

```
  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



NOTA: Foi usado o comando `jitter` na variável idade:

```
> plot(DAC ~ jitter(Idade), data=HosLem.tudo, 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 (numéricos e/ou factores),

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 $\varepsilon_i \cap \mathcal{N}(0, \sigma^2)$ $(i = 1, 2, ..., n)$.
isto é, tal que $E[Y_i | X_1 = x_{1_{(i)}}, ..., X_p = x_{p_{(i)}}]$ é dada por:

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

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.

Modelo Linear Generalizado (MLG)

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

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

1) Componente aleatória: 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.

Havendo p variáveis preditoras e n observações:

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

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.)

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

A componente sistemática do modelo é dada 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: uma função diferenciável e monótona g (e a sua correspondente função vectorial $\vec{g}$) que associa as componentes aleatória e sistemática, através duma relação da forma:

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

sendo:

- $\vec{\mathbf{Y}}$ o vector com as n observações $\{Y_i\}_{i=1}^n$.
- $\vec{\mu} = E[\vec{\mathbf{Y}}] = (\mu_1, \mu_2, ..., \mu_n)^t$ o vector esperado de $\vec{\mathbf{Y}}$;
- $\vec{\mathbf{x}}_{[i]}^t$ a i-ésima linha da matriz $\mathbf{X}$, com os valores dos preditores na i-ésima observação da variável resposta.

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 usados erros aleatórios aditivos. A flutuação aleatória da variável-resposta é dada directamente pela sua distribuição de probabilidades.

Caso a função g seja invertível (o que sucede se a monotonia acima exigida fôr estrita), pode escrever-se:

$$g(\mu_i) = \vec{\mathbf{x}}_{[i]}^t \vec{\boldsymbol{\beta}} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{j(i)} \quad \Leftrightarrow \quad \mu_i = g^{-1}\left(\vec{\mathbf{x}}_{[i]}^t \vec{\boldsymbol{\beta}}\right) = g^{-1}\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{j(i)}\right)$$

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)
- 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).

A família exponencial de distribuições (cont.)

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 pode 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:

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

é da forma indicada, com:

- $\theta = \mu$
- $\phi = \sigma^2$
- $b(\theta) = \frac{\theta^2}{2} = \frac{\mu^2}{2}$
- $a(\phi) = \phi = \sigma^2$
- $c(y,\phi) = \ln\left(\frac{1}{\sqrt{2\pi\phi}}\right) - \frac{y^2}{2\phi} = \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{y^2}{2\sigma^2}$

A Poisson

Recorde-se que uma variável aleatória discreta tem distribuição de Poisson se toma valores em $\mathbb{N}_0$ com função de massa probabilística

$$P[Y = y] = \frac{\lambda^y}{y!} e^{-\lambda} .$$

Para os valores $y \in \{0, 1, 2, ...\}$, podemos escrever a função de massa probabilística duma Poisson como:

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

que é da família exponencial com:

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

A Bernoulli

A variável aleatória dicotómica – ou seja, 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:

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

que é da familia exponencial com:

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

A Binomial/n

A Binomial não pertence à família exponencial de distribuições.
Mas se $X \cap B(n,p)$, então $Y = \frac{1}{n}X$ pertence à família exponencial.

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

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

que é da familia exponencial com:

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

A Gama

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

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

que é da familia exponencial com:

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

A família das distribuições Gama inclui como caso particular a distribuição Qui-quadrado (χ^2_n se $\nu = \frac{n}{2}$ e $\mu = n$) e também a distribuição Exponencial ($\nu = 1$).

Funções de ligação

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

As mais importantes funções de ligação tornam, para cada distribuição da família exponencial, o valor esperado da variável-resposta igual ao parâmetro natural, θ.

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

As funções de ligação canónica são úteis porque simplificam de forma assinalável 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.

Algumas funções de ligação canónicas

Distribuição	$\mu = E[Y]$	Ligação canónica
Normal	μ	Identidade: $g(\mu) = \mu = \Theta$
Poisson	λ	Log: $g(\lambda) = \ln(\lambda) = \Theta$
Bernoulli	p	Logit: $g(p) = \ln\left(\frac{p}{1-p}\right) = \Theta$
Binomial/n	p	Logit: $g(p) = \ln\left(\frac{p}{1-p}\right) = \Theta$
Gama	μ	Recíproco: $g(\mu) = \frac{1}{\mu} = -\Theta$

O Modelo Linear como um MLG

Eis alguns exemplos de MLGs:

1) 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.

A função de ligação identidade é a ligação canónica para a distribuição Normal.

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, com probabilidades p (para 1) e $1-p$ (para 0).

Admite-se que o parâmetro p varia nas n observações de Y, e o valor esperado da i-ésima observação de Y é dado por:

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

Uma função de ligação vai relacionar este valor esperado p_i da variável-resposta com uma combinação linear dos preditores:

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

A Regressão Logística

2) A Regressão Logística.

A função de ligação canónica transforma p no parâmetro natural θ da distribuição Bernoulli: $\theta = \ln\left(\frac{p}{1-p}\right)$. Logo, a função de ligação canónica para variáveis resposta de Bernoulli é a função *logit*:

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

Com estas opções, o MLG 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*. É habitual designar a função de ligação *logit* como um *log-odds ratio*.

A Regressão Logística (cont.)

Consideramos que os *logits* dos valores esperados p_i são combinações lineares das variáveis preditoras $X_0, X_1, ..., X_p$. Concretamente, dado um conjunto $\vec{\mathbf{x}} = (x_1, x_2, ..., x_p)$ de observações nas variáveis preditoras, tem-se:

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

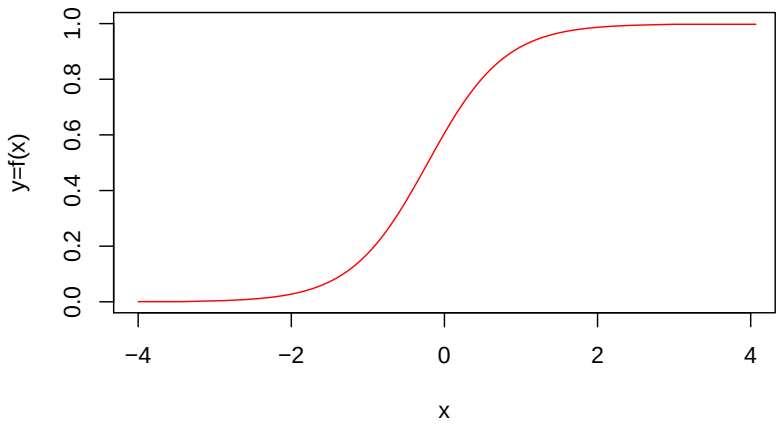
Logo, a relação entre o valor esperado de Y_i (a probabilidade de êxito de Y) e o vector de valores das variáveis preditoras, $\vec{\mathbf{x}}_i$, é:

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

A Regressão Logística (cont.)

No caso duma única variável preditora quantitativa, a relação entre Y e X é uma curva logística, que origina o nome Regressão Logística.

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

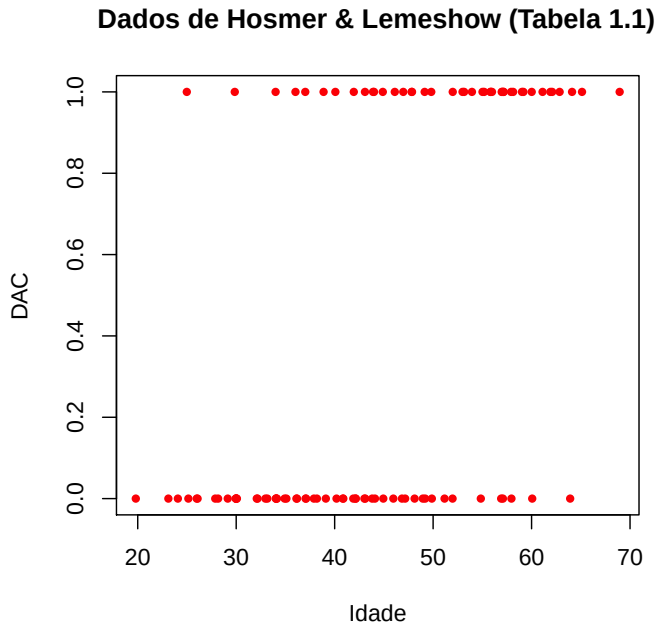


É uma função crescente, caso $\beta_1 > 0$, e decrescente caso $\beta_1 < 0$.

Quando há vários preditores, Y tem relação logística com a parte sistemática $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$.

Novamente o exemplo DAC

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



A variável resposta é dicotómica (binária): aplica-se uma regressão logística? Será preciso relacionar $p = E[Y]$, a probabilidade de ter doença arterial coronária, com a idade X.

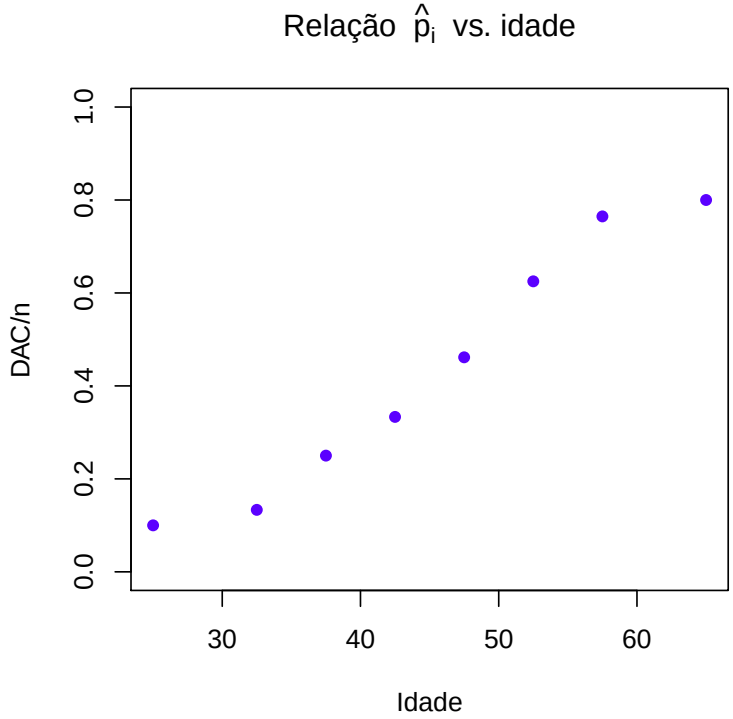
Exemplo: A função de ligação

Para procurar uma função de ligação adequada, é necessário 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



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

Temos uma relação sigmóide, talvez logística.

Exemplo: os dados tabelados

```
> HosLem <-data.frame(Idade=c(25,32.5,37.5,42.5,47.5,52.5,57.5,65),
+                     n=c(10,15,12,15,13,8,17,10),DAC=c(1,2,3,5,6,5,13,8))

> rownames(HosLem) <- c("20-30-","30-35-","35-40-","40-45-","45-50-",
+                       "50-55-","55-60-","60-70-")

> HosLem
```

	Idade	n	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 anterior foi obtido com o comando

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

Resposta dicotómica e Binomial

A tabela anterior, resultante de agrupar as idades em classes, transformou a variável resposta Y_i, Bernoulli (1/0), 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: $Y_j \cap B(n_j, p_j)$, onde p_j é a probabilidade de "êxito" na classe j.

Já vimos que a Binomial não pertence à família de distribuições exponenciais. Mas se $Y \cap B(n, p)$, então a distribuição da proporção de êxitos $W = \frac{1}{n} Y$ pertence à família exponencial.

A distribuição "Binomial/n"

Existem ligações íntimas, no contexto de MLGs, entre considerar que:

- temos n variáveis resposta Bernoulli, com parâmetros p_i; ou
- temos m variáveis resposta $Y_i \cap B(n_i, p_i)$.

O tratamento destas opções alternativas é igual, desde que transforme as Binomiais Y_i em proporções de êxitos, i.e., desde que se considere novas v.a.s resposta $W_i = Y_i/n_i$, cujas distribuições pertencem à família exponencial de distribuições.

Bernoulli e "Binomial/n" podem ser vistas como essencialmente a mesma coisa, apresentada de forma diferente. A ligação canónica, quer da Bernoulli, quer da Binomial/n é a função *logit*:

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

GLMs no R

No R, o comando crucial para o ajustamento de Modelos Lineares Generalizados é o comando `glm`.

Dos numerosos argumentos desta função, dois são cruciais:

formula indica, de forma análoga à usada no modelo linear, qual a componente aleatória (à esquerda dum "$\sim$") e quais os preditores (à direita, e separados por sinais de soma):

$$y \sim x1 + x2 + x3 + \ldots + xp$$

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

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.
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* `HosLem.tudo`);
  - observações tabeladas para valores repetidos do(s) preditor(es)
    (como a *data frame* `HosLem`).

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 correspondem objectos y de natureza diferente:

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

```
> glm(DAC ~ Idade , family=binomial , data=HosLem.tudo)
```

- 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 do(s) preditor(es):

```
> glm(cbind(DAC,n-DAC) ~ Idade , family=binomial, data=HosLem)
```

Exemplo: ajustamento do modelo

Ajustar o modelo com base nas observações dicotómicas individuais:

> glm(DAC ~ idade, family=binomial, data=HosLem.tudo)

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

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:     136.7
Residual Deviance: 107.4  AIC: 111.4
```

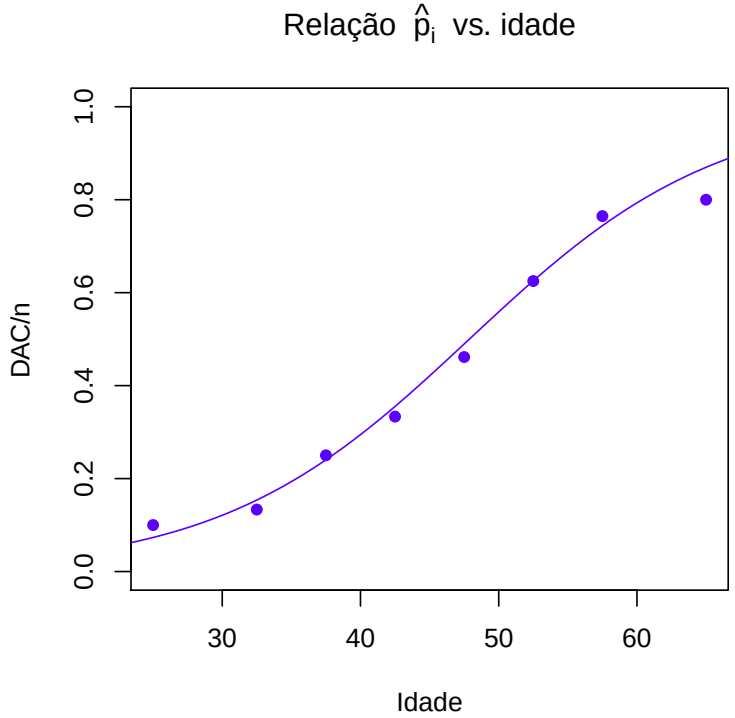
A equação da logística ajustada é:

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

Exemplo: ajustamento do modelo (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.)

Ajustar o modelo com base nos dados tabelados:

```
> glm(cbind(DAC,n-DAC) ~ Idade , family=binomial, data=HosLem)

Call:  glm(formula = cbind(DAC, n - DAC) ~ Idade, family = binomial,
    data = HosLem)
Coefficients:
(Intercept)        Idade
     -5.091        0.105          <----  parâmetros estimados

Degrees of Freedom: 7 Total (i.e. Null);  6 Residual
Null Deviance:     28.7
Residual Deviance: 0.5242  AIC: 25.66
```

A equação da logística ajustada é:

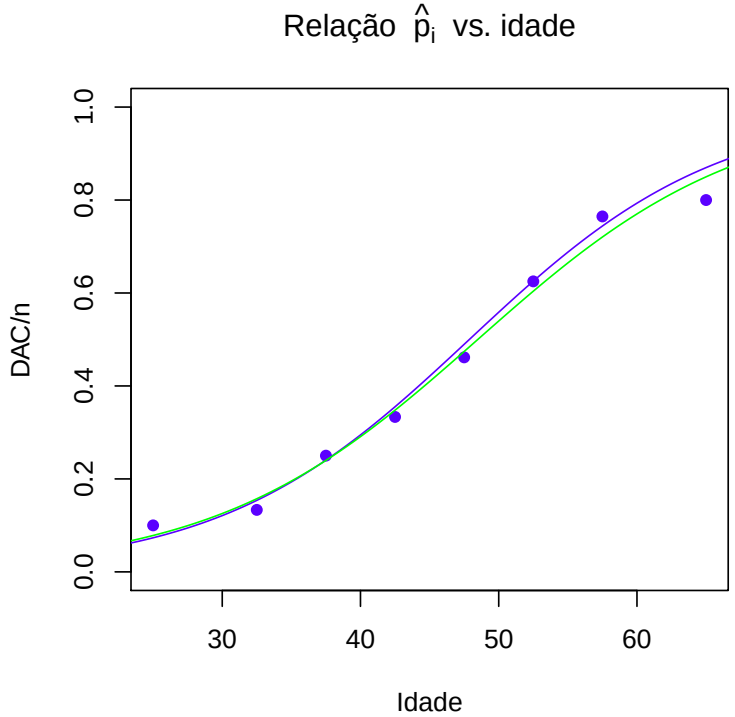
$$y = \frac{1}{1+\mathrm{e}^{-(b_0+b_1 x)}} = \frac{1}{1+\mathrm{e}^{-(-5.091+0.105x)}}$$

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.)

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 da função `glm`

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

```
> HosLem.glm <- glm(cbind(DAC,n-DAC) ~ Idade , family=binomial, data=HosLem)
> names(HosLem.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:

```
> HosLem.glm$coef
```

```
(Intercept)       Idade
 -5.0907332   0.1050191
```

A função `coef`

Tal como para os Modelos Lineares, existem funções para facilitar a extracção de informação dum ajustamento de MLG. Algumas funções iniciais:

`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:

```
> HosLem.tudo.glm <- glm(DAC ~ idade, family=binomial, data=HosLem.tudo)
> coef(HosLem.tudo.glm)

(Intercept)       idade
 -5.3094534   0.1109211
```

A função `predict`

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

```
> predict(HosLem.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(HosLem.tudo.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
```

A função `predict` (cont.)

Pode também 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:

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

```
        1
-2.425504
```

```
> predict(HosLem.glm, newdata=data.frame(Idade=c(26,53,74)))
```

```
         1          2          3
-2.3602358  0.4752807  2.6806824
```

A função `fitted`

`fitted` – devolve um vector com 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(HosLem.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
```

Um resultado análogo pode ser obtido através da função `predict`, indicando a opção `type=‘‘response’’`:

```
> predict(HosLem.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
```

Esta última opção permite também estimar os valores de $\hat{p}$ para novos conjuntos de valores dos preditores:

```
> predict(HosLem.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,

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

  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 que dão à variável aleatória Y os 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

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

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.

A Regressão Logística (cont.)

O modelo de regressão logística é uma opção a considerar sempre que a variável-resposta Y assinala qual de duas categorias de classificação se verifica e se pretende relacionar a probabilidade do acontecimento associado ao valor 1 com um conjunto de variáveis preditoras.

A função logística revela rigidez estrutural, com um ponto de inflexão associado à probabilidade $p = 0.5$, em torno do qual há simetria da curva.

A função de ligação g pode ser substituída por outras funções, que já não serão funções de ligação canónicas para uma distribuição Bernoulli. Nesse caso, já não se fala em regressão logística.

Estimação de parâmetros em MLGs

A estimação de parâmetros em Modelos Lineares Generalizados é feita pelo Método da Máxima Verosimilhança. A estimação incide em primeiro lugar sobre os parâmetros β_j da parte sistemática do modelo.

O facto da estimação se basear na função verosimilhança significa que, ao contrário do que acontece com o Modelo Linear, em GLMs 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:

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

Maximizar a verosimilhança é maximizar a log-verosimilhança:

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

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 desconhecidos $\vec{\boldsymbol{\beta}}$.

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 (pelo menos para os 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:

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

Uma vez que a função de ligação é dada por $g(p) = \ln\left(\frac{p}{1-p}\right) = \vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}$, tem-se a seguinte expressão para a log-verosimilhança como função dos parâmetros $\vec{\boldsymbol{\beta}}$:

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

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:

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

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

Material Complementar: Regressão Logística (cont.)

A não-linearidade nos parâmetros $\vec{\boldsymbol{\beta}}$ não permite explicitar uma solução $\vec{\hat{\boldsymbol{\beta}}}$ do sistema de equações.

Mas existe uma notação mnemónica, definindo o vector $\vec{\hat{\mathbf{p}}}$ de probabilidades estimadas, cuja i-ésima componente é dada por:

$$\hat{p}_i = \frac{e^{\hat{\beta}_0+\sum_{k=1}^{p} x_{k(i)}\hat{\beta}_k}}{1+e^{\hat{\beta}_0+\sum_{k=1}^{p} x_{k(i)}\hat{\beta}_k}}$$

e uma matriz $\mathbf{X}$ que (tal como no Modelo Linear) tem uma primeira coluna de n uns e em cada uma de p colunas adicionais tem as n observações de uma das p variáveis preditoras. Com esta notação, o sistema de $p+1$ equações toma a forma:

$$\mathbf{X}^t\vec{\mathbf{y}} = \mathbf{X}^t\vec{\hat{\mathbf{p}}}$$

Sendo um sistema não-linear, a sua solução exigirá métodos numéricos que serão considerados mais adiante.

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:

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

Um algoritmo numérico de resolução utilizado no contexto de MLGs é uma modificação do algoritmo de Newton-Raphson, conhecida por vários nomes: Método Iterativo de Mínimos Quadrados Ponderados (IWLS) ou Re-ponderados (IRLS), ou ainda Método de Fisher (*Fisher Scoring Method*, em inglês).

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

Algoritmos de estimação (cont.)

Designando por:

- $\vec{\boldsymbol{\beta}}^{[0]}$, a solução inicial para $\vec{\boldsymbol{\beta}}$;
- $\frac{\partial \mathscr{L}}{\partial \vec{\boldsymbol{\beta}}}(\vec{\boldsymbol{\beta}})$ o vector gradiente de $\mathscr{L}(\vec{\boldsymbol{\beta}})$ calculado no ponto $\vec{\boldsymbol{\beta}}$;
- $\mathscr{H}_{\vec{\boldsymbol{\beta}}}$ a matriz Hessiana das segundas derivadas parciais da função $\mathscr{L}(\cdot)$, nesse mesmo ponto,

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

$$\mathscr{L}(\vec{\boldsymbol{\beta}}) \approx \mathscr{L}_0(\vec{\boldsymbol{\beta}}) = \mathscr{L}(\vec{\boldsymbol{\beta}}^{[0]}) + \left(\frac{\partial \mathscr{L}}{\partial \vec{\boldsymbol{\beta}}}(\vec{\boldsymbol{\beta}}^{[0]})\right)^t \left(\vec{\boldsymbol{\beta}} - \vec{\boldsymbol{\beta}}^{[0]}\right) + \\ + \frac{1}{2}\left(\vec{\boldsymbol{\beta}} - \vec{\boldsymbol{\beta}}^{[0]}\right)^t \mathscr{H}_{\vec{\boldsymbol{\beta}}^{[0]}} \left(\vec{\boldsymbol{\beta}} - \vec{\boldsymbol{\beta}}^{[0]}\right)$$

Em vez de maximizar $\mathscr{L}(\vec{\boldsymbol{\beta}})$, maximiza-se a aproximação $\mathscr{L}_0(\vec{\boldsymbol{\beta}})$.

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,

$$\frac{\partial \mathscr{L}_0}{\partial \vec{\boldsymbol{\beta}}}(\vec{\boldsymbol{\beta}}) = \frac{\partial \mathscr{L}}{\partial \vec{\boldsymbol{\beta}}}(\vec{\boldsymbol{\beta}}^{[0]}) + \mathscr{H}_{\vec{\boldsymbol{\beta}}^{[0]}} \left(\vec{\boldsymbol{\beta}} - \vec{\boldsymbol{\beta}}^{[0]} \right).$$

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

$$\frac{\partial \mathscr{L}_0}{\partial \vec{\boldsymbol{\beta}}}(\vec{\boldsymbol{\beta}}) = 0 \quad \Leftrightarrow \quad \vec{\boldsymbol{\beta}} = \vec{\boldsymbol{\beta}}^{[0]} - \mathscr{H}^{-1}_{\vec{\boldsymbol{\beta}}^{[0]}} \left(\frac{\partial \mathscr{L}}{\partial \vec{\boldsymbol{\beta}}}(\vec{\boldsymbol{\beta}}^{[0]}) \right).$$

O algoritmo Newton-Raphson itera esta relação.

Algoritmos de estimação (cont.)

Tome-se:

$$\vec{\boldsymbol{\beta}}^{[i+1]} = \vec{\boldsymbol{\beta}}^{[i]} - \mathscr{H}^{-1}_{\vec{\boldsymbol{\beta}}^{[i]}} \left(\frac{\partial \mathscr{L}}{\partial \vec{\boldsymbol{\beta}}} (\vec{\boldsymbol{\beta}}^{[i]}) \right)$$

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 a partir de qualquer ponto inicial $\vec{\boldsymbol{\beta}}^{[0]}$, mesmo quando existe e é único o máximo da função log-verosimilhança;
- Dada a existência e unicidade do máximo, a convergência é tanto melhor quanto mais próximo $\vec{\boldsymbol{\beta}}^{[0]}$ estiver do máximo.

Algoritmos de estimação (cont.)

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

O algoritmo de Fisher 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:

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

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

$$\vec{\boldsymbol{\beta}}^{[i+1]} = \vec{\boldsymbol{\beta}}^{[i]} + \mathscr{I}^{-1}_{\vec{\boldsymbol{\beta}}^{[i]}}\left(\frac{\partial \mathscr{L}}{\partial \vec{\boldsymbol{\beta}}}(\vec{\boldsymbol{\beta}}^{[i]})\right)$$

Material Complementar: 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.

MC: Algoritmos de estimação (cont.)

O algoritmo de Fisher é também conhecido por Método Iterativo de Mínimos Quadrados Ponderados (IWLS) ou Re-ponderados (IRLS) porque é, em geral, possível re-escrever a expressão anterior para $\vec{\boldsymbol{\beta}}^{[i+1]}$ na forma:

$$\vec{\boldsymbol{\beta}}^{[i+1]} = \left(\mathbf{X}^t \mathbf{W}^{[i]} \mathbf{X}\right)^{-1} \mathbf{X}^t \mathbf{W}^{[i]} \vec{\mathbf{z}}^{[i]}$$

onde:

- $\vec{\mathbf{z}}^{[i]}$ é uma linearização da função de ligação $g(y)$, escrita como função dos parâmetros $\vec{\boldsymbol{\beta}}$; e
- $\mathbf{W}^{[i]}$ é uma matriz diagonal.

Para alguns modelos, as expressões concretas de $\vec{\mathbf{z}}^{[i]}$ e $\mathbf{W}^{[i]}$ serão vistas adiante.

MC: Algoritmos de estimação (cont.)

A expressão anterior significa que o algoritmo de Fisher está associado a uma projecção não-ortogonal, em que, quer o vector $\vec{\mathbf{z}}^{[i]}$, quer os subespaços envolvidos na projecção, são re-definidos em cada iteração do algoritmo.

A matriz $\mathbf{X}\left(\mathbf{X}^t\mathbf{W}^{[i]}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{W}^{[i]}$ é idempotente.

Não é, em geral, simétrica, a não ser que a matriz diagonal $\mathbf{W}^{[i]}$ verifique $\mathbf{X}^t\mathbf{W}^{[i]} = \mathbf{X}^t$.

O Método de Fisher baseia-se em ideias de Mínimos Quadrados em sentido generalizado, isto é, envolvendo projecções não-ortogonais.

MC: IRLS para a Regressão Logística

Viu-se que as derivadas parciais de 1a. ordem da log-verosimilhança, na Regressão Logística (e com a convenção $x_{0(i)} = 1$), são:

$$\frac{\partial \mathscr{L}(\vec{\boldsymbol{\beta}})}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{j(i)} - \sum_{i=1}^{n} \underbrace{\frac{e^{\sum_{k=0}^{p} x_{k(i)}\beta_k}}{1 + e^{\sum_{k=0}^{p} x_{k(i)}\beta_k}}}_{=p_i} \cdot x_{j(i)}, \qquad \forall j = 0 : p$$

$$\Leftrightarrow \quad \frac{\partial \mathscr{L}(\vec{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} = \mathbf{X}^t\mathbf{y} - \mathbf{X}^t\mathbf{p}$$

As derivadas parciais de 2a. ordem (elementos da Hessiana) são:

$$\frac{\partial^2 \mathscr{L}}{\partial \beta_j \partial \beta_l}\left(\vec{\boldsymbol{\beta}}\right) = -\sum_{i=1}^{n} x_{j(i)} x_{l(i)} \cdot \underbrace{\frac{e^{\sum_{k=0}^{p} x_{k(i)}\beta_k}}{1 + e^{\sum_{k=0}^{p} x_{k(i)}\beta_k}}}_{= p_i} \cdot \underbrace{\frac{1}{1 + e^{\sum_{k=0}^{p} x_{k(i)}\beta_k}}}_{= 1 - p_i}$$

$$= -\sum_{i=1}^{n} x_{j(i)} x_{l(i)} \cdot p_i \cdot (1 - p_i)$$

MC: IRLS para a Regressão Logística (cont.)

A matriz Hessiana da função de log-verosimilhança $\mathscr{L}$, nos pontos correspondentes às iterações $\vec{\boldsymbol{\beta}}^{[i]}$, é constituída pelos valores destas derivadas parciais de segunda ordem.

Como acontece sempre quando se trabalha com Modelos que utilizam a função de ligação canónica, estes elementos das matrizes Hessianas não dependem dos valores observados da variável resposta Y, pelo que a Hessiana e o seu valor esperado coincidem (os Métodos de Newton-Raphson e de Fisher coincidem).

Defina-se a matriz $n \times n$ diagonal $\mathbf{W}$, cujos elementos diagonais são dados pelos n valores $p_i(1-p_i)$.

MC: IRLS para a Regressão Logística (cont.)

A matriz Hessiana e a matriz de informação de Fisher associada, podem escrever-se, em termos matriciais, como:

$$\mathbf{H} = -\mathbf{X}^t\mathbf{W}\mathbf{X} \qquad \mathscr{I} = \mathbf{X}^t\mathbf{W}\mathbf{X}$$

A equação que define a iteração dos vectores $\vec{\boldsymbol{\beta}}$ no algoritmo IRLS para a Regressão Logística é assim:

$$\vec{\boldsymbol{\beta}}^{[i+1]} = \vec{\boldsymbol{\beta}}^{[i]} + \left(\mathbf{X}^t\mathbf{W}^{[i]}\mathbf{X}\right)^{-1}\mathbf{X}^t\left(\vec{\mathbf{y}} - \vec{\mathbf{p}}^{[i]}\right)$$

Definindo o vector $\vec{\mathbf{z}}^{[i]} = \mathbf{X}\vec{\boldsymbol{\beta}}^{[i]} + \left(\mathbf{W}^{[i]}\right)^{-1}\left(\vec{\mathbf{y}} - \vec{\mathbf{p}}^{[i]}\right)$, tem-se:

$$\vec{\boldsymbol{\beta}}^{[i+1]} = \left(\mathbf{X}^t\mathbf{W}^{[i]}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{W}^{[i]}\vec{\mathbf{z}}^{[i]}$$

A Regressão Probit

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

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 como uma função logística da componente sistemática $\vec{\eta} = \vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}}$.

Aqui, escolhe-se uma outra relação sigmóide: a função de distribuição cumulativa (f.d.c.), Φ, duma Normal Reduzida.

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

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

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

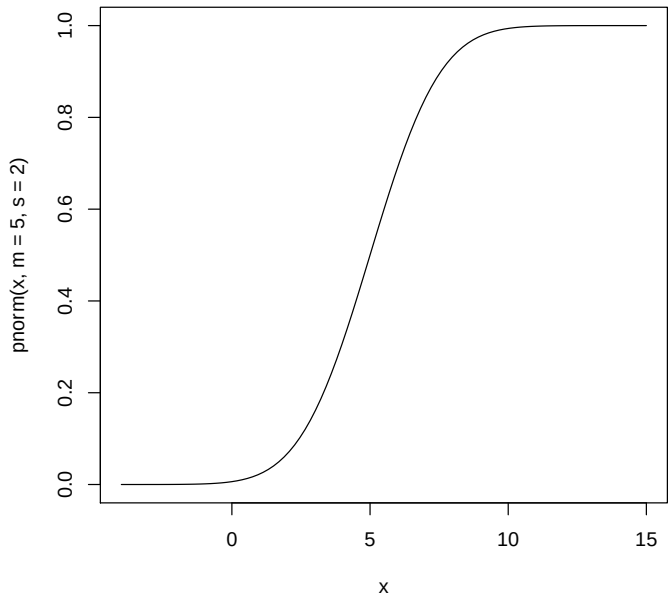
$$\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}} \quad = \quad g\left(p(\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}})\right) \quad = \quad \Phi^{-1}\left(p(\vec{\mathbf{x}}^t\vec{\boldsymbol{\beta}})\right) \ .$$

A Regressão Probit (cont.)

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

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

onde $\beta_0 = -\frac{\mu}{\sigma}$ e $\beta_1 = \frac{1}{\sigma}$, i.e., 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 (cont.)

Em geral, 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,
- com 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, $p(-\vec{\boldsymbol{\eta}}) = 1 - p(\vec{\boldsymbol{\eta}})$, para qualquer $\vec{\boldsymbol{\eta}}$.

Inconvenientes:

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

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 dosagem há um nível de tolerância t: o limiar acima do qual o produto tóxico provoca a morte do indivíduo;
- esse nível de tolerância varia entre indivíduos e pode ser representado por uma v.a. T.

Definindo a v.a. binária Y:

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

A Regressão Probit em toxicologia (cont.)

Tem-se:

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

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.

Os coeficientes verificam $\beta_0 = -\frac{\mu}{\sigma}$ e $\beta_1 = \frac{1}{\sigma}$, estando pois associados aos parâmetros da distribuição de T.

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

Regressão Probit no 

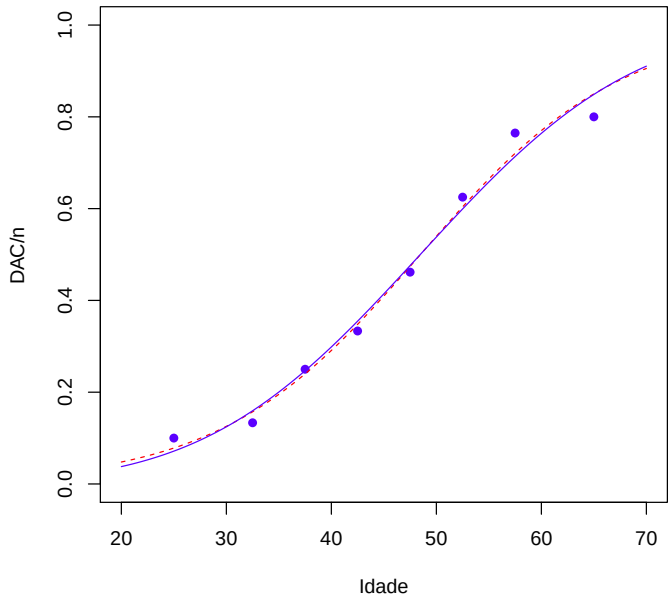
Numa regressão probit, há que especificar a respectiva função de ligação, como opção do argumento `family`, da seguinte forma:

```
> glm(cbind(DAC,n-DAC)~Idade, family=binomial(link=probit), data=HosLem)
Call:  glm(formula = cbind(DAC, n - DAC) ~ Idade,
                family = binomial(link = probit), data = HosLem)
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
```

Tal como no caso da Regressão Logística, a variável resposta pode ser explicitada sob a forma duma matriz de duas colunas, indicando o número de "êxitos" e o número de "fracassos" (como acima) ou, alternativamente, como uma coluna 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=T, col="blue")
```

O 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 ($Y = 1$):

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

A função p é a diferença para 1 duma curva de Gompertz com valor assintótico $\alpha = 1$. O valor assintótico em 1 é natural, uma vez que a função p descreve *probabilidades*.

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

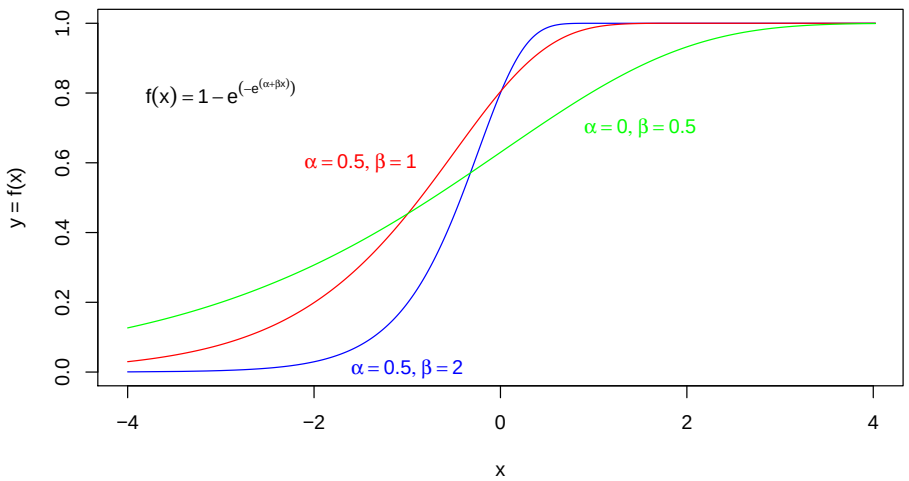
O modelo log-log do complementar (cont.)

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

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

donde a designação do modelo que usa esta função de ligação.

No caso de haver uma única variável preditora X, 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 $\vec{\eta} = 0$;
- mas o valor de probabilidade associado já não se encontra a meio caminho na escala de probabilidades, sendo $p(0) = 1 - \frac{1}{\mathrm{e}}$;
- isso significa que a "fase de aceleração" da curva de probabilidades decorre até um valor superior da probabilidade ($1 - 1/\mathrm{e} \approx 0.632$) 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`:

```
> glm(cbind(DAC,n-DAC)~Idade, family=binomial(link=cloglog), data=HosLem)

Call: glm(formula=cbind(DAC, n-DAC)~Idade, family=binomial(link=cloglog),
                                                        data=HosLem)


Coefficients:
(Intercept)        Idade
   -4.00470      0.07311

Degrees of Freedom: 7 Total (i.e. Null);  6 Residual
Null Deviance:      28.7
Residual Deviance: 1.148  AIC: 26.29
```

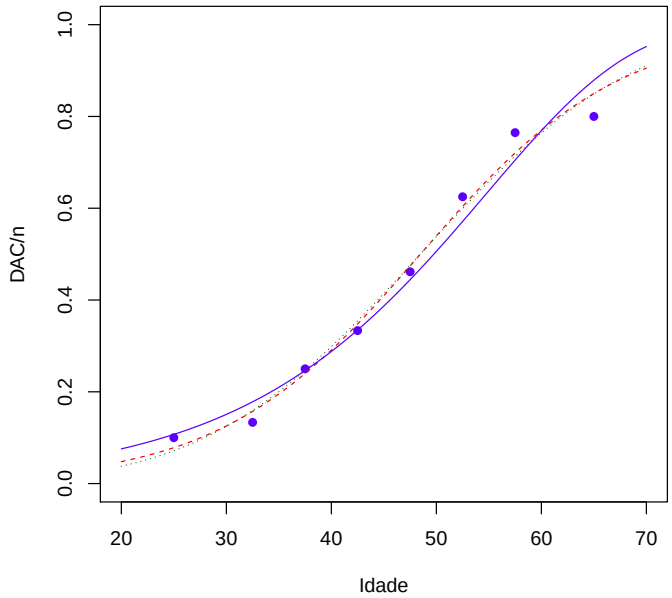
A curva ajustada é:

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

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=T, col="blue")
```

Outras funções de ligação para respostas binárias

Foram consideradas três funções de ligação em modelos de resposta Bernoulli, 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.

Outras funções de ligação (cont.)

Outra possível generalização das funções de ligação para dados binários consiste em considerar a seguinte família de funções de ligação, que depende de um parâmetro, δ:

$$g(p;\delta) \quad = \quad \ln\left[\frac{(1/(1-p))^{\delta}-1}{\delta}\right]$$

A função de ligação logit corresponde a tomar $\delta = 1$.
A função de ligação log-log do complementar corresponde ao limite quando $\delta \to 0$.

Inferência: propriedades dos estimadores MV

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

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

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

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

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

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

Inferência em MLGs

Aplicando estes resultados gerais aos estimadores $\vec{\hat{\boldsymbol{\beta}}}$, obtém-se, assintoticamente;

$$\vec{\hat{\boldsymbol{\beta}}} \sim \mathscr{N}_{(p+1)}(\vec{\boldsymbol{\beta}}, \mathscr{I}_{\beta}^{-1})$$

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

A dimensão da amostra tem uma importância grande para garantir a fiabilidade destes resultados.

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.)

Teorema

Dado um MLG (e admitindo certas condições de regularidade), os estimadores de Máxima Verosimilhança $\vec{\hat{\boldsymbol{\beta}}}$ verificam, assintoticamente:

- *Dado um vector não-aleatório* $\mathbf{a}_{p+1}$*:* $\dfrac{\vec{\mathbf{a}}^t\vec{\hat{\boldsymbol{\beta}}}-\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}}{\sqrt{\vec{\mathbf{a}}^t\mathscr{I}_\beta^{-1}\vec{\mathbf{a}}}} \sim \mathscr{N}(0,1)$.
- *Dada uma matriz não-aleatória* $\mathbf{C}_{q\times(p+1)}$ *de característica q,* $\mathbf{C}\vec{\hat{\boldsymbol{\beta}}} \sim \mathscr{N}_q\left(\mathbf{C}\vec{\boldsymbol{\beta}}, \mathbf{C}\mathscr{I}_\beta^{-1}\mathbf{C}^t\right)$.

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

Inferência em MLGs (cont.)

A derivação de resultados para combinações lineares dos parâmetros inclui como casos particulares importantes, resultados sobre parâmetros individuais e sobre somas ou diferenças de parâmetros.

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:

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

sendo $\mathscr{I}^{-1}_{\hat{\beta}}$ 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:

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

- 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 $\mathscr{I}_{\vec{\boldsymbol{\beta}}}^{-1}$;
- `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 

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

```
> DAC.logistica <- glm(cbind(DAC,n-DAC) ~ Idade,
                       family=binomial, data=HosLem)
> summary(DAC.logistica)

Call:
glm(formula=cbind(DAC, n-DAC) ~ Idade, family=binomial, data=HosLem)
[...]
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 

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

```
> vcov(DAC.logistica)
             (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 $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$:

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

Intervalos de confiança para β_j no 

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

```
> confint.default(DAC.logistica)
                  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(DAC.logistica)
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 surge com muita frequência, associada à contagem de acontecimentos aleatórios (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.

Funções de ligação e ligação canónica

O valor esperado de $Y \cap Po(\lambda)$ é o parâmetro λ.

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

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

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

São modelos com:

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

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 λ 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_j}.

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 da variável preditora 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: Modelos log-lineares

Num Modelo Log-Linear, as n observações independentes são duma variável aleatória com distribuição de Poisson.
A função de verosimilhança destas n observações é dada por:

$$\mathbf{L}(\lambda\ ;\ \vec{\mathbf{y}})\ =\ \prod_{i=1}^{n} \mathrm{e}^{-\lambda_i} \frac{\lambda_i^{y_i}}{y_i!}$$

E a log-verosimilhança por:

$$\mathscr{L}(\lambda\ ;\ \vec{\mathbf{y}})\ =\ \sum_{i=1}^{n} (-\lambda_i + y_i \ln \lambda_i - \ln y_i!)$$

A função de ligação é dada por $g(\lambda) = \ln(\lambda) = \vec{\mathbf{x}}^t \vec{\boldsymbol{\beta}}$. Eis a expressão para a log-verosimilhança como função dos parâmetros $\vec{\boldsymbol{\beta}}$:

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

Estimação em modelos log-lineares (cont.)

Deixando cair a última parcela, que é constante nos parâmetros β_j, logo dispensável na identificação dos máximos:

$$\mathscr{L}(\vec{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \left(-\mathrm{e}^{\sum_{k=0}^{p} x_{k(i)}\beta_k} + y_i \sum_{k=0}^{p} x_{k(i)}\beta_k \right)$$

(com a convenção $x_{0(i)} = 1, \forall\, i$). Condição necessária para a existência de extremo da log-verosimilhança no ponto $\vec{\boldsymbol{\beta}} = \vec{\hat{\boldsymbol{\beta}}}$ é que:

$$\frac{\partial \mathscr{L}(\vec{\hat{\boldsymbol{\beta}}})}{\partial \beta_j} = \sum_{i=1}^{n} x_{j(i)} \left[y_i - \mathrm{e}^{\sum_{k=0}^{p} x_{k(i)}\hat{\beta}_k} \right] = 0 \qquad \forall j = 0 : p$$

[MC:] Estimação em modelos log-lineares (cont.)

Tal como no caso anterior, estas $p+1$ equações formam um sistema não-linear de equações nas $p+1$ incógnitas $\hat{\beta}_j$, $j = 0 : p$.

De novo, embora o sistema de equações seja não linear, é possível utilizar uma notação mnemónica matricial, definindo o vector $\vec{\boldsymbol{\lambda}}$ de probabilidades estimadas, cuja i-ésima componente é dada por:

$$\hat{\lambda}_i \quad = \quad e^{\sum_{k=0}^{p} x_{k(i)} \hat{\beta}_k}$$

Com esta notação, o sistema de $p+1$ equações toma a forma:

$$\mathbf{X}^t \vec{\mathbf{y}} = \mathbf{X}^t \vec{\boldsymbol{\lambda}}$$

A não-linearidade do sistema exige métodos numéricos.

[MC:] IRLS para modelos log-lineares

No contexto do Modelo Log-Linear, as derivadas parciais de primeira ordem da log-verosimilhança são:

$$\frac{\partial \mathscr{L}(\vec{\boldsymbol{\beta}})}{\partial \beta_j} = \sum_{i=1}^{n} x_{j(i)} \underbrace{\left[y_i - \mathrm{e}^{\sum_{k=0}^{p} x_{k(i)} \beta_k} \right]}_{=y_i - \lambda_i}, \qquad \forall j = 0 : p$$

$$\Leftrightarrow \qquad \frac{\partial \mathscr{L}(\vec{\boldsymbol{\beta}})}{\partial \vec{\boldsymbol{\beta}}} = \mathbf{X}^t \vec{\mathbf{y}} - \mathbf{X}^t \vec{\boldsymbol{\lambda}}$$

Assim, as derivadas parciais de segunda ordem são:

$$\frac{\partial^2 \mathscr{L}}{\partial \beta_l \partial \beta_j} \left(\vec{\boldsymbol{\beta}} \right) = - \sum_{i=1}^{n} x_{j(i)} x_{l(i)} \, \mathrm{e}^{\sum_{k=0}^{p} x_{k(i)} \beta_k}$$

$$\frac{\partial^2 \mathscr{L}}{\partial \beta_l \partial \beta_j} \left(\vec{\boldsymbol{\beta}} \right) = - \sum_{i=1}^{n} x_{j(i)} x_{l(i)} \, \lambda_i, \qquad \forall j, l = 0 : p$$

[MC:] IRLS para modelos log-lineares (cont.)

De novo, a função de ligação é canónica e os elementos da Hessiana não dependem de Y, pelo que Hessiana e seu valor esperado são iguais, i.e., os métodos de Newton-Raphson e Fisher coincidem.

Defina-se a matriz $n \times n$ diagonal $\mathbf{W}$, cujos elementos diagonais são dados pelos n valores λ_i. A matriz Hessiana e a correspondente matriz de informação de Fisher podem escrever-se como:

$$\mathscr{H} = -\mathbf{X}^t\mathbf{W}\mathbf{X} \qquad \mathscr{I} = \mathbf{X}^t\mathbf{W}\mathbf{X}$$

A equação que define a iteração dos vectores $\vec{\boldsymbol{\beta}}$ no algoritmo IRLS é:

$$\vec{\boldsymbol{\beta}}^{[i+1]} = \vec{\boldsymbol{\beta}}^{[i]} + \left(\mathbf{X}^t\mathbf{W}^{[i]}\mathbf{X}\right)^{-1}\mathbf{X}^t\left(\vec{\mathbf{y}} - \vec{\boldsymbol{\lambda}}^{[i]}\right)$$

Definindo o vector $\vec{\mathbf{z}}^{[i]} = \mathbf{X}\vec{\boldsymbol{\beta}}^{[i]} + \left(\mathbf{W}^{[i]}\right)^{-1}\left(\vec{\mathbf{y}} - \vec{\boldsymbol{\lambda}}^{[i]}\right)$, tem-se uma expressão de transição idêntica à do Modelo Logit:

$$\vec{\boldsymbol{\beta}}^{[i+1]} = \left(\mathbf{X}^t\mathbf{W}^{[i]}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{W}^{[i]}\vec{\mathbf{z}}^{[i]}$$

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.)

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, $E[\hat{Y}_i] = Y_i$.

Recorde-se que, quer no Modelo Logístico, quer no Modelo Log-Linear, o sistema de equações normais resultante da condição necessária para a existência de máximo da log-verosimilhança toma a forma $\mathbf{X}\vec{\mathbf{y}} = \mathbf{X}\hat{\vec{\boldsymbol{\mu}}}$, onde $\hat{\vec{\boldsymbol{\mu}}}$ indica o vector estimado de $E[Y_i]$ para as n observações (Acetatos 65 e 101).

Num modelo saturado, com tantos parâmetros quantas observações, $\mathbf{X}$ é de tipo $n \times n$ e, em geral, invertível. Nesse caso, $\hat{\vec{\boldsymbol{\mu}}} = \vec{\mathbf{y}}$.

Desvios

Assim, 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.

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

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

Desvios (cont.)

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

Seja $\vec{\hat{\boldsymbol{\beta}}}_M$ o vector estimado dos seus parâmetros e $\mathscr{L}_M(\vec{\hat{\boldsymbol{\beta}}}_M)$ a respectiva log-verosimilhança (máxima).

Considere-se um modelo saturado com n parâmetros. Designe-se por $\mathscr{L}_T(\vec{\hat{\boldsymbol{\beta}}}_T)$ a log-verosimilhança correspondente (isto é, a log-verosimilhança obtida substituindo cada parâmetro estimado $\hat{\mu}_i$ pela observação correspondente y_i).

Define-se o desvio como sendo:

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

Desvios e desvios reduzidos

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

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

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

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

Na expressão do desvio surge o parâmetro de dispersão ϕ.

Desvios e desvios reduzidos

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

- $\phi = 1$ na Poisson;
- $\phi = 1$ na Bernoulli;
- $\phi = \frac{1}{n}$ na Binomial/n.

Mas, 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.

Desvios na Poisson e Binomial/n

Substituindo as expressões já antes obtida na definição geral do Desvio (acetato 107), têm-se as seguintes expressões para MLGs em que Y tem:

- distribuição de Poisson:

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

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

- distribuição Binomial/n:

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

Desvios e desvios reduzidos (cont.)

Para distribuições em que seja necessário estimar ϕ, é hábito definir um conceito alternativo de Desvio. Admitindo que

$$a(\phi_i) = \frac{\phi}{w_i} \ ,$$

para constantes w_i conhecidas e ϕ comum a todas as observações:

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

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

$$D^* = \frac{D}{\phi} \ ,$$

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

NOTA: Na Poisson e Binomial/n, desvio e desvio reduzido coincidem.

Desvio e desvio reduzido na Normal

O desvio reduzido na Normal, (admitindo a variância σ_i^2 de cada observação conhecida e escrevendo $\hat{\mu}_i^M$ apenas como $\hat{\mu}_i$) é:

$$D^* = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\sigma_i^2} .$$

Com a hipótese usual do Modelo Linear de que $\sigma_i^2 = \sigma^2 = \phi$ para todas as observações, o desvio da Normal vem:

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

ou seja, o desvio e a tradicional Soma de Quadrados Residual coincidem.

O AIC em GLMs

O Critério de Informação de Akaike (AIC) define-se, num MLG com $p+1$ parâmetros, como

$$AIC = -2 \cdot \mathscr{L}(\vec{\hat{\boldsymbol{\beta}}}; \vec{\mathbf{Y}}) + 2(p+1) \, .$$

- 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 e submodelos.
- Note-se a relação entre o desvio reduzido D^* dum GLM o seu AIC: ambos definidos à custa da log-verosimilhança.

A razão de verosimilhanças

Um teste à admissibilidade de um Submodelo pode ser obtido com base num resultado mais geral: o teste à razão de Verosimilhanças.

Seja $(Y_1, Y_2, ..., Y_n)$ uma amostra aleatória. Seja $L(\boldsymbol{\theta}|\mathbf{x})$ a sua função verosimilhança, onde $\boldsymbol{\theta}$ 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 os valores admissíveis com a restrição $\vec{\boldsymbol{\beta}}_{\overline{S}} = \vec{\mathbf{0}}$.
Por Θ_1 indica-se a condição complementar: $\vec{\boldsymbol{\beta}}_{\overline{S}} \neq \vec{\mathbf{0}}$.
Por $\Theta_0 \cup \Theta_1$ indica-se qualquer vector $\vec{\boldsymbol{\beta}}_{\overline{S}}$, sem restrições.

Teorema de Wilks

Designa-se razão de verosimilhanças 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:

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

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, a variação de $E[Y]$ não depende de variáveis preditoras.

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

Teste de Wilks ao Ajustamento Global (cont.)

Teste de Wilk ao Ajustamento de um MLG

Hipóteses:

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

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: Exercícios 1 e 10

Do livro de Venables e Ripley. Uma experiência estuda a resistência da larva do tabaco *heliothis virescens* a doses de uma substância tóxica.

Lotes de 20 traças de cada sexo foram expostas, durante 3 dias, a doses da referida substância. Registou-se o número de individuos de cada lote que morria até ao fim desse período de exposição. Os resultados são sintetizados na seguinte tabela (doses em μg).

	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.)

Criação de `data.frame` com os dados:

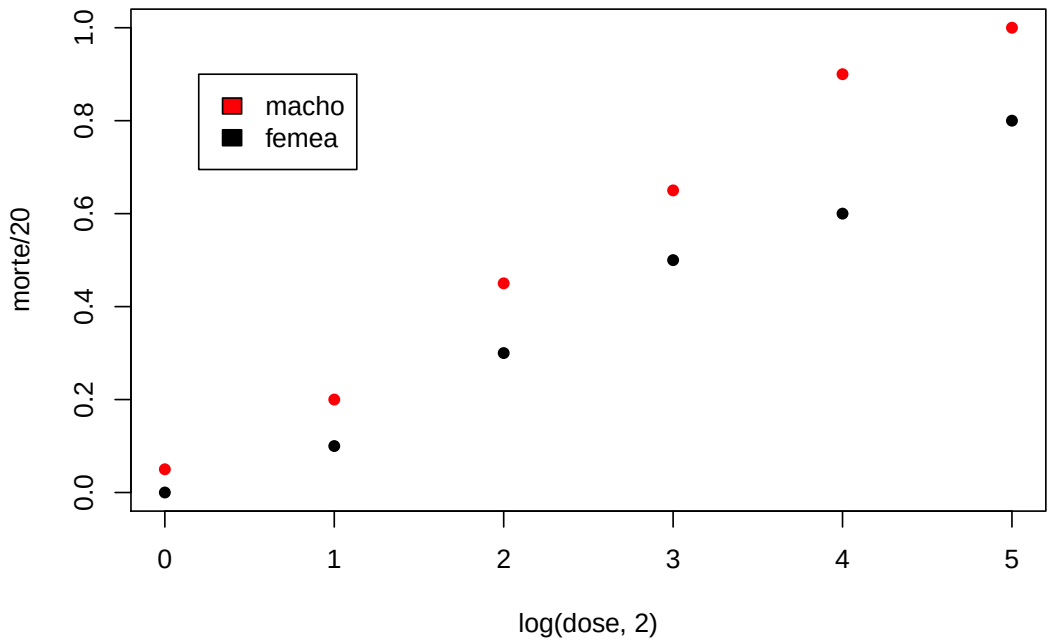
```
> morte <- c(1,4,9,13,18,20,0,2,6,10,12,16)
> sexo <- factor(rep(c("macho","femea"),c(6,6)))
> dose <- rep(2^(0:5),2)
> tabaco <- data.frame(morte,sexo,dose)
> tabaco
```

```
   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

```
> attach(tabaco)
> plot(log(dose,2),morte/20,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 2 no R (cont.)

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

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

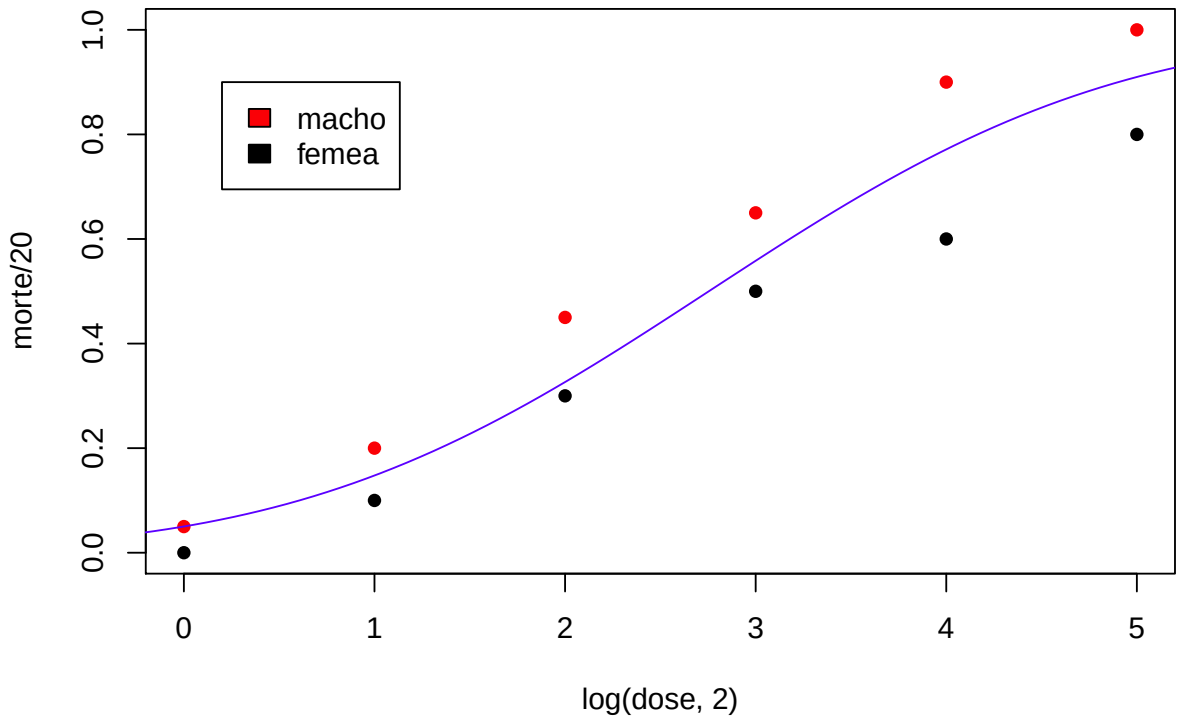
```
Call:  glm(formula = cbind(morte, 20 - morte) ~ log(dose, 2),
+        family = binomial(link = "probit"), data = tabaco)

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.

Exercício 1 no R (cont.)

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)
```



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''`.

```
> 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, sem preditores.

Exercício 10

Também é possível cruzar factores com preditores numéricos, como numa Análise de Covariância.

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

Call: glm(formula = cbind(morte, 20 - morte) ~ log(dose, 2) * sexo,
    family = binomial(link = "probit"), data = tabaco)
(...)
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
AIC: 41.878
```

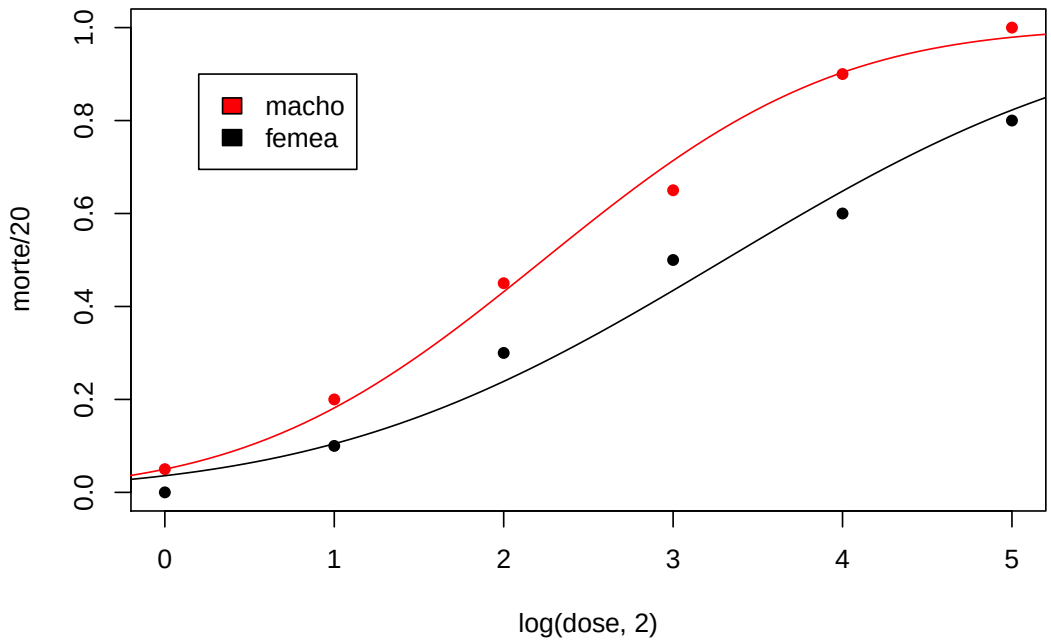
As relações estimadas são:

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

Repare-se como o desvio baixou de 16.41 para apenas 3.768.

Exercício 10 no R (cont.)

```
> plot(morte/20 ~ log(dose,2),col=sexo, data=tabaco,pch=16)
> curve(pnorm(-1.80072+0.54523*x), from=-1, to=6, col="black",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"))
```



Exercícios 1 e 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.

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

Há vantagens evidentes na distinção por sexo (previsível pelo ajustamento gráfico).

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.

Algoritmos sequenciais no 

No  ,

- o comando `anova` fornece a informação básica para efectuar um Teste de razão de verosimilhanças a Submodelos encaixados (indicando os submodelos como argumentos do comando); e
- os comandos `drop1` e `add1` fornecem a informação básica para proceder aos algoritmos de exclusão/inclusão sequenciais de variáveis preditoras, na escolha de Submodelos.
- o comando `step` automatiza os algoritmos de selecção sequencial com base no AIC. É respeitada a natureza dos preditores categóricos e a hierarquia dos tipos de efeitos que lhe estão associados.

Algoritmos de selecção de preditores no R

Para ilustrar a aplicação do algoritmo de exclusão sequencial, vejamos o exemplo já considerado, associado ao Exercício 10:

```
> step(tabaco.glmS)
Start:  AIC=41.88
cbind(morte, 20 - morte) ~ log(dose, 2) * sexo
                     Df Deviance    AIC
- log(dose, 2):sexo   1    5.566 41.676
<none>                     3.768 41.878

Step:  AIC=41.68
cbind(morte, 20 - morte) ~ log(dose, 2) + sexo
               Df Deviance     AIC
<none>              5.566  41.676
- 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.

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

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

que é da familia exponencial com:

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

A família das distribuições Gama inclui como caso particular a distribuição Qui-quadrado (χ^2_n se $\nu = \frac{n}{2}$ e $\mu = n$) e a distribuição Exponencial ($\nu = 1$).

Modelos com variável resposta Gama

Vejamos agora um exemplo de MLG com variável resposta contínua, não Normal. Consideremos uma componente aleatória Y com distribuição Gama (que, como sabemos, inclui como casos particulares uma Exponencial ou uma Qui-quadrado).

Se $Y \cap G(\mu, \nu)$, tem-se:

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

Assim, na distribuição Gama a variância é proporcional ao quadrado da média. Esta propriedade sugere que MLGs com componente aleatória Gama podem ser úteis em situações onde a variância dos dados não seja constante, mas proporcional ao quadrado da média.

Funções de ligação e ligação canónica na Gama

Uma vez que para $Y \cap G(\mu, \nu)$ 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:

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

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

Como o sinal negativo não é relevante na discussão, é hábito definir a função de ligação canónica para modelos com variável resposta Gama apenas como a função recíproco:

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

Um único preditor na Gama

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

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

No caso particular de haver uma única variável preditora, a relação que acabámos de estabelecer 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).

Um preditor transformado

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).

Nota: embora o valor esperado da variável resposta Y tenha de ser positivo (uma vez que uma variável Y com distribuição Gama só toma valores positivos), na relação estabelecida o valor esperado pode ser negativo para alguns valores da(s) variável(is) preditora(s).

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

Desvio e desvio reduzido na Gama

Tem-se, a partir das expressões para D^* do Acetato 107 e para D do Acetato 110, e tendo em conta que $\theta = \frac{1}{\mu}$, $b(\theta) = -\ln(-\theta) = \ln(\mu)$, $\phi = \frac{1}{\nu}$ e $a(\phi) = \phi = \frac{1}{\nu}$:

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

Admitindo que $a(\phi_i) = \frac{\phi}{w_i}$, para algum conjunto de constantes w_i, o desvio não vem muito diferente (apenas substituindo ν_i por w_i).

Com a hipótese da igualdade de parâmetros de dispersão nas n observações, fica-se com uma expressão mais simples para o desvio:

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

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 $a(\phi) = \phi$.

Resíduos e Validação do Modelo

O conceito usual de resíduos, $e_i = y_i - \hat{y}_i$, usado no Modelo Linear como ferramenta para a validação das hipóteses subjacentes ao Modelo, 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.

Resíduos de Pearson

Como base da ideia de resíduos de Pearson está a comparação "normalizada" entre valores observados de Y_i e correspondentes estimativas dos seus valores esperados, $\widehat{E[Y_i]} = \hat{\mu}_i$.

Resíduos de Pearson

Seja $Y_1, Y_2, ..., Y_n$ uma amostra aleatória de uma Componente Aleatória dum Modelo Linear Generalizado. Designa-se resíduos de Pearson de cada observação às quantidades:

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

sendo $f_v(\mu) = \frac{V[Y_i]}{\phi_i}$ a chamada função de variância ($V[Y] = f_v(\mu)\,\phi$).

À soma de quadrados destes resíduos dá-se o nome de estatística de Pearson generalizada, X^2.

Resíduos de Pearson (cont.)

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

- 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:

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

- 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:

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

Resíduos de Pearson (cont.)

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

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

- 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.)

Mas 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 140 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:

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

Estimação do parâmetro de dispersão ϕ

Os resíduos de Pearson, em modelos com parâmetro de dispersão ϕ desconhecido, são uma das formas usada para estimar ϕ.

Admitindo ϕ comum a todas as observações Y_i, toma-se:

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

onde m indica o número de parâmetros do modelo. No caso particular do Modelo Linear esta estimativa reduz-se ao *QMRE*.

Nota: Em modelos de resposta Binomial ou Poisson, $\phi = 1$ (estimação desnecessária). Mas um valor de $\hat{\phi}$ muito superior a 1 sugere a existência de sobredispersão, aconselhando modificações ao modelo.

Nota: Existem outras formas de estimar ϕ, nomeadamente por máxima verosimilhança, com resultados diferentes.

Resíduos do Desvio

Um conceito alternativo de resíduo baseia-se na analogia entre o Desvio no estudo dum MLG, e da Soma de Quadrados dos Resíduos no Modelo Linear.

Resíduos do Desvio

Seja $Y_1, Y_2, ..., Y_n$ uma amostra aleatória de uma 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:

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

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 no Modelo Linear são os resíduos usuais.

- 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 estandardizados

Tal como no estudo do modelo linear, é usual definir normalizações dos resíduos.

No modelo linear, os resíduos estandardizados definem-se como:

$$r_i = \frac{e_i}{\sqrt{QMRE\,(1-h_{ii})}},$$

onde

- $QMRE$ estima a variância σ^2 dos erros aleatórios; e
- $h_{i,i}$ é a *leverage* (efeito alavanca) da i-ésima observação, dada pelo i-ésimo elemento diagonal da matriz de projecção ortogonal, $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$.

Os Resíduos estandardizados

Nos MLGs, define-se um conceito análogo, com as devidas substituições:

- em vez de *QMRE*, a estimativa do parâmetro de dispersão, $\hat{\phi}$.
- em vez da matriz $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$, a matriz

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W}^{1/2} \ ,$$

sendo $\mathbf{W}$ a matriz referida aquando da discussão do Método de Fisher (Acetatos 63 e seguintes).

Resíduos estandardizados de Pearson e do Desvio

Os resíduos de Pearson estandardizados definem-se como:

$$r_i^{P'} = \frac{r_i^P}{\sqrt{\hat{\phi}\,(1-h_{ii})}} = \frac{(Y_i - \hat{\mu}_i)}{\sqrt{\hat{\phi}\,(1-h_{ii})\,f_v(\hat{\mu}_i)}}$$

Os resíduos do Desvio estandardizados definem-se como:

$$r_i^{D'} = \frac{r_i^D}{\sqrt{\hat{\phi}\,(1-h_{ii})}}$$

Também se podem definir resíduos studentizados, resultantes de estimativas de ϕ obtidas sem a i-ésima observação, embora sejam computacionalmente pesadas.

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(toxico.glm)
         1          2          3          4          5          6          7
-0.1785893 -1.5621554  0.6421616 -0.6229189  1.2091715 -0.1955369  0.7903046
         8          9         10         11         12
-0.5679427  1.4211256 -1.7989247  1.9850357 -1.4380000
```

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

```
> residuals(toxico.glm, type="pearson")
         1          2          3          4          5          6          7
-0.1740663 -1.1216744  0.6710920 -0.5922706  1.2433189 -0.1944047  0.7822631
         8          9         10         11         12
-0.5702404  1.3022474 -1.9324076  1.4389199 -1.6287271
```

Os Resíduos no R (cont.)

- Os resíduos estandardizados do desvio podem ser obtidos através do comando `rstandard`:

```
> rstandard(toxico.glm)
         1          2          3          4          5          6          7
-0.1940009 -1.6969641  0.7077878 -0.6865786  1.3122925 -0.2122128  0.8543712
         8          9         10         11         12
-0.6139834  1.5767825 -1.9959623  2.2022768 -1.5953739
```

- Resíduos externamente estandardizados obtêm-se através do comando `rstudent`:

```
> rstudent(toxico.glm)
         1          2          3          4          5          6          7
-0.1932594 -1.6330486  0.7135305 -0.6807268  1.3179548 -0.2120277  0.8531218
         8          9         10         11         12
-0.6143426  1.5528805 -2.0245860  2.1019875 -1.6371460
```

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. Sugere-se a consulta de

- McCullagh & Nelder (1989);
- Turkman & Silva (2000)

para uma discussão mais aprofundada.

O estudo dos resíduos (cont.)

São frequentes as seguintes inspecções gráficas:

1) Resíduos contra transformações das esperanças estimadas: é o gráfico correspondente ao gráfico de resíduos *vs.* valores ajustados no Modelo Linear. Em MLGs, estas transformações diferem consoante a distribuição dos Y_i, na tentativa de fazer com que os gráficos tenham uma leitura semelhante à que se fazia no Modelo Linear.

As transformações sugeridas por McCullagh & Nelder (1989) são:

- $\hat{\mu}$ para Y Normal de média μ;
- $2\sqrt{\hat{\lambda}}$ para Y Poisson de média λ;
- $2\arcsin(\hat{p})$ para Y Bernoulli de média p.
- $2\ln\hat{\mu}$ para Y Gama de média μ.

O estudo dos resíduos (cont.)

Num bom ajustamento do Modelo Linear Generalizado, os resíduos devem dispersar-se em torno de zero, sem ordem aparente, e dentro de uma banda horizontal de amplitude constante.

Curvaturas em gráficos deste tipo sugerem a possibilidade de escolha errada de função de ligação ou a necessidade de transformação de uma ou mais variáveis preditoras.

McCullagh & Nelder sugerem a utilização dos resíduos do desvio estandardizados neste tipo de gráficos.

O estudo dos resíduos (cont.)

2) resíduos contra cada variável preditora: trata-se dum tipo de gráfico que também pode ser usado nos Modelos Lineares, para sugerir transformações de algum(ns) preditor(es).

A sua utilidade é tanto maior quanto menor fôr o número de variáveis preditoras.

Um padrão evidente neste gráfico indicia ou uma função de ligação errada, ou a necessidade duma transformação do preditor.

O estudo dos resíduos (cont.)

3) resíduos contra ordem de observação: caso faça sentido, este tipo de gráfico pode indicar a presença de correlação entre observações que se desejam independentes.

4) módulo dos resíduos contra os valores ajustados de $\hat{\mu}$: é útil para estudar se a função de variância admitida é plausível, em cujo caso os pontos devem dispersar-se numa banda horizontal.

Observações influentes

No modelo linear, o conceito de influência indica uma observação cuja exclusão do conjunto de dados conduziria a alterações importantes nos valores ajustados. A forma usual de medir a influência de observações no modelo linear é através da distância de Cook.

Em MLGs, um conceito análogo resulta de considerar, para a observação i a seguinte analogia com a distância de Cook:

$$D_i = \frac{(\vec{\hat{\boldsymbol{\beta}}}_{[-i]} - \vec{\hat{\boldsymbol{\beta}}})^t (\mathbf{X}^t \mathbf{W} \mathbf{X}) (\vec{\hat{\boldsymbol{\beta}}}_{[-i]} - \vec{\hat{\boldsymbol{\beta}}})}{(p+1)\,\hat{\phi}},$$

onde $\vec{\hat{\boldsymbol{\beta}}}_{[-i]}$ indica o vector de estimativas dos parâmetros que resultaria de omitir a i-ésima observação e sendo $\mathbf{W}$ a matriz referida na discussão do Método de Fisher (Acetatos 63 e seguintes).

No R estas quantidades obtêm-se pelo comando `cooks.distance`.

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, e merecem uma referência especial.

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

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	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot(b-1)}$	$n_{\cdot b}$	$n = n_{\cdot\cdot}$

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

A hipótese de independência

Mais útil serão modelos associados a hipóteses mais específicas 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.

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 $E[Y_{ij}]$:

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

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.

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, podemos 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.

As restrições aos parâmetros

Consideramos

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

Logaritmizando, temos as relações

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

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

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 serão dados de forma directa pelas frequências relativas marginais:

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

pelo que

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

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.

Exemplo 3: modelo para tabela de contingência

Na *data frame* `cabelo.olho` encontram-se dados relativos a $n = 16$ 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
    castanho     68       119      26      7
      azuis      20        84      17     94
   cinzentos     15        54      14     10
     verdes       5        29      14     16
```

Exemplo 3 (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 $\frac{N}{n} = \frac{592}{16} = 37$.

Exemplo 3 (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: 1.0436e-14  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 ***
```

Exemplo 3 (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
```

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.

[MC]: Tabelas de contingência com três factores

Vejamos agora o contexto de tabelas de contingência com três factores de classificação:

- um factor A com a níveis,
- um factor B com b níveis, e
- um factor C com c níveis.

Os dados são contagens n_{ijk} do número de observações na célula (i,j,k) ($i = 1:a$, $j = 1:b$ e $k = 1:c$).

Uma tabela deste tipo corresponde a uma matriz tri-dimensional.

[MC]: Tabelas de contingência com três factores (cont.)

Admita-se que as contagens em cada célula duma tabela com três factores de classificação são observações independentes com distribuição de Poisson, de parâmetros λ_{ijk}.

O modelo mais geral é um modelo log-linear do tipo ANOVA factorial, a 3 factores, com todas as possíveis interacções (tripla e os três tipos de interacção dupla):

$$log(E[Y_{ijk}]) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} \ .$$

O modelo tem *abc* parâmetros, e neste contexto é saturado.

De novo, os modelos úteis correspondem a modelos com algum tipo de estrutura associada à tabela.

[MC]: Conceitos de independência com 3 factores

Consideremos agora vários conceitos de independência relacionados com três factores A, B e C.

Sejam

- A, B e C três factores com, respectivamente, a, b e c níveis;
- p_{ijk} a probabilidade duma observação pertencer ao nível i do factor A, j do factor B e k do factor C;
- $p_{ij.}$ a probabilidade (marginal) de uma observação recair no nível i do factor A e j do factor B, qualquer que seja o nível do factor C associado. Sejam $p_{i.k}$ e $p_{.jk}$ probabilidades definidas de forma análoga.
- $p_{i..}$ a probabilidade (marginal) da observação recair no nível i do factor A, qualquer que sejam os níveis dos outros dois factores. Sejam $p_{.j.}$ e $p_{..k}$ as probabilidades marginais análogas para B e C.

[MC]: Conceitos de independência (cont.)

1. diz-se que A, B e C são mutuamente independentes se

$$p_{ijk} = p_{i..} \cdot p_{.j.} \cdot p_{..k} \quad \forall\, i,j,k\ ;$$

2. diz-se que A é conjuntamente independente de B e C se

$$p_{ijk} = p_{i..} \cdot p_{.jk} \quad \forall\, i,j,k$$

(definições análogas para os outros casos análogos);

3. diz-se que A e B são condicionalmente independentes de C se

$$p_{ij|k} = p_{i.|k} \cdot p_{.j|k} \quad \forall\, i,j,k$$

(definições análogas para os outros casos análogos);

4. diz-se que A e B são marginalmente independentes se

$$p_{ij.} = p_{i..} \cdot p_{.j.} \quad \forall\, i,j$$

(definições análogas para os outros casos análogos);

[MC]: Conceitos de independência (cont.)

5. diz-se que A, B e C são independentes se forem
   - mutuamente independentes e
   - os três pares (A,B), (A,C) e (B,C) forem marginalmente independentes.

Existem relações de implicação entre vários destes tipos de independência.

É imediato a partir da definição que a independência implica a independência mútua e ainda a independência marginal de qualquer dos possíveis pares de factores.

[MC]: Relações de conceitos de independência

1. Se A, B e C são factores mutuamente independentes, cada factor é conjuntamente independente dos outros dois. I.e., (isolando C):

$$p_{ijk} = p_{i..} \cdot p_{.j.} \cdot p_{..k} \quad \Longrightarrow \quad p_{ijk} = p_{ij.} \cdot p_{..k} \ ,$$

Dem.: Basta mostrar que A e B são marginalmente independentes ($p_{ij.} = p_{i..} \cdot p_{.j.}$). Ora, se A, B e C são mutuamente independentes,

$$p_{ij.} = \sum_{k=1}^{c} p_{ijk} = \sum_{k=1}^{c} p_{i..} \cdot p_{.j.} \cdot p_{..k} = p_{i..} \cdot p_{.j.} \sum_{k=1}^{c} p_{..k} = p_{i..} \cdot p_{.j.} \ ,$$

uma vez que necessariamente $\sum_{k=1}^{c} p_{..k} = 1$. A demonstração é análoga para qualquer outra das independências conjuntas.

[MC]: Relações de conceitos de independência (cont.)

2. Se A é conjuntamente independente de (B,C), então
   - (A,B) são condicionalmente independentes de C; e
   - (A,C) são condicionalmente independentes de B.

   Ou seja,

$$p_{ijk} = p_{i..} \cdot p_{.jk} \quad \Longrightarrow \quad \begin{cases} p_{ij.|k} = p_{i..|k} \cdot p_{.j.|k} \\ p_{i.k|j} = p_{i..|j} \cdot p_{..k|j} \end{cases}$$

Dem: Tem-se (no primeiro caso),

$$p_{ij|k} = \frac{p_{ijk}}{p_{..k}} = \frac{p_{i..} \cdot p_{.jk}}{p_{..k}} = p_{i..} \cdot p_{.j|k} \ ,$$

donde, somando ao longo do índice j se tem

$$p_{i.|k} = \sum_{j=1} p_{ij|k} = p_{i..} \cdot \sum_{j=1}^{b} \frac{p_{.jk}}{p_{..k}} = p_{i..} \ .$$

Substituindo a expressão para $p_{i..}$, obtem-se o resultado desejado:

$$p_{ij|k} = p_{i.|k} \cdot p_{.j|k} \ .$$

[MC]: Relações de conceitos de independência (cont.)

3. A independência conjunta de (A,B) com C implica
   - a independência marginal de A e C; e
   - a independência marginal de B e C.

   ou seja,

$$p_{ijk} = p_{ij.} \cdot p_{..k} \quad \Longrightarrow \quad \begin{cases} p_{i.k} = p_{i..} \cdot p_{..k} & , \forall\, i,j,k\,. \\ p_{.jk} = p_{.j.} \cdot p_{..k} & , \forall\, i,j,k\,. \end{cases}$$

Dem.: O resultado é evidente somando (no primeiro caso) a equação inicial em j:

$$p_{i.k} = \sum_{j=1}^{b} p_{ijk} = \sum_{j=1}^{b} p_{ij.} \cdot p_{..k} = p_{i..} \cdot p_{..k}\,.$$

A independência marginal de B e C sai de forma análoga.

[MC]: Notas

1. Como já se tinha mostrado que a independência mútua dos três factores implica a independência conjunta de, digamos, (A,B) com C, o último ponto do Teorema anterior mostra que a independência mútua dos três factores implica a independência marginal de qualquer par desses factores.
2. É possível exemplificar que a independência marginal de, digamos, A e C não é implicada pela independência condicional de (A,B) face a C.
3. A independência condicional pode escrever-se apenas à custa de probabilidades marginais. De facto, a partir da definição de independência condicional tem-se a seguinte expressão alternativa para a definição de A e B serem independentes condicionalmente a C:

$$p_{ij|k} = \frac{p_{i.k} \cdot p_{.jk}}{p_{..k}}$$

[MC]: Modelo para a independência mútua

Vejamos como, associados a cada uma destes tipos de independência, se pode definir um modelo log-linear adequado, de tal forma que às implicações referidas correspondam submodelos encaixados.

Por analogia com o caso a dois factores, a independência mútua dos três factores significa que o valor esperado do número de observações na célula (i,j,k) é dado por

$$E[Y_{ijk}] \quad = \quad n \cdot p_{ijk} \quad = \quad n \cdot p_{i..} \cdot p_{.j.} \cdot p_{..k} \ .$$

Logaritmizando, tem-se

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

que é uma equação do tipo de um modelo ANOVA para três factores, sem qualquer tipo de interacção:

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

[MC]: Modelo para a independência mútua (cont.)

Tendo mais uma vez em conta a necessidade de evitar dependências lineares nas colunas da matriz do delineamento, já estudados em Modelação Estatística I, iremos re-escrever a equação base da relação sob a forma

$$\begin{aligned} \lambda_{111} &= E[Y_{111}] &&= n \cdot p_{1..} \cdot p_{.1.} \cdot p_{..1} \\ \lambda_{ijk} &= E[Y_{ijk}] &&= n \cdot p_{i..} \cdot p_{.j.} \cdot p_{..k} \\ & &&= \lambda_{111} \cdot \frac{p_{i..}}{p_{1..}} \cdot \frac{p_{.j.}}{p_{.1.}} \cdot \frac{p_{..k}}{p_{..1}} \quad , \quad \forall\, i = 2:a\,,\ j = 2:b\,,\ k = 2:c \end{aligned}$$

Logaritmizando, temos as relações

$$\begin{aligned} \ln(\lambda_{111}) &= \ln(E[Y_{111}]) &&= \ln(n) + \ln(p_{1..}) + \ln(p_{.1.}) + \ln(p_{..1}) \\ \ln(\lambda_{ijk}) &= \ln(E[Y_{ijk}]) &&= \ln(\lambda_{111}) + \ln\left(\frac{p_{i..}}{p_{1..}}\right) + \ln\left(\frac{p_{.j.}}{p_{.1.}}\right) + \ln\left(\frac{p_{..k}}{p_{..1}}\right) \\ & &&\quad , \quad \forall\, i = 2:a\,,\ j = 2:b\,,\ k = 2:c \end{aligned}$$

[MC]: Modelo para a independência mútua (cont.)

Assim, o modelo associado à independência mútua dos três factores é um modelo tipo ANOVA a 3 factores, sem qualquer tipo de interacção,

$$\ln(\lambda_{ijk}) \quad = \quad \mu + \alpha_i + \beta_j + \gamma_k \qquad , \qquad i = 2:a \ , \ j = 2:b \ , \ k = 2:c \ ,$$

onde

$$\begin{array}{rcccccl}
\mu & = & \ln(\lambda_{111}) & \Longleftrightarrow & e^{\mu} & = & \lambda_{111} = n \cdot p_{1..} \cdot p_{.1.} \cdot p_{..1} \\
\alpha_i & = & \ln\left(\frac{p_{i..}}{p_{1..}}\right) & \Longleftrightarrow & e^{\alpha_i} & = & \frac{p_{i..}}{p_{1..}} \qquad (i = 2:a) \\
\beta_j & = & \ln\left(\frac{p_{.j.}}{p_{.1.}}\right) & \Longleftrightarrow & e^{\beta_j} & = & \frac{p_{.j.}}{p_{.1.}} \qquad (j = 2:b) \\
\gamma_k & = & \ln\left(\frac{p_{..k}}{p_{..1}}\right) & \Longleftrightarrow & e^{\gamma_k} & = & \frac{p_{..k}}{p_{..1}} \qquad (k = 2:c) \ .
\end{array}$$

[MC]: Modelo para a independência mútua (cont.)

Neste modelo, os três tipos de efeitos, α_i, β_j e γ_k são log-razões de probabilidades. Uma "transição" de uma observação do primeiro nível de referência do factor A para o nível *i* desse mesmo factor corresponde (mantendo o resto igual) a multiplicar por e^{α_i} o valor esperado da contagem de célula .

Os estimadores de máxima verosimilhança de cada um destes efeitos resultam de substituir cada uma das probabilidades marginais pela frequência relativa correspondente. Por exemplo, para qualquer *i*, a probabilidade marginal $p_{i..}$ é estimada por $\hat{p}_{i..} = \frac{n_{i..}}{n_{...}}$.

O modelo log-linear para a independência mútua dos factores A,B e C pode ser representado, de forma mnemónica, como (A,B,C), indicando a existência de apenas três efeitos principais dos níveis de cada factor.

[MC]: Modelos para independências conjuntas

Como vimos na definição do conceito, a independência conjunta de, digamos, o factor A face ao par (B,C) significa que $p_{ijk} = p_{i..} \cdot p_{.jk}$, para qualquer i, j, k.

Logo, o número esperado de observações na célula (i, j, k) é

$$\lambda_{ijk} = E[Y_{ijk}] = n \cdot p_{ijk} = n \cdot p_{i..} \cdot p_{.jk} \ .$$

Para modelar esta relação, iremos admitir, para o logaritmo deste valor esperado, um modelo tipo ANOVA com:

- uma parcela comum a todas as observações;
- parcelas de efeitos principais de cada factor; e
- parcelas de interacção entre os factores B e C.

$$\ln\left(\lambda_{ijk}\right) \quad = \quad \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} \ .$$

[MC]: Modelos para independências conjuntas (cont.)

Este tipo de modelo justifica-se porque:

- Em relação ao modelo saturado, que admite todos os tipos de efeitos, a independência conjunta de (B,C) com A significa que:
  - não é necessária a tripla interacção;
  - tendo em conta que a independência conjunta implica a independência marginal, quer de A e B, quer de A e C, também as parcelas das duplas interacções referidas são dispensáveis.
- um modelo para a relação B-C sem qualquer tipo especial de estrutura seria um modelo com parcelas de efeitos principais dos factores B e C e ainda de interacção B-C. Falta ainda cobrir os efeitos do factor A, $p_{i..}$, tornando-se assim necessário acrescentar parcelas de efeitos principais do factor A.

Modelos em que, havendo efeitos de interacção, há efeitos dos factores individuais envolvidos nessas interacções chamam-se modelos hierarquizados.

[MC]: Modelos para independências conjuntas (cont.)

Pode construir-se o modelo a partir da ideia-base que

$$\lambda_{ijk} = E[Y_{ijk}] = n p_{i..} p_{.jk} \ .$$

Considerando a célula de cruzamento dos níveis $i = j = k = 1$ como célula de referência, tem-se:

$$\lambda_{111} = E[Y_{111}] = n p_{111} = n p_{1..} p_{.11} \iff \ln(\lambda_{111}) = \ln(n p_{1..} p_{.11}) = \mu$$

Agora, consideremos as células em que a esta parcela se acrescenta apenas um dos efeitos principais do factor A, ou seja, uma célula em que $j = k = 1$, mas $i > 1$. Teremos então, para $i = 2 : a$,

$$\lambda_{i11} = E[Y_{i11}] = n p_{i11} = n p_{i..} p_{.11} = (n p_{1..} p_{.11}) \frac{p_{i..}}{p_{1..}}$$

$$\iff \ln(\lambda_{i11}) = \ln(n p_{1..} p_{.11}) + \ln\left(\frac{p_{i..}}{p_{1..}}\right) = \mu + \underbrace{\ln\left(\frac{p_{i..}}{p_{1..}}\right)}_{= \alpha_i}$$

[MC]: Modelos para independências conjuntas (cont.)

Para obter as parcelas do tipo β_j, efeitos principais do factor B, considerem-se as parcelas associadas a células com $i = k = 1$, mas $j > 1$. Teremos então, para $j = 2 : b$,

$$\lambda_{1j1} = E[Y_{1j1}] = np_{1j1} = np_{1..}p_{.j1} = (np_{1..}p_{.j1})\frac{p_{.j1}}{p_{.11}}$$

$$\iff \ln(\lambda_{1j1}) = \ln(np_{1..}p_{.11}) + \ln\left(\frac{p_{.j1}}{p_{.11}}\right) = \mu + \underbrace{\ln\left(\frac{p_{.j1}}{p_{.11}}\right)}_{=\beta_j}$$

Consideremos ainda as células em que a μ apenas se acrescenta um dos efeitos principais do factor C, ou seja, uma célula em que $i = j = 1$, mas $k > 1$. Teremos então, para $k = 2 : c$,

$$\lambda_{11k} = E[Y_{11k}] = np_{11k} = np_{1..}p_{.1k} = (np_{1..}p_{.11})\frac{p_{.1k}}{p_{.11}}$$

$$\iff \ln(\lambda_{11k}) = \ln(np_{1..}p_{.11}) + \ln\left(\frac{p_{.1k}}{p_{.11}}\right) = \mu + \underbrace{\ln\left(\frac{p_{.1k}}{p_{.11}}\right)}_{=\gamma_k}$$

[MC]: Modelos para independências conjuntas (cont.)

Falta apenas obter as parcelas de interacção B-C, $(\beta\gamma)_{jk}$.

Consideremos uma célula em que $i = 1$, mas $j, k \neq 1$: Nesse caso, temos, para $j = 2 : b$ e $k = 2 : c$,

$$\begin{aligned}
\lambda_{1jk} = E[Y_{1jk}] &= np_{1jk} = np_{1..}p_{.jk} = (np_{1..}p_{.11})\,\frac{p_{.j1}}{p_{.11}}\,\frac{p_{.1k}}{p_{.11}}\,\frac{p_{.jk}p_{.11}}{p_{.j1}p_{.1k}} \\
\iff \quad \ln(\lambda_{11k}) &= \ln(np_{1..}p_{.11}) + \ln\left(\frac{p_{.j1}}{p_{.11}}\right) + \ln\left(\frac{p_{.1k}}{p_{.11}}\right) + \ln\left(\frac{p_{.jk}p_{.11}}{p_{.j1}p_{.1k}}\right) \\
&= \mu + \beta_j + \gamma_k + \underbrace{\ln\left(\frac{p_{.jk}\cdot p_{.11}}{p_{.j1}\cdot p_{.1k}}\right)}_{=(\beta\gamma)_{jk}}
\end{aligned}$$

Os valores esperados do número de observações em outras células, $\lambda_{ijk} = E[Y_{ijk}]$, obtêm-se somando as correspondentes parcelas do tipo já referido.

[MC]: Modelos para independências conjuntas (cont.)

É este o modelo associado à independência conjunta de (B,C) com A:

$$\ln(\lambda_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} \ , \qquad i = 2:a,\ j = 2:b,\ k = 2:c \ .$$

sendo

$$\begin{aligned}
\mu &= \ln(n \cdot p_{1..} \cdot p_{.11}) \\
\alpha_i &= \ln\left(\frac{p_{i..}}{p_{1..}}\right) \qquad \forall i = 2:a \\
\beta_j &= \ln\left(\frac{p_{.j1}}{p_{.11}}\right) \qquad \forall j = 2:b \\
\gamma_k &= \ln\left(\frac{p_{.1k}}{p_{.11}}\right) \qquad \forall k = 2:c \\
(\beta\gamma)_{jk} &= \ln\left(\frac{p_{.jk} \cdot p_{.11}}{p_{.j1} \cdot p_{.1k}}\right) \qquad \forall j = 2:b\ , k = 2:c
\end{aligned}$$

Os restantes modelos de independência conjunta – de B face a (A,C) ou C face a (A,B) – são análogos, trocando o papel de cada factor.

[MC]: Modelos para independências conjuntas (cont.)

Para este tipo de modelos, os efeitos principais de cada factor mantêm a sua natureza de log-razões de probabilidades, embora a interpretação do efeito de interacção, $(\beta\gamma)_{jk}$ seja mais complexa.

As estimativas de máxima verosimilhança são as que se obtêm substituíndo cada probabilidade p pela respectiva estimativa $\hat{p}$ resultante de tomar a proporção de observações na célula ou margem correspondente. Assim, por exemplo, $\hat{p}_{.jk} = \frac{n_{.jk}}{n_{...}}$.

[MC]: Testes a tipos de independência

Como se viu anteriormente, a independência mútua implica a independência conjunta de cada factor com o par restante (embora a implicação inversa não seja verdadeira).

Tendo em conta a relação dos modelos acima expostos com as hipóteses de independência mútua e independência conjunta de A com (B,C), poderemos testar estas hipóteses, em alternativa, verificando se os correspondentes modelos encaixados diferem significativamente, para o que podemos utilizar a teoria geral dos MLGs anteriormente estudada.

[MC]: Testes a tipos de independência (cont.)

Ou seja, podemos comparar o desvio do modelo de independência conjunta de (B,C) com A (acetato 191):

$$\ln\left(\lambda_{ijk}\right) \quad = \quad \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} \ ,$$

com o desvio do submodelo da independência mútua (acetato 184):

$$\ln(\lambda_{ijk}) \quad = \quad \mu + \alpha_i + \beta_j + \gamma_k$$

Se os modelos diferem significativamente, a hipótese de independência mútua deve ser rejeitada a favor da independência conjunta.

Os modelos log-lineares de independência conjunta de um par com o factor restante podem ser indicados de forma mnemónica com a indicação de qual o par de factores que é, conjuntamente, independente do terceiro. Assim, por exemplo, o modelo acima A pode ser referenciado como modelo (B:C).

[MC]: Modelos para independências condicionais

Consideremos agora a independência de um par de factores, condicional ao terceiro factor, por exemplo, a independência de (A,B), condicional a C.

Como foi salientado, esta independência condicional pode escrever-se apenas em termos das probabilidades conjuntas e marginais:

$$p_{ijk} = \frac{p_{i.k} \cdot p_{.jk}}{p_{..k}}$$

Tendo este facto em conta, será necessário que existam dois termos de dupla interacção num modelo log-linear associado a esta hipótese: a interacção A-C e a interacção B-C, que são ambas necessárias para se poder dispensar a tripla interacção.

[MC]: Modelos para independências condicionais (cont.)

Por um raciocínio análogo ao utilizado no caso das independências conjuntas, o valor esperado na célula (i,j,k), no caso de haver independência de (A,B) condicional a C, será da forma

$$\lambda_{ijk} = E[Y_{ijk}] = n\,p_{ijk} = n\,\frac{p_{i.k}\,p_{.jk}}{p_{..k}} \ .$$

Para modelar esta relação, admite-se que o logaritmo deste valor esperado é uma soma tipo ANOVA, com:

- uma parcela comum a todas as observações;
- parcelas de efeitos principais de cada factor; e ainda
- parcelas de interacção entre os factores A-C e B-C.

[MC]: Modelos para independências condicionais (cont.)

Obtem-se o modelo da independência de (A,B) condicional a C:

$$\ln(\lambda_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} \ ,$$

sendo

$$\begin{aligned}
\mu &= \ln\left(n \cdot \frac{p_{11.} \cdot p_{1.1}}{p_{1..}}\right) \\
\alpha_i &= \ln\left(\frac{p_{i.1}}{p_{1.1}}\right) \qquad \forall i = 2:a \qquad (\alpha_1 = 0) \\
\beta_j &= \ln\left(\frac{p_{.j1}}{p_{.11}}\right) \qquad \forall j = 2:b \qquad (\beta_1 = 0) \\
\gamma_k &= \ln\left(\frac{p_{1.k} \cdot p_{.1k} \cdot p_{..1}}{p_{1.1} \cdot p_{.11} \cdot p_{..k}}\right) \qquad \forall i = 2:a \qquad (\gamma_1 = 0) \\
(\alpha\gamma)_{ik} &= \ln\left(\frac{p_{i.k} \cdot p_{1.1}}{p_{1.k} \cdot p_{i.1}}\right) \qquad \forall i = 2:a \ , k = 2:c \qquad [(\alpha\gamma)_{1k} = (\alpha\gamma)_{i1} = 0] \\
(\beta\gamma)_{jk} &= \ln\left(\frac{p_{.jk} \cdot p_{.11}}{p_{.1k} \cdot p_{.j1}}\right) \qquad \forall j = 2:b \ , k = 2:c \qquad [(\beta\gamma)_{1k} = (\beta\gamma)_{j1} = 0]
\end{aligned}$$

[MC]: Modelos para independências condicionais (cont.)

A justificação para esta opção de modelo está, como já se indicou, no facto de ser possível recuperar as probabilidades p_{ijk} , desde que se mantenham as duas interacções duplas indicadas.

A justificação para estes parâmetros do modelo está num raciocínio análogo ao que se utilizou no caso de modelos para independências conjuntas.

Os estimadores de máxima verosimilhança dos parâmetros resultam ser, mais uma vez, os que resultam de substituir cada probabilidade p pela correspondente probabilidade estimada $\hat{p}$, dada pela frequência relativa correspondente na tabela.

[MC]: Testes a tipos de independências

O modelo agora discutido contém como submodelos:

- o modelo de independência mútua
  (se todas as interacções são nulas);
- o modelo de independência conjunta de (B,C) com A
  (se $(\alpha\gamma)_{ik} = 0$, para todo o i e k);
- o modelo de independência conjunta de (A,C) com B
  (se $(\beta\gamma)_{jk} = 0$, para todo o j e k).

Pode-se testar a independência condicional em relação às duas independências conjuntas que surgem como casos particulares deste modelo anulando, ou uma ou outra, das duplas interacões presentes.

[MC]:

O facto dos modelos surgirem como modelos encaixados está associado às implicações entre os tipos de independência considerados atrás.

Tal como para os modelos associados aos tipos anteriores de independências, pode recorrer-se a uma notação compacta, utilizando os termos de dupla interacção presentes no modelo, para o descrever. Assim, podemos representar o modelo da independência de (A,B) condicional a C como o modelo (A:C,B:C).

[MC]: 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	$\mu+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\alpha\gamma)_{ik}+(\beta\gamma)_{jk}+(\alpha\beta\gamma)_{ijk}$	

[MC]: Um exemplo famoso

Completemos a discussão de modelos log-lineares para tabelas de contingência com três factores de classificação, com um exemplo famoso, a que está associado o chamado paradoxo de Simpson. O exemplo pode ser visto em mais pormenor no livro de A. Agresti referido na bibliografia.

O exemplo tem por base dados reais relacionados com o sistema jurídico dos EUA: 326 julgamentos em que o réu foi considerado culpado de homícidio foram classificados de acordo com três factores, cada um dos quais possui apenas dois níveis.

- sentença do réu (condenação à morte, ou não);
- raça do réu (branco ou negro);
- raça da vítima (branco ou negro).

[MC]: Um exemplo famoso (cont.)

		Sentença	
Raça Réu	Raça Vítima	Pena de Morte	Outra Pena
Branco	Branco	19	132
	Negro	0	9
Negro	Branco	11	52
	Negro	6	97

Tabela: Dados de 326 julgamentos por homicídio nos EUA de Radelet, M. *Racial characteristics and the imposition of the death penalty*, American Sociology Review, 1981, 46: 918-927.

Comecemos por analisar a tabela criando a *data frame*

```
> radelet
  contagens sentenca raca.reu raca.vitima
1        19    Morte   branco      branco
2         0    Morte   branco       negro
3        11    Morte    negro      branco
4         6    Morte    negro       negro
5       132    Outra   branco      branco
6         9    Outra   branco       negro
7        52    Outra    negro      branco
8        97    Outra    negro       negro
```

[MC]: Um exemplo famoso (cont.)

Foi efectuada no R a análise a um modelo log-linear apenas abaixo do modelo saturado: um modelo com todas as duplas interacções, mas sem tripla interacção. Os resultados obtidos foram os seguintes.

```
Call: glm(formula = contagens ~ sentenca + raca.reu + raca.vitima +
    sentenca:raca.reu + sentenca:raca.vitima + raca.reu:raca.vitima,
    family = poisson)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                     2.9272     0.2297  12.746  < 2e-16 ***
sentencaOutra                   1.9581     0.2451   7.991 1.34e-15 ***
raca.reunegro                  -0.5001     0.3690  -1.355   0.1753
raca.vitimanegro               -4.0491     0.6065  -6.676 2.46e-11 ***
sentencaOutra:raca.reunegro    -0.4402     0.4009  -1.098   0.2722
sentencaOutra:raca.vitimanegro  1.3242     0.5193   2.550   0.0108 *
raca.reunegro:raca.vitimanegro  3.3580     0.3820   8.791  < 2e-16 ***
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 395.91531  on 7  degrees of freedom
Residual deviance:   0.70074  on 1  degrees of freedom
AIC: 50.382
Number of Fisher Scoring iterations: 4
```

Como a tabela de contingências é do tipo $2 \times 2 \times 2$, cada linha dos resultados está associada a um tipo de efeitos.

[MC]: Um exemplo famoso (cont.)

Os resultados sugerem que a interacção "sentença:raça do réu" é a menos significativa de todas, tendo-se repetido a análise na sua ausência. Os resultados obtidos foram os seguintes.

```
Call: glm(formula = contagens ~ sentenca + raca.reu + raca.vitima +
    sentenca:raca.vitima + raca.reu:raca.vitima, family = poisson)

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                      3.0525     0.1878  16.251  < 2e-16 ***
sentencaOutra                    1.8137     0.1969   9.212  < 2e-16 ***
raca.reunegro                   -0.8741     0.1500  -5.828 5.60e-09 ***
raca.vitimanegro                -3.7820     0.5515  -6.858 6.99e-12 ***
sentencaOutra:raca.vitimanegro   1.0579     0.4635   2.282   0.0225 *
raca.reunegro:raca.vitimanegro   3.3116     0.3786   8.748  < 2e-16 ***
---
    Null deviance: 395.9153  on 7  degrees of freedom
Residual deviance:   1.8819  on 2  degrees of freedom
AIC: 49.563
Number of Fisher Scoring iterations: 4
```

O modelo ajustado é um modelo de independência condicional dos factores (Raça do réu,Sentença), face ao factor Raça da vítima.

[MC]: Um exemplo famoso (cont.)

Os valores estimados dos parâmetros do modelo têm a interpretação indicada no Acetato 197, simplificada pelo facto de haver apenas dois níveis em todos os factores.

O valor estimado $\hat{\mu} = \ln(\hat{\lambda}_{111}) = 3.0525$ significa que o valor esperado na célula de referência (a célula de condenação, para réus brancos e vítimas brancas) é $e^{3.0525} = 21.1682$, próximo do valor observado (19).

No caso do primeiro factor (Sentença), o valor $\hat{\alpha}_2 = 1.8137$ significa que, em relação ao valor esperado para a célula de referência, o valor esperado na célula resultante de transitar para "Outra sentença" (mantendo réu e vítima brancos) é $e^{1.8137} = 6.133$ vezes maior, ou seja, é $e^{1.8137} * 21.1682 = 126.3050$, próximo do valor observado (132).

[MC]: Paradoxo de Simpson

Vimos que a sentenças e raça do réu podem ser consideradas independentes, dada a raça da vítima.

Mas olhando para a tabela verifica-se que em nenhum caso, houve condenação à morte de um réu branco quando a vítima era negra, enquanto que no caso de um réu negro e vítima branca, a proporção de condenações à morte era mais elevada do que o habitual: 17.5%, comparado com os 11.4% de condenações à morte globais, sendo a mais alta das percentagens também de qualquer das combinações de raça do réu e raça da vítima.

Este exemplo ilustra uma situação conhecida por paradoxo de Simpson.

[MC]: Tabelas parciais

Comecemos por introduzir um conceito auxiliar. Designa-se por tabela parcial uma sub-tabela resultante de fixar um nível de um dos factores.

Por exemplo, a tabela parcial resultante de fixar o nível "Branco" do factor "Raça da vítima" é a seguinte:

		Sentença	
Raça Vítima	Raça Réu	Pena de Morte	Outra Pena
Branco	Branco	19	132
	Negro	11	52

E a tabela parcial associada a fixar o nível "Negro" do factor "Raça da vítima" é a seguinte:

		Sentença	
Raça Vítima	Raça Réu	Pena de Morte	Outra Pena
Negro	Branco	0	9
	Negro	6	97

[MC]: Tabelas marginais

O conceito de tabela parcial não deve ser confundido com o de tabela marginal, que se obtém somando as contagens ao longo de todos os níveis de um dos factores.

Assim, por exemplo, a tabela marginal correspondente a Sentença vs. Raça do réu obtém-se somando as entradas correspondentes para ambas as raças da vítima e é dada por:

	Sentença		
Raça Réu	Pena de Morte	Outra Pena	Freq. marginal
Branco	19	141	160
Negro	17	149	166
Freq. Marginal	36	290	326

[MC]: O paradoxo de Simpson

Analisando as tabelas parciais e marginal surge um resultado aparentemente contraditório.

Ao inspeccionar a tabela marginal, vemos que a proporção de réus brancos condenados à morte foi de $\frac{19}{160} = 11.875\%$. A mesma proporção para réus negros foi de $\frac{17}{166} = 10.241\%$.

Ou seja, juntando as vítimas das duas raças, a percentagem de brancos condenados à morte é superior à percentagem de negros condenados à morte.

[MC]: O paradoxo de Simpson (cont.)

Mas analisemos agora as tabelas parciais, em que se consideram apenas as vítimas de uma ou outra côr. A tabela parcial para vítimas de raça branca mostra como, nesse caso, a percentagem de réus brancos condenados à morte é de $\frac{19}{19+132} = 12.58\%$, sendo a percentagem para os réus negros de $\frac{11}{11+52} = 17.46\%$, e portanto superior.

Analisando a tabela parcial para vítimas de raça negra temos que, nesse caso, a percentagem de réus brancos condenados à morte é de 0%, enquanto que a percentagem de réus negros condenados à morte é de $\frac{6}{6+97} = 5.83\%$. Assim, controlando a raça da vítima, e qualquer que esta seja a percentagem de negros condenados à morte é superior: o contrário do que se tinha concluído quando se ignorou a raça da vítima.

[MC]: O paradoxo de Simpson (cont.)

Ou seja, as associações nas tabelas parciais `Sentença-Raça do réu` são ao contrário das associações na tabela marginal `Sentença-Raça do réu`. É esta a situação conhecida pela designação de paradoxo de Simpson.

Este exemplo mostra que tabelas parciais e tabelas marginais podem ter diferentes tipos de associação. Ou seja, pode ser enganador analisar apenas tabelas marginais.

Em particular, a independência de A e B condicional a C não implica a independência marginal de A e B.