Apontamentos de
MODELAÇÃO ESTATÍSTICA II

Jorge Cadima
Departamento de Matemática – Instituto Superior de Agronomia

Outubro 2010
Mestrado em Matemática Aplicada às Ciências Biológicas

Conteúdo

Capítulo 1

Introdução

A disciplina de Modelação Estatística II integra-se no Mestrado em Matemática Aplicada às Ciências Biológicas, do Departamento de Matemática do Instituto Superior de Agronomia. Vem no seguimento das disciplinas de Complementos de Probabilidades e Estatística e de Complementos de Álgebra e Análise, Modelação Estatística I e Estatística Multivariada, cuja matéria será usada frequentemente.

Na disciplina de Modelação Estatística II é utilizado o **programa informático R**. Trata-se de um programa baseado na linguagem computacional S, especialmente concebida para aplicações estatísticas, e exposta nos livros:

- Becker, R.A.; Chambers, J.M. & Wilks, A.R. (1988) *The S Language.* Wadsworth & Brooks/Cole
- Chambers, J.M. & Hastie, T. (1992) *Statistical Models in S.* Wadsworth & Brooks/Cole

A linguagem S conhece duas concretizações na forma de programas informáticos: uma comercial, e outra o programa (gratuito e de código público) R. Os dois programas diferem em vários aspectos de funcionalidade, compatibilidades, etc. Mas trata-se, no fundamental, de dois "dialectos" da linguagem S. John Chambers, co-autor dos dois livros acima referidos, integra o núcleo central de desenvolvimento do programa R.

O programa R pode ser descarregado gratuitamente através da Internet, a partir do *site*:

http://cran.r-project.org

ou em vários outros *sites* que reproduzem o conteúdo do endereço atrás referido (*mirror sites*, cujos endereços estão indicados no *site* acima referido). Existem versões do programa R já compiladas para execução nos principais sistemas operativos (Linux, Macintosh, Windows).

Informação vária sobre o programa (Manuais, respostas a perguntas frequentes, páginas de Ajuda, Boletim informativo) podem ser também obtidos através da rede, a partir do endereço acima, ou em:

http://www.r-project.org

1.1 O Programa da disciplina

Na disciplina de Modelação Estatística II são abordadas várias técnicas de modelação que generalizam, de forma diferente, o Modelo Linear estudado na disciplina de Modelação Estatística I: regressões robustas, testes não-paraméticas do tipo ANOVA, a Regressão Não Linear, os Modelos Lineares Generalizados, correspondem a técnicas em que se modificam uma ou várias das hipóteses do Modelo Linear, resultando técnicas de aplicação em contextos diferentes ou mais gerais.

Capítulo 2

Métodos Não-paramétricos

(Em construção)

2.1 A regressão robusta

(Em construção)

2.2 As ANOVAs não paramétricas

(Em construção)

Capítulo 3

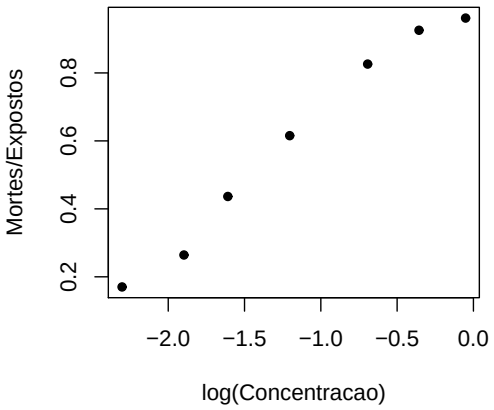
A Regressão Não Linear

(Em construção)

3.1 Exercícios

1. Uma experiência visa estudar a toxicidade da nicotina para moscas de fruta. Um dado número de moscas são expostas a várias concentrações de nicotina (medidas em $g/100cc$), e contabiliza-se a proporção de moscas que morrem após um intervalo de tempo protocolar. Os resultados obtidos, e a sua representação gráfica, utilizando a transformação logarítmica das concentrações no eixo horizontal, foram:

Concentração	Mortalidade
0.10	0.1702
0.15	0.2642
0.20	0.4364
0.30	0.6154
0.50	0.8261
0.70	0.9259
0.95	0.9615



Decide-se ajustar uma relação Não-Linear aos pontos representados no gráfico.

(a) Considere o modelo logístico com equação:

$$y = \frac{\alpha}{1 + \mathrm{e}^{-k(x-\gamma)}}.$$

Que interpretação pode ser dada aos parâmetros α, k e γ? (Pode admitir que $k > 0$). Justifique. Diga porque é que é lógico considerar, no nosso caso, que $\alpha = 1$. Estime empiricamente valores para k e γ, com base nas interpretações sugeridas.

(b) Admitindo que $\alpha = 1$, determine uma transformação linearizante da relação logística entre Mortalidade (y) e log-concentração (x), que permita obter estimativas iniciais dos valores dos parâmetros k e γ através duma Regressão Linear. Ajuste esta relação linearizada, obtendo estimativas para k e γ. Compare com as estimativas empíricas que sugeriu na alínea anterior e comente.

(c) Ajustou-se uma regressão não linear, fixando $\alpha = 1$ e utlizando o algoritmo de Gauss-Newton com estimativas iniciais de k e γ dadas pelos valores obtidos com a metodologia descrita na alínea anterior. Obtiveram-se os seguintes resultados:

```
Formula: Mortes/Expostos ~ 1/(1 + exp(-k * (log(Concentracao) - gama)))

Parameters:
     Estimate Std. Error t value Pr(>|t|)
k     2.08363    0.08869   23.49 2.60e-06 ***
gama -1.45827    0.02097  -69.54 1.16e-08 ***
```

```
---
Residual standard error: 0.01955 on 5 degrees of freedom
```

Comente os resultados obtidos.

2. Os processos de fermentação provocados por diferentes tipos de alimentos influenciam a capacidade de assimilação de nutrientes pelos animais. Uma forma de avaliar os processos de fermentação consiste em medir o volume de gases produzido no estômago de animais ao longo de várias horas após a ingestão de alimentos. Os dados seguintes referem-se à produção cumulativa de gás em ruminantes alimentados com vagens de Espinheiro da Virgínia (*Gleditsia triacanthos*). As unidades de produção de gás são ml/(200 mg de matéria seca).

Horas após ingestão	2h	4h	8h	12H	24h	48H	56h	72h
Produção de gás	7,1	11,1	17,5	24,3	37,3	44,4	45,2	46,6

Foi ajustada uma curva de Gompertz para descrever a relação entre a variável preditora Horas e a variável resposta Produção cumulativa de gás. Os resultados obtidos foram os seguintes:

```
> summary(nls(Producao ~ a*exp(-exp(-k*(Horas-g))),
start=list(a=45,g=1,k=0.05)))
Formula: Producao ~ a * exp(-exp(-k * (Horas - g)))
Parameters:
   Estimate Std. Error t value Pr(>|t|)
a 45.728899   0.605152   75.57 7.69e-09 ***
g  7.394692   0.342236   21.61 3.94e-06 ***
k  0.100158   0.006519   15.36 2.12e-05 ***
Residual standard error: 0.9564 on 5 degrees of freedom
```

(a) Interprete o significado biológico dos parâmetros estimados.

(b) Indique, justificando, um intervalo a 95% de confiança (aproximado) para a produção total de gás após a ingestão. Considera que estão reunidas todas as condições para garantir a validade deste intervalo de confiança?

3. Um estudo morfométrico em lagostins considera a relação alométrica $y = \alpha x^{\beta}$ entre o **Comprimento da tenaz** (y) e o **Comprimento dactil** (x).

(a) Deduza a justificação para uma relação deste tipo, bem como a interpretação biológica que se pode dar ao parâmetro β.

(b) Ajustou-se o Modelo Não-Linear, tendo por base a relação alométrica, a um conjunto de n=63 pares de observações (x_i, y_i) $(i = 1, ..., 63)$. Foram obtidos os seguintes resultados:

```
> summary(nls(crayfish[,13] ~ a*crayfish[,15]^b, start=list(a=1,b=1)))
Parameters:
  Estimate Std. Error t value Pr(>|t|)
a  2.64470    0.34596   7.644  1.8e-10 ***
b  0.83153    0.04622  17.991  < 2e-16 ***
Residual standard error: 1.27 on 61 degrees of freedom
```

i. Comente o significado biológico da relação estimada.

ii. Será admissível falar-se numa relação isométrica entre os comprimentos considerados? Comente a validade da aplicação do procedimento utilizado neste caso.

(c) A relação alométrica acima considerada é linearizável. Descreva essa linearização e relacione os parâmetros α e β da relação não-linear com os novos parâmetros que obtém após a linearização.

(d) Efectuou-se o ajustamento da relação linearizada considerada na alínea anterior. Foram obtidos os seguintes resultados:

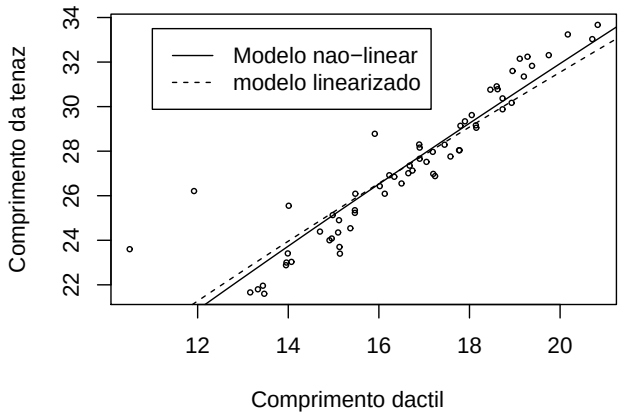
```
> summary(lm(log(crayfish[,13]) ~log(crayfish[,15])))
Call: lm(formula = log(crayfish[, 13]) ~ log(crayfish[, 15]))
Residuals:
      Min        1Q    Median        3Q       Max
-0.083989 -0.027676 -0.004472  0.020479  0.213780
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.14165    0.13653   8.362 1.05e-11 ***
log(crayfish[, 15])  0.77100    0.04872  15.825  < 2e-16 ***

Residual standard error: 0.05265 on 61 degrees of freedom
Multiple R-Squared: 0.8041,     Adjusted R-squared: 0.8009
F-statistic: 250.4 on 1 and 61 DF,  p-value:     0
```

i. Comente os valores estimados para os parâmetros do modelo linearizado. Compare com os valores obtidos nos parâmetros do modelo não-linear original.

ii. Construa um intervalo de confiança a 95% que lhe permita estudar a hipótese de isometria a partir do modelo linearizado. Comente os seus resultados, tendo em conta os resultados obtidos com base no modelo não-linear.

iii. Comente brevemente as causas das diferenças nas estimativas obtidas em cada caso (modelo não-linear e modelo linearizado), tendo também em conta o seguinte gráfico onde se reproduz a nuvem dos 63 pontos observados e as relações estimadas (nas unidades originais) entre as duas variáveis.



4. Uma relação não-linear habitualmente utilizada para descrever contagens de reacções enzimáticas (y) em função de concentração de subestrato (x) é dada pela equação de Michaelis-Menten.

   No contexto de modelos de rendimento total de culturas agrícolas, o Modelo de Michaelis-Menten é conhecido pela designação de modelo de Shinozaki & Kira, sendo usual a seguinte parametrização:

   $$y = \frac{x}{a + bx} \qquad (a > 0, b > 0 \quad \text{e} \quad x > 0)$$

   No contexto das pescas, o Modelo Michaelis-Menten surge com a designação de Modelo Beverton-Holt, no estudo de Recrutamento *vs.* dimensão dos mananciais de progenitores. Neste contexto, a parametrização usual é:

   $$y = \frac{\alpha x}{1 + \frac{x}{k}} \qquad (\alpha > 0, k > 0 \quad \text{e} \quad x > 0)$$

   (a) Indique, justificando, quais as interpretações possíveis para os parâmetros nas parametrizações de Shinozaki & Kira e de Beverton & Holt.

   (b) O facto de o Modelo de Michaelis-Menten ser linearizável depende da parametrização específica que é utilizada? Justifique, e em caso de os modelos com as novas parametrizações serem linearizáveis, indique qual a relação entre as variáveis e parâmetros das relações originais e das relações linearizadas.

   Considere os dados *Puromycin*, estudados nas aulas com o Modelo Michaelis-Menten, *apenas para os casos das células tratadas*, i.e., apenas para as 12 primeiras observações do conjunto de dados *Puromycin* do *R*. Para cada uma das parametrizações alternativas agora referidas (Shinozaki & Kira e Beverton & Holt), responda às seguintes alíneas.

   (c) Ajuste regressões não-lineares. Compare os resultados com os resultados obtidos nas aulas com a parametrização usual do Modelo Michaelis-Menten e comente.

(d) Caso esteja interessado em obter um intervalo de confiança para o valor assintótico da velocidade de reacção enzimática, para grandes concentrações, qual a parametrização mais conveniente? Estude a possibilidade de obter os referidos intervalos de confiança com cada uma das parametrizações e apresente as suas conclusões.

(e) Determine as funções perfil dos parâmetros em cada uma das parametrizações, e trace os respectivos gráficos ("traços de perfil"). Comente os resultados.

(f) Calcule as medidas de curvatura RMS e compare-as com as obtidas na parametrização usual do Modelo Michaelis-Menten. Comente os resultados.

(g) Tendo em conta as suas respostas às alíneas anteriores, qual a parametrização que considera mais conveniente?

Capítulo 4

Introdução aos Modelos Lineares Generalizados

Neste Capítulo far-se-á uma introdução a uma gama muito vasta de modelos que generalizam o Modelo Linear: os Modelos Lineares Generalizados (MLGs, ou, usando as iniciais em língua inglesa, GLMs). O Modelo Linear é um caso particular de Modelo Linear Generalizado.

O conceito de Modelo Linear Generalizado foi introduzido e formalizado por McCullagh e Nelder (1989), sintetizando de forma notável os aspectos comuns a muitos modelos já estudados e que, nalguns casos, eram utilizados há largas décadas, como sejam os modelos de *probits* e de *logits* ou os *modelos log-lineares*, para não falar já do próprio Modelo Linear.

Como veremos, a generalização permitida pelos Modelos Lineares Generalizados incide essencialmente sobre dois aspectos fundamentais: (i) a distribuição de probabilidades associada à variável-resposta aleatória Y já não se restringe à Normal, podendo ser qualquer distribuição numa classe designada *família exponencial de distribuições* (que será estudada na Secção 4.1); e (ii) a relação entre a combinação linear das variáveis preditoras e a variável-resposta pode ser mais geral do que a prevista no Modelo Linear.

4.1 Conceitos básicos

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

Componente aleatória Y: É a variável-resposta que se deseja estudar, tratando-se duma **variável aleatória**, da qual se recolhem n **observações independentes** e cuja **distribuição faz parte da família exponencial de distribuições** (que será definida mais adiante);

Componente Sistemática: Consiste numa **combinação linear de variáveis preditoras**. Havendo p variáveis preditoras e dadas n combinações de valores das p variáveis, pode construir-se a matriz $\mathbf{X}_{n\times(p+1)}$ de forma idêntica ao que se fez aquando do estudo do Modelo Linear (incluíndo uma

primeira coluna de uns, associada a uma constante aditiva). Nesse caso, a componente sistemática do modelo é dada pela relação:

$$\eta = \mathbf{X}\boldsymbol{\beta} \tag{4.1}$$

Função de ligação: Função *diferenciável* e *monótona* g que associa as componentes aleatória e sistemática, através duma relação da forma:

$$g(E[Y]) = \mathbf{X}\boldsymbol{\beta} \qquad \Longleftrightarrow \qquad g(\mu_i) \;=\; \mathbf{x}_i^t\boldsymbol{\beta} \;=\; \sum_{j=0}^{p}\beta_j x_{ij} \qquad (i = 1:n) \tag{4.2}$$

onde $\mu = E[Y]$ será o vector de valores esperados das n observações de Y e μ_i cada uma das suas componentes, e $\mathbf{x}_i$ representa a i-ésima linha da matriz $\mathbf{X}$ (enquanto vector-coluna), isto é, o conjunto de valores das variáveis preditoras para os quais se efectuou a i-ésima observação da variável-resposta. Caso a função g seja invertível (o que sucede se a monotonia acima exigida fôr estrita), pode escrever-se:

$$E[Y] = g^{-1}\left(\mathbf{X}\boldsymbol{\beta}\right) \qquad \Longleftrightarrow \qquad \mu_i \;=\; g^{-1}\left(\mathbf{x}_i^t\boldsymbol{\beta}\right) \;=\; g^{-1}\left(\sum_{j=0}^{p}\beta_j x_{ij}\right) \qquad (i = 1:n) \tag{4.3}$$

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.*

Assinale-se que, ao contrário do que foi feito aquando do estudo do Modelo Linear, nesta forma de apresentação dos Modelos Lineares Generalizados não são explicitados erros aleatórios aditivos. A flutuação aleatória da variável-resposta é dada directamente pela sua distribuição de probabilidades, sendo a relação com as variáveis preditoras estabelecida apenas para uma transformação do seu valor esperado.

4.1.1 A família exponencial de distribuições

A definição da família exponencial de distribuições é apresentada aqui na forma bi-paramétrica usada por McCullagh & Nelder (1989) e ainda por Turkman & Silva (2000).

Definição 4.1 *Seja Y uma variável aleatória, cuja função densidade ou de massa probabilística se pode escrever na forma:*

$$f(\, y \mid \theta, \phi \,) \quad = \quad \mathrm{e}^{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)} \tag{4.4}$$

*onde θ e ϕ são parâmetros (escalares reais) e $a(\cdot), b(\cdot)$ e $c(\cdot)$ são funções reais conhecidas. Então diz-se que a distribuição de Y pertence à **família exponencial de distribuições**. O parâmetro θ designa-se **parâmetro natural** da distribuição, e ϕ é designado o **parâmetro de dispersão**.*

Admite-se que as funções que definem esta relação são o suficientemente bem comportadas para que seja possível efectuar as operações que seguidamente se estudarão. Tal é o caso para as distribuições consideradas nesta disciplina.

A família exponencial de distribuições é vasta e inclui algumas das mais importantes e conhecidas distribuições, quer de variáveis aleatórias contínuas, quer de variáveis aleatórias discretas. Inclui os seguintes casos particulares:

Distribuição Normal. A conhecida função densidade duma Normal (univariada) pode ser escrita da seguinte forma:

$$\begin{aligned} f(y|\mu,\sigma) &= \frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} \\ &= \mathrm{e}^{\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)-\frac{y^2-2y\mu+\mu^2}{2\sigma^2}} \\ &= \mathrm{e}^{\frac{y\mu-\frac{\mu^2}{2}}{\sigma^2}+\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)-\frac{y^2}{2\sigma^2}} \end{aligned}$$

que é da forma (4.4) com:

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

Distribuição Poisson. Recorde-se que uma variável aleatória discreta tem distribuição de Poisson se toma valores em $\mathbb{N}_0$ com função de massa probabilística $P[Y=k] = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}$. Para os valores $y \in \{0,1,2,...\}$, podemos escrever a função de massa probabilística duma Poisson como:

$$\begin{aligned} f(y|\lambda) &= \mathrm{e}^{-\lambda}\frac{\lambda^y}{y!} \\ &= \mathrm{e}^{-\lambda+y\log(\lambda)-\log(y!)} \end{aligned}$$

que é da forma (4.4) com:

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

Distribuição 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:

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

que é da forma (4.4) com:

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

Distribuição Binomial/n. Embora a distribuição Binomial não pertença à família de distribuições exponenciais, uma pequena transformação de variáveis com distribuição Binomial origina distribuições nessa família. Seja X uma variável aleatória com distribuição Binomial de parâmetros n e p, isto é, $X \cap B(n,p)$. Defina-se agora uma nova variável aleatória $Y = \frac{1}{n}X$. Tem-se $P[Y = y] = P[X = ny]$. A distribuição de Y pertence à família exponencial. De facto, Y toma valores no conjunto $F = \{0, \frac{1}{n}, \frac{2}{n}, ..., 1\}$. Para $y \in F$, a função de massa probabilística de Y pode escrever-se da seguinte forma:

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

que é da forma (4.4) com:

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

Distribuição Gama. A família das distribuições Gama inclui como caso particular a distribuição Quiquadrado e também a distribuição Exponencial. O facto de a distribuição Gama também pertencer à família exponencial de distribuições representa assim um alargamento assinalável do campo de aplicabilidade dos MLGs. Como foi visto na disciplina de Complementos de Probabilidades e Estatística, uma variável aleatória Y diz-se ter distribuição Gama com parâmetros α e β se apenas

tomar valores em $\mathbb{R}^+$, com função densidade da forma

$$\begin{aligned} f(y \mid \alpha, \beta) &= \frac{1}{\beta^\alpha \Gamma(\alpha)}\, y^{\alpha-1}\, \mathrm{e}^{-\frac{y}{\beta}} \\ &= \mathrm{e}^{-\log \beta^\alpha - \log \Gamma(\alpha) + (\alpha-1)\log y - \frac{y}{\beta}} \\ &= \mathrm{e}^{\left(-\frac{1}{\beta}\right) y - \alpha \log \beta - \log \Gamma(\alpha) + (\alpha-1)\log y} \\ &= \mathrm{e}^{\alpha\left(-\frac{1}{\alpha\beta}\right) y + \alpha \log\left(\frac{1}{\beta}\right) - \log \Gamma(\alpha) + (\alpha-1)\log y} \\ &= \mathrm{e}^{\alpha\left(-\frac{1}{\alpha\beta}\right) y + \alpha \log\left(\frac{\alpha}{\alpha\beta}\right) - \log \Gamma(\alpha) + (\alpha-1)\log y} \\ &= \mathrm{e}^{\alpha\left(-\frac{1}{\alpha\beta}\right) y + \alpha \log\left(\frac{1}{\alpha\beta}\right) + \alpha \log \alpha - \log \Gamma(\alpha) + (\alpha-1)\log y} \\ &= \mathrm{e}^{\frac{\left(-\frac{1}{\alpha\beta}\right) y + \log\left(\frac{1}{\alpha\beta}\right)}{\frac{1}{\alpha}} + \alpha \log \alpha - \log \Gamma(\alpha) + (\alpha-1)\log y} \end{aligned}$$

que é da forma (4.4) com:

- $\theta = -\frac{1}{\alpha\beta}$
- $\phi = \frac{1}{\alpha}$
- $b(\theta) = -\log\left(\frac{1}{\alpha\beta}\right) = -\log(-\theta)$
- $a(\phi) = \phi = \frac{1}{\alpha}$
- $c(y, \phi) = \alpha \log \alpha - \log \Gamma(\alpha) + (\alpha - 1)\log y$

4.1.2 A componente sistemática

As variáveis preditoras cuja combinação linear define a chamada *componente sistemática* do Modelo, podem, em geral, ser variáveis quantitativas ou variáveis indicatrizes, à semelhança do que sucede no Modelo Linear. O seu papel, e a interpertação dos coeficientes que lhes estão associados, far-se-á de forma mais adequada ao estudar cada caso concreto de Modelo Linear Generalizado.

Tal como no Modelo Linear, considera-se que as variáveis preditoras não são aleatórias, ou então que a relação estabelecida entre variável-resposta e variáveis preditoras é *condicional* aos valores observados das variáveis preditoras. Em qualquer dos casos, tratam-se os valores das variáveis preditoras como constantes.

4.1.3 A função de ligação

Como se viu anteriormente, a relação entre variável-resposta e variáveis predictoras é efectuada através duma *função de ligação* que relaciona uma combinação linear das variáveis predictoras com o valor esperado da variável resposta (ver a equação (4.2), na página 12).

A mais simples de todas as funções de ligação será a *ligação identidade*, quando $g(\mu) = \mu$. Essa é a função ligação utilizada no Modelo Linear, em que a relação de base entre valor esperado da variável resposta e variáveis preditoras é apenas $E[Y] = \mathbf{X}\boldsymbol{\beta}$.

No entanto, as mais importantes funções de ligação são funções que tornam, para cada distribuição da família exponencial, o valor esperado da variável-resposta igual ao parâmetro natural, θ.

Definição 4.2 *Num Modelo Linear Generalizado, a função $g(\cdot)$ diz-se uma* ***função de ligação canónica*** *para a variável-resposta Y, se $g(E[Y]) = \theta$.*

Repare-se que 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. Pode dizer-se que a ligação canónica representa de alguma forma uma função de ligação "natural" para o respectivo tipo de distribuição da variável-resposta.

4.2 Exemplos de Modelos Lineares Generalizados

Vejamos alguns exemplos de Modelos Lineares Generalizados que são considerados em mais pormenor nesta disciplina.

4.2.1 O Modelo Linear

O Modelo Linear estudado na disciplina de Modelação Estatística I é um caso particular de Modelo Linear Generalizado, em que cada uma das n observações da variável-resposta Y tem distribuição Normal, com variância constante σ^2, e onde a função de ligação é a função identidade[1]. De facto, o Modelo Linear considera $E[Y] = \sum_{j=0}^{p} \beta_j \mathbf{x}_j$, pelo que é o próprio valor esperado de $E[Y]$ a ser uma combinação linear das variáveis preditoras. Note-se ainda que *a função de ligação identidade é a ligação canónica para a distribuição Normal*, já que $g(E[Y]) = E[Y] = \mu$, que é o parâmetro natural θ da definição geral da família exponencial no caso de uma distribuição Normal (ver página 13).

Assim, justifica-se a afirmação de que os MLG são uma generalização do Modelo Linear.

4.2.2 A Regressão Logística

Considere-se um Modelo com variável resposta dicotómica (binária), isto é, que apenas toma dois possíveis valores que, por conveniência, representaremos por 0 e 1. Nesse caso, é evidente a inadequação do Modelo Linear que prevê uma distribuição Normal para a variável-resposta. Uma variável-resposta aleatória dicotómica terá uma distribuição de Bernoulli, com probabilidades p e $1-p$ associadas aos seus possíveis valores 1 e 0, respectivamente. Admitindo que o parâmetro p varia nas n observações de Y, o valor esperado da i-ésima observação de Y será dado por:

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

[1] Recorde-se que a exigência de observações independentes da variável-resposta é comum a todos os Modelos Lineares Generalizados.

Uma função de ligação será, neste caso, uma função relacionando este valor esperado p_i da variável-resposta com uma combinação linear das variáveis preditoras, ou seja, uma função $g(\cdot)$ tal que:

$$g(p(\mathbf{x})) \quad = \quad \mathbf{x}^t\boldsymbol{\beta} \qquad \Longleftrightarrow \qquad p(\mathbf{x}) \quad = \quad g^{-1}\left(\mathbf{x}^t\boldsymbol{\beta}\right) \tag{4.5}$$

A função de ligação canónica será a função $g(\cdot)$ que transforme $p(\mathbf{x})$ no parâmetro natural θ da distribuição Bernoulli, visto na página 14, ou seja, $\theta = \log\left(\frac{p}{1-p}\right)$. Logo, para garantir que $g(E[Y_i]) = g(p_i) = \theta_i$, toma-se:

$$g(p) \quad = \quad \log\left(\frac{p}{1-p}\right) \tag{4.6}$$

Esta é a *função de ligação canónica para variáveis resposta de Bernoulli* e designa-se a **função *logit***. Com estas opções, o MLG é conhecido pela designação **Regressão Logística**.

A função de ligação *logit* representa o logaritmo do quociente entre a probabilidade de a variável Y tomar o valor 1 e a probabilidade de tomar o valor 0. Esse quociente de probabilidades é conhecido na literatura anglo-saxónica por *odds ratio*[2]. É habitual designar a função de ligação *logit* como um *log-odds ratio*.

Como se viu, uma função de ligação no contexto de MLGs serve para relacionar o valor esperado de cada Y_i com combinações lineares das variáveis preditoras. No caso em apreço, consideramos que os logits dos valores esperados p_i são combinações lineares das variáveis preditoras $X_0, X_1, ..., X_p$. Mais concretamente, dado um conjunto $\mathbf{x}$ de observações nas variáveis preditoras, temos:

$$\begin{aligned}
& g(p) \quad = \quad \log\left(\frac{p}{1-p}\right) \quad = \quad \mathbf{x}^t\boldsymbol{\beta} \\
\Longleftrightarrow \quad & \frac{p}{1-p} \quad = \quad \mathrm{e}^{\mathbf{x}^t\boldsymbol{\beta}} \\
\Longleftrightarrow \quad & \frac{1}{1-p} \quad = \quad 1 + \frac{p}{1-p} \quad = \quad 1 + \mathrm{e}^{\mathbf{x}^t\boldsymbol{\beta}} \\
\Longleftrightarrow \quad & 1-p \quad = \quad \frac{1}{1+\mathrm{e}^{\mathbf{x}^t\boldsymbol{\beta}}} \\
\Longleftrightarrow \quad & p \quad = \quad 1 - \frac{1}{1+\mathrm{e}^{\mathbf{x}^t\boldsymbol{\beta}}} \quad = \quad \frac{\mathrm{e}^{\mathbf{x}^t\boldsymbol{\beta}}}{1+\mathrm{e}^{\mathbf{x}^t\boldsymbol{\beta}}}
\end{aligned}$$

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

$$p(\mathbf{x}_i^t\boldsymbol{\beta}) \quad = \quad g^{-1}\left(\mathbf{x}_i^t\boldsymbol{\beta}\right) \quad = \quad \frac{\mathrm{e}^{\mathbf{x}_i^t\boldsymbol{\beta}}}{1+\mathrm{e}^{\mathbf{x}_i^t\boldsymbol{\beta}}} \quad = \quad \frac{1}{1+\mathrm{e}^{-\mathbf{x}_i^t\boldsymbol{\beta}}} \tag{4.7}$$

No caso de uma componente sistemática constituída por uma constante aditiva e uma única variável preditora *quantitativa*, isto é, se:

$$\mathbf{x}^t\boldsymbol{\beta} \quad = \quad \beta_0 + \beta_1 x,$$

[2]Este conceito é de difícil tradução para português. O conceito de *odds*, como razão entre a probabilidade de se verificar um determinado acontecimento e a probabilidade de esse mesmo acontecimento não se verificar, é utilizado na linguagem corrente dos países anglo-saxónicos. Em português não existe uma palavra sinónima.

a relação (4.7) resulta ser a equação de uma *curva logística*, e está na origem da designação "Regressão Logística":

$$p(x) \quad = \quad g^{-1}\left(\beta_0 + \beta_1 x\right) \quad = \quad \frac{\mathrm{e}^{\beta_0+\beta_1 x}}{1+\mathrm{e}^{\beta_0+\beta_1 x}} \quad = \quad \frac{1}{1+\mathrm{e}^{-(\beta_0+\beta_1 x)}} \tag{4.8}$$

Esta relação logística entre variável resposta e variável preditora é ilustrada na Figura 4.1.

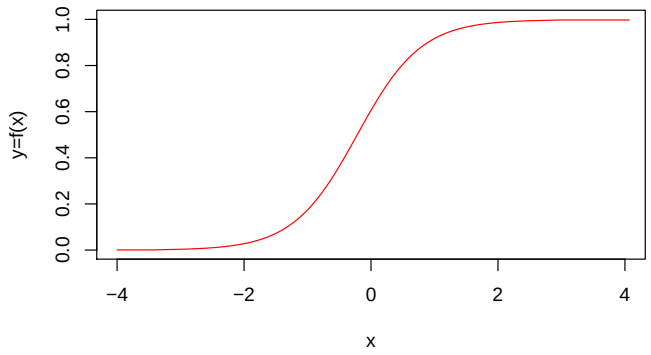


Figura 4.1: Gráfico da função $y = f(x) = \frac{1}{1+\mathrm{e}^{-(\beta_0+\beta_1 x)}}$, com $\beta_0 = 0.5$ e $\beta_1 = 2$.

Esta curva descreve a relação entre a probabilidade de a variável-resposta tomar o valor 1 (o seu valor esperado) e os valores da (neste caso única) variável preditora X. Trata-se de uma função crescente em x, caso $\beta_1 > 0$, e decrescente em x, caso $\beta_1 < 0$. O caso de uma função decrescente para as probabilidades $p = P[Y = 1]$ pode ser sempre transformado num caso equivalente com função crescente de probabilidades, trocando os acontecimentos que dão à variável aleatória Y os valores 0 e 1, respectivamente.

Ainda no caso de haver uma única variável preditora *quantitativa* (além de constante aditiva), o parâmetro β_1 tem a seguinte interpretação: como $\frac{p(x)}{1-p(x)} = \mathrm{e}^{\beta_0} \cdot \mathrm{e}^{\beta_1 x}$, cada aumento de uma unidade na variável preditora X traduz-se num efeito multiplicativo sobre o *odds ratio* de e^{β_1}. Caso a variável preditora X seja uma variável *indicatriz* (associada a um factor preditor), β_1 indica a diferença do *log-odds ratio* quando a indicatriz toma o valor 1 e quando toma o valor zero, ou seja, indica o incremento no *log-odds ratio* resultante de a observação em questão pertencer à categoria de que X é variável indicatriz.

Numa caso mais geral, a relação (4.7) entre uma variável resposta e *várias* variáveis preditoras quantitativas, significa que a probabilidade de a variável-resposta tomar o valor 1 (o seu valor esperado) descreve uma relação logística do tipo indicado na Figura 4.1 como função dos valores da *combinação linear* das variáveis preditoras, $\eta = \mathbf{x}^t\boldsymbol{\beta}$. Trata-se duma **função estritamente monótona**, que tem um único **ponto de inflexão quando o preditor linear** $\mathbf{x}^t\boldsymbol{\beta} = 0$, ponto de inflexão esse a que corresponde uma probabilidade de êxito $p(0) = 0.5$ (recorde-se o estudo da função logística feito no Capítulo da Regressão Não Linear). A função (4.7) tem ainda uma certa rigidez estrutural, associada à sua **simetria em torno do ponto de inflexão**, isto é, ao facto de que, vista como função do preditor linear $\eta = \mathbf{x}^t\boldsymbol{\beta}$, a função de probabilidade verifica $p(-\eta) = 1 - p(\eta)$.

Assim, a função de ligação *logit* gera uma relação logística – equação (4.7) – para a probabilidade de êxito

p como função dos valores da combinação linear das variáveis preditoras. Esta função logística tem boas propriedades para representar uma probabilidade: para *qualquer* valor da componente sistemática, toma valores entre 0 e 1. O mesmo não acontece com uma relação linear $p(\mathbf{x}^t\boldsymbol{\beta}) = \mathbf{x}^t\boldsymbol{\beta} = \sum_{j=0}^{p} \beta_j x_j$, que pode tomar valores em toda a recta real $\mathbb{R}$, embora isso não faça sentido em termos de uma probabilidade.

As **interpretações dos coeficientes** β_j dadas para o caso de haver uma única variável preditora generalizam-se, quando existe mais do que uma variável preditora quantitativa: 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}.

Em muitas situações práticas é concebível aplicar Modelos de Regressão Logística. Se a variável-resposta Y regista a presença ou ausência de alguma característica ou fenómeno nas unidades experimentais (doença, morte, presença física duma espécie num dado território, etc.) ou assinala qual de duas categorias de classificação se verifica (por exemplo, um indivíduo observado tem, ou não, uma dada patologia), e caso se pretenda relacionar o valor esperado dessa variável resposta (isto é, a probabilidade do acontecimento associado ao valor 1) com valores de um conjunto de variáveis preditoras, então a Regressão Logística é uma opção a considerar.

A forma geral da relação (4.7) pode ser substituída por outras funções de comportamento análogo, embora nesse caso já não se trate de funções de ligação canónicas para uma distribuição Bernoulli. Na próxima Subsecção veremos dois exemplos desta situação.

4.2.3 Outros Modelos com variável-resposta binária

Admita-se de novo uma variável-resposta dicotómica (binária) com distribuição de Bernoulli e parâmetro p_i para a i-ésima de entre as n observações da amostra. Já se viu no caso anterior que a escolha de uma função de ligação *logit* gera uma relação logística entre a componente sistemática do modelo e o valor esperado da variável-resposta. Outras funções de ligação g, cujas funções inversas g^{-1} tenham a mesma forma geral que a logística, isto é que sejam curvas **sigmóides** (curvas monótonas, balisadas por duas assintotas horizontais e com um único ponto de inflexão) podem ser consideradas, embora já não sejam funções de ligação *canónicas*. Duas escolhas frequentes são indicadas a seguir.

O Modelo de Regressão Probit

A função de distribuição cumulativa (f.d.c.) de uma Normal Reduzida tem uma forma semelhante à indicada na Figura 4.1. Assim, seria possível substituir a opção $p(\mathbf{x}^t\boldsymbol{\beta}) = \frac{1}{1+\mathrm{e}^{-\mathbf{x}^t\beta}}$ feita no Modelo de Regressão Logística pela opção:

$$p(\mathbf{x}^t\boldsymbol{\beta}) \quad = \quad g^{-1}\left(\mathbf{x}^t\boldsymbol{\beta}\right) \quad = \quad \Phi(\mathbf{x}^t\boldsymbol{\beta}) \tag{4.9}$$

onde Φ indica a f.d.c. duma $\mathcal{N}(0,1)$. Esta opção significa considerar como *função de ligação* a inversa da f.d.c. duma Normal reduzida:

$$\mathbf{x}^t\boldsymbol{\beta} \quad = \quad g\left(p(\mathbf{x}^t\boldsymbol{\beta})\right) \quad = \quad \Phi^{-1}\left(p(\mathbf{x}^t\boldsymbol{\beta})\right) \tag{4.10}$$

Esta opção de função de ligação dá origem ao chamado **Modelo de Regressão Probit**, ou apenas **Modelo Probit**. Trata-se de um Modelo muito usado em estudos de Toxicologia e que remonta a

1935 (para uma pequena nota histórica, veja-se McCullagh & Nelder, 1989). No contexto toxicológico, é frequente a existência de uma variável preditora X que indica a *dosagem* (ou log-dosagem) de um determinado produto tóxico. Para cada nível existe um *nível de tolerância* t ao tóxico, definido como sendo o limiar acima do qual o produto tóxico provoca a morte do indivíduo. Mas esse nível de tolerância varia entre indivíduos. Assim, e num contexto onde se escolhem aleatoriamente indivíduos, pode-se definir uma variável aleatória T que indique os níveis de tolerância. Definindo uma outra variável aleatória binária Y, que toma valor 0 caso um indivíduo sobreviva a uma dada dosagem de tóxico, e 1 em caso contrário, as variáveis T e Y estarão relacionadas através da seguinte relação:

$$P[\, Y = 1 \mid x \,] \quad = \quad P[T \leq x] \quad = \quad p(x) \tag{4.11}$$

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

$$p(x) \quad = \quad \Phi\left(\frac{x-\mu}{\sigma}\right) \, . \tag{4.12}$$

Tem-se então o Modelo Probit para esta única variável preditora. Neste caso, os coeficientes habituais da relação linear, β_0 e β_1 aparecem como $\beta_0 = -\frac{\mu}{\sigma}$ e $\beta_1 = \frac{1}{\sigma}$. A estimação dos parâmetros β_0 e β_1 está, assim, associada à estimação dos parâmetros da distribuição de T.

Em geral, para qualquer número de variáveis preditoras, o Modelo Probit produz uma função $p(\mathbf{x}^t\boldsymbol{\beta})$ de probabilidade de êxito de Y cujo comportamento é muito semelhante à da sua congénere no Modelo Logit: trata-se duma **função estritamente crescente**, com um único **ponto de inflexão quando o preditor linear $\mathbf{x}^t\boldsymbol{\beta}$ se anula**, a que corresponde uma probabilidade de êxito $p(0) = 0.5$. A função (4.10) revela ainda a rigidez estrutural idêntica à logística, na medida em que possui simetria em torno do ponto de inflexão, isto é, $p(-\eta) = 1 - p(\eta)$, para qualquer η. De inconveniente, em comparação com a função usada na Regressão Logística, tem o facto de não permitir uma interpretação tão fácil do significado dos parâmetros β_j da componente sistemática do Modelo, e ainda o facto de a função de ligação ser não-canónica.

O Modelo log-log do Complementar

Outra escolha frequente e com tradição histórica desde 1922 em estudos de organismos infecciosos (de novo, veja-se McCullagh & Nelder, 1989, para uma pequena resenha histórica) consiste em definir a probabilidade de êxito (ou valor esperado da variável-resposta binária Y) como sendo:

$$p(\mathbf{x}^t\boldsymbol{\beta}) \quad = \quad g^{-1}\left(\mathbf{x}^t\boldsymbol{\beta}\right) \quad = \quad 1 - \mathrm{e}^{-\mathrm{e}^{\mathbf{x}^t\beta}} \tag{4.13}$$

A função p que descreve a probabilidade de êxito é assim considerada como a diferença entre uma curva de Gompertz com valor assintótico um[3] e esse mesmo valor assintótico. O facto de se fixar o valor assintótico em 1 é natural, uma vez que a função p descreve *probabilidades*. Assinale-se que o contradomínio da função agora definida é o intervalo $]0, 1[$. A função de ligação será, neste caso, da forma:

$$\mathbf{x}^t\boldsymbol{\beta} \quad = \quad g\left(p(\mathbf{x}^t\boldsymbol{\beta})\right) \quad = \quad \log\left(-\log(1 - p(\mathbf{x}^t\boldsymbol{\beta}))\right) \tag{4.14}$$

expressão esta que justifica a designação do Modelo (por "complementar" entende-se o complementar do acontecimento associado aos "êxitos", e cuja probabilidade de ocorrência é $1 - p$).

[3] $\alpha = 1$, na notação usada aquando do estudo do modelo de Gompertz, no Capítulo da Regressão Não Linear.

No caso de haver uma única variável preditora X, ou seja, de $\mathbf{x}^t\boldsymbol{\beta}$ ser da forma $\mathbf{x}^t\boldsymbol{\beta} = \alpha + \beta x$, a função que devolve $p(x)$ para cada valor de x é a função distribuição cumulativa da distribuição de Gumbel. O aspecto gráfico da relação 4.13 é, nesse caso, indicado na Figura 4.2. Esta escolha para função inversa da função

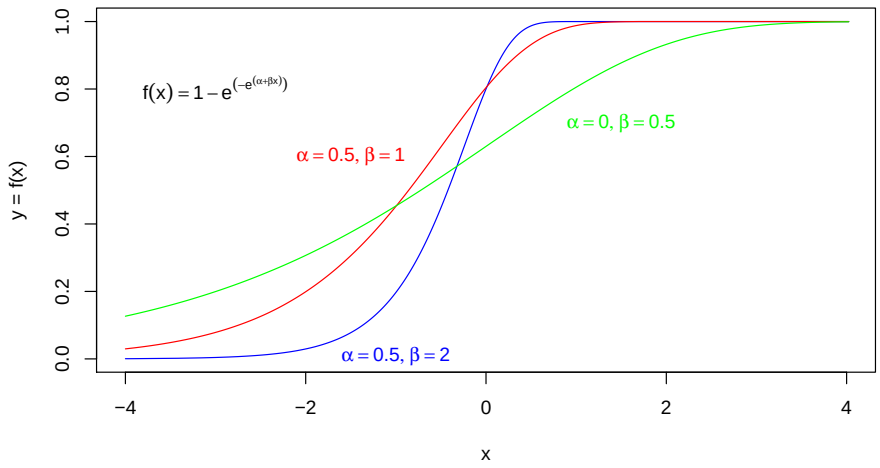


Figura 4.2: Gráfico da função $y = g^{-1}(\alpha + \beta x) = 1 - \mathrm{e}^{-\mathrm{e}^{(\alpha+\beta x)}}$, com valores de α e β indicados no gráfico.

ligação introduz algumas diferenças de comportamento, em relação aos Modelos Logit e Probit, no que respeita à relação da probabilidade de êxito da variável resposta Y com os valores do preditor linear $\mathbf{x}^t\boldsymbol{\beta}$. Embora a relação que exprime $p(\cdot)$ como função do preditor linear $\eta = \mathbf{x}^t\boldsymbol{\beta}$ seja igualmente estritamente monótona, e tenha igualmente um único ponto de inflexão, quando $\eta = 0$, o valor de probabilidade associado a esse ponto de inflexão já não se encontra a meio caminho na escala de probabilidades, sendo neste caso $1 - \frac{1}{\mathrm{e}}$. 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 os do Modelo da Regressão Logística.

Outras escolhas de função ligação

Assinale-se que outras funções de distribuição cumulativa de variáveis aleatórias contínuas têm uma forma genérica compatível com as exigências de funções inversas de funções de ligação para o contexto que está sendo considerado: crescentes e balizadas por assintotas horizontais em 0 e 1. Assim, em substituição da f.d.c. duma Normal reduzida feita no Modelo Probit, ou da f.d.c. duma Gumbel, feita no Modelo log-log do Complementar, poderia ser utilizada outra f.d.c. duma variável aleatória contínua, gerando um novo Modelo para este contexto.

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, que designamos por δ:

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

A função de ligação *logit*, usada na Regressão Logística, corresponde a um caso particular desta família, tomando $\delta = 1$, enquanto que a função de ligação log-log do complementar corresponde ao limite quando $\delta \longrightarrow 0$.

Dados binários e a distribuição Binomial

Nesta Subsecção consideraram-se variáveis resposta Y dicotómicas (binárias), associadas a n conjuntos de valores $\mathbf{x}_i$ das variáveis preditoras. No caso de haver repetições desses conjuntos das variáveis preditoras, a apresentação dos resultados das n observações pode ser feita de forma diferente. Admita-se que, em vez de termos n observações-resposta de valor 0 ou 1, foram contabilizadas quantas de entre n_i $(i = 1 : m)$ observações efectuadas para cada um de m diferentes conjuntos de valores $\mathbf{x}_i$ das variáveis preditoras corresponderam a "êxitos" (categoria "1"), sendo $n = \sum_{i=1}^{m} n_i$. Nesse caso, haverá m valores-resposta Y_i, que serão observações duma distribuição Binomial $B(n_i, p_i)$, uma vez que para cada um dos m conjuntos de valores observados dos preditores haverá uma mesma probabilidade p_i de "êxito". Como seria desejável, existem ligações íntimas na modelação, no contexto de MLGs, destas duas formas alternativas de encarar os dados. Por razões que serão explicadas em pormenor na Subsecção 4.7.4 (página 45), convém considerar como variável-resposta não as contagens de número de "êxitos" em cada caso, mas sim a *proporção* de "êxitos" em cada uma das m situações experimentais.

4.2.4 Modelos Log-Lineares

Considere-se agora uma situação onde **a variável-resposta Y é uma variável de contagem de acontecimentos**, com distribuição de Poisson. Por exemplo, considere-se uma experiência em que se deseja contar o número de ovos de um determinado insecto parasita que se encontram nas folhas de uma certa espécie vegetal.

Se Y tem uma distribuição de Poisson, toma valores em $\mathbb{N}_0$ com probabilidades $P[Y = k] = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}$. Assinale-se desde já que, podendo uma variável com distribuição de Poisson tomar valores ilimitadamente grandes, esta distribuição não será a escolha indicada para situações em que seja fixado, à partida, o número máximo de observações ou realizações do fenómeno que esteja a ser contado. Afirmar que as contagens não podem exceder um determinado limiar máximo não é a mesma coisa que afirmar que a probabilidade de determinados valores muito elevados de contagens se verificarem seja muito baixo, ou praticamente nulo. Esta última situação é perfeitamente compatível com uma distribuição de Poisson[4] para Y. O que se pretende excluir são situações onde existe um limiar estrutural, impossível de franquear, como por exemplo quando se fixa previamente um número n de indivíduos a observar e mais tarde se efectuam contagens sobre esses indivíduos, tornando impossível que as contagens excedam o valor n.

[4] Aliás, acontecerá sempre que $P[Y > k] \approx 0$, a partir de determinado valor de k que dependerá do parâmetro λ da distribuição Poisson de Y.

Como é sabido, o valor esperado de $Y \cap Po(\lambda)$ é igual ao valor do parâmetro, isto é, a λ. Assim, uma função de ligação será uma função $g(\cdot)$ tal que:

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

onde $\mathbf{x}^t\boldsymbol{\beta}$ é a componente sistemática do Modelo, isto é a combinação linear das variáveis preditoras.

Por outro lado, e como se viu na Secção 4.1.1, a distribuição de Poisson tem parâmetro natural θ dado por $\log(\lambda)$. Assim, a função de ligação canónica para uma variável-resposta com distribuição de Poisson é a ligação logarítmica:

$$g(\lambda) = \log(\lambda) = \mathbf{x}^t\boldsymbol{\beta} \iff \lambda(\mathbf{x}^t\boldsymbol{\beta}) = g^{-1}\left(\mathbf{x}^t\boldsymbol{\beta}\right) = \mathrm{e}^{\mathbf{x}^t\boldsymbol{\beta}} \tag{4.15}$$

Um Modelo assim definido designa-se um Modelo Log-Linear. Assinale-se que a relação (4.15) apenas permite valores postivos (mas ilimitados) para o parâmetro λ, o que está estruturalmente de acordo com as características do parâmetro λ duma distribuição Poisson.

No caso de haver uma única variável preditora X, a relação entre parâmetro λ da distribuição Poisson e variáveis preditoras fica, no Modelo Log-Linear, da forma: $\lambda = \mathrm{e}^{\alpha} \cdot \mathrm{e}^{\beta x}$. O significado de parâmetro β é fácil de deduzir: um aumento de uma unidade no valor da variável preditora multiplica o valor esperado da variável resposta por e^{β}. Como no caso anterior, a expressão "um aumento de uma unidade na variável preditora", que é clara no caso de uma variável preditora quantitativa, deve ser entendida, no caso de uma variável indicatriz X, como indicando o facto de uma observação passar a ser considerada como pertencendo à categoria de que X é indicatriz. Repare-se que, nesse caso, tem-se $\lambda = \mathrm{e}^{a}$ para observações que não pertencem à categoria assinalada pela indicatriz X, e $\lambda = \mathrm{e}^{a+b}$, caso contrário.

Estas interpretações generalizam-se para o caso de haver mais do que uma variável preditora. Nesse caso, 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}.

Tal como nos casos anteriores, outras funções de ligação são concebíveis para variáveis-resposta com distribuição de Poisson, mas nesta disciplina apenas será estudado o caso do Modelo Log-Linear, associado à função de ligação canónica para a distribuição de Poisson.

4.2.5 Modelos com variável resposta de distribuição Gama

Com a excepção do Modelo Linear, os restantes MLGs considerados até aqui tinham variável resposta discreta. Vejamos agora um exemplo de MLG com variável resposta contínua, mas não Normal (como no Modelo Linear e Regressão Não Linear). Em particular, consideremos uma variável resposta Y com distribuição Gama (que, como sabemos, inclui como casos particulares uma Exponencial ou uma Qui-quadrado).

Se $Y \cap G(\alpha, \beta)$, sabemos da disciplina de Complementos de Probabilidades e Estatística que $\mu = E[Y] = \alpha\beta$ e que $V[Y] = \alpha\beta^2 = \mu\beta$. Assim, *é característica das variáveis com distribuição Gama que a sua variância é proporcional à sua média*. Esta propriedade sugere que para situações onde a variância dos dados não pareça ser constante, como se admite no Modelo Linear e no Modelo Não Linear, mas varie proporcionalmente com a média, se considere a utilização de MLGs com variável resposta Gama. Esta característica estava associada aos "gráficos em funil" observados no estudo dos resíduos desses Modelos.

Naturalmente que a existência desse tipo de gráficos não chega, por si só, para garantir que a opção por uma variável resposta Gama seja adequada. Mas é seguramente uma opção a considerar.

Uma vez que para $Y \cap G(\alpha, \beta)$ se verifica $\mu = E[Y] = \alpha\beta$, as funções de ligação g num MLG com variável resposta Gama irão relacionar o produto $\mu = \alpha\beta$ com as combinações lineares das variáveis preditoras[5], $\mathbf{x}^t\boldsymbol{\beta}$.

A **função de ligação canónica** para modelos com distribuição Gama será a função g que transforma o valor esperado de Y, $\alpha\beta$, no parâmetro natural $\theta = -\frac{1}{\alpha\beta}$ (veja-se a página 15). Assim, e uma vez que o sinal negativo não é relevante na discussão[6], a função de ligação canónica para modelos com variável resposta Gama é a **função recíproco**:

$$g(\mu) \;=\; \frac{1}{\mu} \tag{4.16}$$

que transforma $\mu = \alpha\beta$ em $\frac{1}{\alpha\beta} = -\theta$.

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

$$g(\mu) \;=\; g(\alpha\beta) \;=\; \frac{1}{\alpha\beta} \;=\; \mathbf{x}^t\boldsymbol{\beta} \qquad \Longleftrightarrow \qquad \mu(\mathbf{x}^t\boldsymbol{\beta}) \;=\; g^{-1}\left(\mathbf{x}^t\boldsymbol{\beta}\right) \;=\; \frac{1}{\mathbf{x}^t\boldsymbol{\beta}} \tag{4.17}$$

sendo $\mu = \alpha\beta$.

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 hiperbólica,

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

Esta função está relacionada com a *curva de rendimento por planta de Shinozaki & Kira*, e também com a *curva de recrutamento por progenitor de Beverton & Holt*, estudadas no Capítulo da Regresão Não Linear. No contexto do Modelo Michaelis-Menten, aplicando à velocidade de reacções enzimáticas, a relação acima referida corresponde à velocidade da reacção por unidade de concentração.

Refira-se um aspecto pouco simpático associado à relação canónica dos MLGs com resposta Gama, referida na equação (4.17): embora o valor esperado da variável resposta Y tenha de ser positivo (uma vez que Y só toma valores positivos), a equação (4.17) permite que esse valor esperado seja negativo para alguns valores da variável preditora x (situação que depende dos parâmetros β_0 e β_1). Assim, e ao contrário do que acontecia nos modelos anteriores analisados, não existe uma "garantia estrutural" de que os valores de $E[Y]$ estimados façam sentido.

Nas Secções seguintes veremos como estimar parâmetros e fazer inferência sobre os parâmetros de Modelos Lineares Generalizados, como testar o ajustamento de um MLG e de seus Submodelos, e ainda, como estudar a validade das hipóteses subjacentes a cada Modelo. Serão considerados essencialmente os casos concretos de MLGs acima referidos, embora a discussão se processe num contexto mais geral de MLGs sempre que tal opção não acarrete dificuldades adicionais.

[5] Não confundir, neste contexto, o parâmetro β da distribuição Gama, com os parâmetros $\beta_0, \beta_1, ..., \beta_p$ associados às variáveis preditoras na combinação linear $\mathbf{x}^t\boldsymbol{\beta}$.

[6] O sinal poderá sempre ficar associado aos parâmetros da combinação linear sem perda de generalidade.

4.3 Estimação de parâmetros em MLGs

A estimação de parâmetros em Modelos Lineares Generalizados é feita pelo Método da Máxima Verosimilhança. Assim, a função verosimilhança é uma ferramenta fundamental no estudo de Modelos Lineares Generalizados.

No caso do Modelo Linear, a estimação dos parâmetros foi introduzida pelo Método dos Mínimos Quadrados, mas como foi visto, esses estimadores e os de Máxima Verosimilhança coincidem (Ver os apontamentos da disciplina de Modelação Estatística I). Para estimar parâmetros nos restantes MLGs, comecemos por considerar a função verosimilhança para n observações independentes $y_1, y_2, ..., y_n$ numa qualquer distribuição de probabilidades da família exponencial de distribuições. Tem-se:

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

Maximizar a verosimilhança equivale a maximizar o logaritmo da verosimilhança[7], ou seja, o expoente da última expressão.

A log-verosimilhança é dada por:

$$\ell(\theta, \phi \ ; \ y_1, y_2, ..., y_n) = \sum_{i=1}^{n} \left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i) \right] \tag{4.18}$$

Ora, num MLG, a componente sistemática e o valor esperado da variável resposta estão relacionados por $g(E[Y]) = \mathbf{x}^t\boldsymbol{\beta}$. No caso de uma função de ligação canónica tem-se $\theta = \mathbf{x}^t\boldsymbol{\beta}$. Em geral, pode escrever-se a log-verosimilhança como função dos parâmetros desconhecidos $\boldsymbol{\beta}$. O Método da Máxima Verosimilhança de estimar esses parâmetros consistirá então em escolher o vector $\boldsymbol{\beta}$ que torne máxima a função de log-verosimilhança $\ell(\boldsymbol{\beta})$.

É sabido que a maximização da função de $p+1$ variáveis $\ell(\boldsymbol{\beta})$ tem como condição necessária[8]:

$$\frac{\partial \ell(\hat{\boldsymbol{\beta}})}{\partial \beta_j} = 0, \qquad \forall \ j = 0 : p$$

Importa referir que, 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 $\boldsymbol{\beta}$), nem que, existindo máximo, este seja único. Nos casos concretos abordados nesta disciplina, a situação não apresenta grandes dificuldades[9].

Vejamos agora a como este processo se concretiza para alguns MLGs específicos.

[7] O logaritmo é uma função estritamente crescente.

[8] Recorde-se que se admite que as funções $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$ são suficientemente regulares para que as operações envolvidas sejam permitidas. Neste caso, admite-se que a log-verosimilhança é uma função diferenciável no seu domínio.

[9] Para uma discussão mais pormenorizada desta questão, veja-se o ponto 2.2.4. de Turkman e Silva (2000) e as referências que aí se encontram.

4.3.1 O Modelo de Regressão Logística

No Modelo de Regressão Logística, as n observações independentes referem-se a uma Variável aleatória com distribuição de Bernoulli. Pelo que já foi visto na Secção 4.1.1, a função de verosimilhança destas n observações será dada por:

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

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

$$\begin{aligned} \ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left(-\log\left(1+\mathrm{e}^{\mathbf{x}_i^t\boldsymbol{\beta}}\right) + y_i\mathbf{x}_i^t\boldsymbol{\beta}\right) \\ \iff \ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{j=0}^{p} y_i x_{ij}\beta_j - \sum_{i=1}^{n} \log\left(1+\mathrm{e}^{\sum_{j=0}^{p} x_{ij}\beta_j}\right) \end{aligned} \tag{4.19}$$

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

$$\frac{\partial \ell(\hat{\boldsymbol{\beta}})}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} \frac{\mathrm{e}^{\sum_{k=0}^{p} x_{ik}\hat{\beta}_k}}{1+\mathrm{e}^{\sum_{k=0}^{p} x_{ik}\hat{\beta}_k}} \cdot x_{ij} = 0 \qquad \forall j = 0:p \tag{4.20}$$

Ao contrário do que acontece para o Modelo Linear, estas $p+1$ equações normais formam um *sistema não-linear* de equações nas $p+1$ incógnitas $\hat{\beta}_j$, $j = 0:p$.

A não-linearidade nos parâmetros estimados $\hat{\boldsymbol{\beta}}$ impede de escrever o sistema de equações (4.20) como uma equação matricial envolvendo o vector $\hat{\boldsymbol{\beta}}$. Mas é possível utilizar uma notação mnemónica matricial, definindo o vector $\hat{\mathbf{p}}$ de probabilidades estimadas, cuja i-ésima componente é dada por:

$$\hat{p}_i = \frac{\mathrm{e}^{\sum_{j=0}^{p} x_{ij}\hat{\beta}_j}}{1+\mathrm{e}^{\sum_{j=0}^{p} x_{ij}\hat{\beta}_j}} \tag{4.21}$$

De facto, com esta notação, o sistema de $p+1$ equações (4.20) toma a forma:

$$\mathbf{X}^t\mathbf{y} = \mathbf{X}^t\hat{\mathbf{p}} \tag{4.22}$$

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

4.3.2 O Modelo Log-Linear

No Modelo Log-Linear, as n observações independentes referem-se a uma variável aleatória com distribuição de Poisson. Pelo que já foi visto na Secção 4.1.1, a função de verosimilhança destas n observações será dada por:

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

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

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

Deixando cair a última parcela, que é constante nos parâmetros β_j e, como tal, dispensável no processo de identificar os pontos máximos, toma-se:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(-\mathrm{e}^{\sum_{j=0}^{p} x_{ij}\beta_j} + y_i \sum_{j=0}^{p} x_{ij}\beta_j\right) \tag{4.23}$$

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

$$\begin{aligned} \frac{\partial \ell(\hat{\boldsymbol{\beta}})}{\partial \beta_j} &= \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} x_{ij}\mathrm{e}^{\sum_{k=0}^{p} x_{ik}\hat{\beta}_k} = 0 \qquad \forall j = 0:p \\ \iff \frac{\partial \ell(\hat{\boldsymbol{\beta}})}{\partial \beta_j} &= \sum_{i=1}^{n} x_{ij}\left[y_i - \mathrm{e}^{\sum_{k=0}^{p} x_{ik}\hat{\beta}_k}\right] = 0 \qquad \forall j = 0:p \end{aligned} \tag{4.24}$$

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

A não-linearidade nos parâmetros estimados $\hat{\boldsymbol{\beta}}$ também impede de escrever o sistema de equações (4.24) como uma equação matricial envolvendo o vector $\hat{\boldsymbol{\beta}}$. Mas é possível utilizar uma notação mnemónica matricial, definindo o vector $\hat{\boldsymbol{\lambda}}$ de probabilidades estimadas, cuja i-ésima componente é dada por:

$$\hat{\lambda}_i = \mathrm{e}^{\sum_{j=0}^{p} x_{ij}\hat{\beta}_j}$$

De facto, com esta notação, o sistema de $p+1$ equações (4.24) toma a forma:

$$\mathbf{X}^t\mathbf{y} = \mathbf{X}^t\hat{\boldsymbol{\lambda}} \tag{4.25}$$

A não-linearidade do sistema exigirá métodos numéricos que serão considerados em seguida.

4.4 Algoritmos de Estimação

Na Subsecção anterior foi visto que, em geral, o sistema de $p+1$ equações normais resultante de procurar maximizar a função de log-verosimilhança num Modelo Linear generalizado é um sistema não-linear:

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

A solução deste sistema tem de ser obtida por métodos numéricos, havendo um algoritmo geral de resolução no contexto de MLGs. Este algoritmo é uma modificação do algoritmo de Newton-Raphson, e é conhecido por vários nomes: Método Iterativo de Mínimos Quadrados Ponderados ou Re-ponderados

(IWLS ou IRLS, respectivamente, utilizando as iniciais em língua inglesa), ou ainda o Método de Fisher (*Fisher Scoring Method*, em inglês).

O Método de Newton-Raphson baseia-se na ideia de trabalhar com uma aproximação de segunda ordem da função log-verosimilhança, utilizando a fórmula de Taylor com desenvolvimento em torno duma estimativa inicial do vector $\hat{\boldsymbol{\beta}}$, o vector $\boldsymbol{\beta}^{[0]}$. Designando por $\frac{\partial \ell}{\partial \beta}(\boldsymbol{\beta})$ o vector gradiente de $\ell(\boldsymbol{\beta})$ calculado no ponto $\boldsymbol{\beta}$, e por $H_{\boldsymbol{\beta}}$ a matriz Hessiana da função $\ell(\cdot)$, nesse mesmo ponto, tem-se a aproximação:

$$\ell(\boldsymbol{\beta}) \approx \ell_0(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}^{[0]}) + \left(\frac{\partial \ell}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}^{[0]})\right)^t \left(\boldsymbol{\beta} - \boldsymbol{\beta}^{[0]}\right) + \frac{1}{2}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{[0]}\right)^t H_{\boldsymbol{\beta}^{[0]}} \left(\boldsymbol{\beta} - \boldsymbol{\beta}^{[0]}\right) \tag{4.26}$$

Em vez de proceder à maximização de $\ell(\boldsymbol{\beta})$, maximiza-se a aproximação $\ell_0(\boldsymbol{\beta})$. Como foi estudado na disciplina de Complementos de Álgebra e Análise, é condição necessária para que exista máximo de ℓ_0 em $\boldsymbol{\beta}$ que o vector gradiente se anule nesse ponto, isto é, que: $\frac{\partial \ell}{\partial \beta}(\boldsymbol{\beta}) = \mathbf{0}$.

Como pode ser visto no Anexo A, quando a função que se pretende maximizar é uma combinação linear ou uma forma quadrática das variáveis, o cálculo do vector gradiente é particularmente simples. De facto, pode-se facilmente comprovar que:

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

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

Assim, resulta que:

$$\frac{\partial \ell_0}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) = \frac{\partial \ell}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}^{[0]}) + H_{\boldsymbol{\beta}^{[0]}}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{[0]}\right).$$

Logo, *admitindo a invertibilidade de* $H_{\boldsymbol{\beta}^{[0]}}$, tem-se:

$$\frac{\partial \ell_0}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) = 0 \iff \boldsymbol{\beta} = \boldsymbol{\beta}^{[0]} - H^{-1}_{\boldsymbol{\beta}^{[0]}}\left(\frac{\partial \ell}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}^{[0]})\right).$$

Esta relação fornece a base do processo iterativo no **algoritmo de Newton-Raphson**, tomando-se:

$$\boldsymbol{\beta}^{[i+1]} = \boldsymbol{\beta}^{[i]} - H^{-1}_{\boldsymbol{\beta}^{[i]}}\left(\frac{\partial \ell}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}^{[i]})\right) \tag{4.27}$$

Naturalmente que a possibilidade de aplicar com êxito este algoritmo exige a existência e invertibilidade das matrizes Hessianas de ℓ nos sucessivos pontos $\boldsymbol{\beta}^{[i]}$. Como é próprio do método de Newton-Raphson, em qualquer contexto, não está automaticamente garantida a convergência do algoritmo para qualquer ponto inicial (neste caso, para qualquer $\boldsymbol{\beta}^{[0]}$), mesmo quando existe e é único o máximo da função log-verosimilhança. É, no entanto, de esperar que, dada a existência e unicidade do máximo, a convergência seja melhor quanto mais próximo $\boldsymbol{\beta}^{[0]}$ estiver do ponto onde a função tem o seu máximo.

Como se verá adiante, aquando da concretização dos cálculos para alguns dos Modelos, o cálculo da matriz Hessiana da log-verosimilhança nos pontos $\boldsymbol{\beta}^{[i]}$ pode ser computacionalmente exigente. O **algoritmo de Fisher** é uma modificação do algoritmo de Newton-Raphson, que consiste em substituir, na expressão (4.27), a matriz Hessiana pela **matriz de informação de Fisher**, definida como sendo o simétrico do valor esperado da matriz Hessiana:

$$\mathbf{I}_{\boldsymbol{\beta}^{[i]}} = -E\left[H_{\boldsymbol{\beta}^{[i]}}\right] \tag{4.28}$$

A matriz de informação de Fisher define-se à custa do conceito de *quantidade de informação de Fisher*, dado na disciplina de Complementos de Probabilidades e Estatística. A sua utilização neste contexto resulta da simplificação computacional que, em geral, é introduzida pela substituição da matriz Hessiana por menos o seu valor esperado.

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

$$\boldsymbol{\beta}^{[i+1]} = \boldsymbol{\beta}^{[i]} + I^{-1}_{\boldsymbol{\beta}^{[i]}} \left(\frac{\partial \ell}{\partial \boldsymbol{\beta}} (\boldsymbol{\beta}^{[i]}) \right) \tag{4.29}$$

Assinale-se que, 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. Nesse caso, o método de Fisher e o de Newton-Raphson coincidem. Esta é uma das razões que confere às ligações canónicas a sua importância.

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

$$\boldsymbol{\beta}^{[i+1]} = \left(\mathbf{X}^t \mathbf{W}^{[i]} \mathbf{X} \right)^{-1} \mathbf{X}^t \mathbf{W}^{[i]} \mathbf{z}^{[i]} \tag{4.30}$$

onde $\mathbf{z}^{[i]}$ é uma linearização da função de ligação $g(y)$, escrita como função dos parâmetros $\boldsymbol{\beta}$ e $\mathbf{W}^{[i]}$ é uma matriz diagonal. As expressões gerais para $\mathbf{z}^{[i]}$ e $\mathbf{W}^{[i]}$ podem ser encontradas em McCullagh & Nelder (1989) ou Turkman e Silva (2000). Para os Modelos que serão estudados em mais pormenor nesta disciplina, podem ser encontradas as expressões concretas nas próximas Subsecções.

A expressão (4.30) significa que o algoritmo de Fisher está associado a uma **projecção não-ortogonal**, em que, quer o vector $\mathbf{z}^{[i]}$, quer os subespaços envolvidos na projecção, são re-definidos em cada iteração do algoritmo. De facto, a combinação linear $\mathbf{X}\boldsymbol{\beta}^{[i+1]}$, com $\boldsymbol{\beta}^{[i+1]}$ dado pela equação (4.30) é o produto de $\mathbf{z}^{[i]}$ pela matriz $\mathbf{X} \left(\mathbf{X}^t \mathbf{W}^{[i]} \mathbf{X} \right)^{-1} \mathbf{X}^t \mathbf{W}^{[i]}$. Ora, esta matriz é idempotente (como se pode facilmente verificar), pelo que se trata duma matriz de projecção. Não é, em geral, simétrica, a não ser que a matriz diagonal $\mathbf{W}^{[i]}$ verifique $\mathbf{X}^t \mathbf{W}^{[i]} = \mathbf{X}^t$. De qualquer forma, o Método de Fisher baseia-se em ideias de Mínimos Quadrados em sentido generalizado, isto é, envolvendo projecções não-ortogonais.

4.4.1 Algoritmo IRLS para alguns Modelos

Vejamos a concretização do algoritmo Iterativo de Mínimos Quadrados Ponderados para dois dos Modelos Lineares Generalizados considerados anteriormente.

Modelo Logit

Já foi visto na expressão (4.20) que as derivadas parciais de primeira ordem da log-verosimilhança de n observações, no contexto do Modelo Logit (Regressão Logística) são dadas por:

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} \frac{\mathrm{e}^{\sum_{k=0}^{p} x_{ik}\beta_k}}{1 + \mathrm{e}^{\sum_{k=0}^{p} x_{ik}\beta_k}} \cdot x_{ij}, \qquad \forall j = 0 : p$$

As derivadas parciais de segunda ordem, que definem os elementos da matriz Hessiana, são dadas por:

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \beta_j \partial \beta_l}(\boldsymbol{\beta}) &= -\sum_{i=1}^{n} x_{ij} x_{il} \cdot \frac{\mathrm{e}^{\sum_{k=0}^{p} x_{ik}\beta_k}}{1+\mathrm{e}^{\sum_{k=0}^{p} x_{ik}\beta_k}} \cdot \frac{1}{1+\mathrm{e}^{\sum_{k=0}^{p} x_{ik}\beta_k}} \\ \iff &= -\sum_{i=1}^{n} x_{ij} x_{il} \cdot p_i \cdot (1-p_i) \end{aligned} \tag{4.31}$$

onde p_i indica o valor da probabilidade associada ao i-ésimo conjunto de valores das variáveis preditoras.

A matriz Hessiana da função de log-verosimilhança ℓ, nos pontos correspondentes às iterações $\boldsymbol{\beta}^{[i]}$, será assim constituída pelos valores destas derivadas parciais de segunda ordem. Saliente-se que, tal 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 (ou seja, os Métodos de Newton-Raphson e de Fisher coincidem).

Defina-se agora a matriz $n \times n$ diagonal $\mathbf{W}$, cujos elementos diagonais são dados pelos n valores $p_i(1-p_i)$. A matriz Hessiana definida pelas equações (4.31), e a matriz de informação de Fisher associada, podem escrever-se, em termos matriciais, como:

$$H = -\mathbf{X}^t \mathbf{W} \mathbf{X} \tag{4.32}$$

$$I = \mathbf{X}^t \mathbf{W} \mathbf{X} \tag{4.33}$$

A equação que define a iteração dos vectores $\boldsymbol{\beta}$ no algoritmo IRLS para a Regressão Logística é assim dada, na passagem da i-ésima para a $(i+1)$-ésima iteração, por:

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

Definindo o vector $\mathbf{z}^{[i]} = \mathbf{X}\boldsymbol{\beta}^{[i]} + \left(\mathbf{W}^{[i]}\right)^{-1}\left(\mathbf{y} - \mathbf{p}^{[i]}\right)$, tem-se uma expressão de transição da iterada i para a iterada $i+1$ que salienta que o algoritmo IRLS está associado a projecções não-ortogonais:

$$\boldsymbol{\beta}^{[i+1]} = \left(\mathbf{X}^t \mathbf{W}^{[i]} \mathbf{X}\right)^{-1} \mathbf{X}^t \mathbf{W}^{[i]} \mathbf{z}^{[i]} \tag{4.34}$$

A expressão indicada para o vector $\mathbf{z}^{[i]}$ pode ser entendida como uma aproximação linear da função de ligação do Modelo Logit, em torno do ponto $\mathbf{p}^{[i]}$. De facto,

$$g(p) = \log\left(\frac{p}{1-p}\right) \qquad \Longrightarrow \qquad g'(p) = \frac{1}{p(1-p)}.$$

Logo, designando:

$$z^{[i]} = g\left(p^{[i]}\right) + g'\left(p^{[i]}\right)\left(y - p^{[i]}\right)$$

e, recordando as relações entre função de ligação e parte sistemática, bem como a definição da matriz $\mathbf{W}$, tem-se, em termos matriciais:

$$\mathbf{z}^{[i]} = \mathbf{X}\boldsymbol{\beta}^{[i]} + \left(\mathbf{W}^{[i]}\right)^{-1}\left(\mathbf{y} - \mathbf{p}^{[i]}\right) \tag{4.35}$$

Modelo Log-Linear

Na discussão anterior sobre a log-verosimilhança o contexto do Modelo Log-Linear, foi visto que as derivadas parciais de primeira ordem da log-verosimilhança são dadas por:

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij} \left[y_i - \mathrm{e}^{\sum_{k=0}^{p} x_{ik}\beta_k} \right], \qquad \forall j = 0 : p$$

Assim, as derivadas parciais de segunda ordem serão:

$$\frac{\partial^2 \ell}{\partial \beta_l \partial \beta_j}(\boldsymbol{\beta}) = -\sum_{i=1}^{n} x_{ij} x_{il} \mathrm{e}^{\sum_{k=0}^{p} x_{ik}\beta_k}$$

Escrevendo $\lambda_i = \mathrm{e}^{\sum_{k=0}^{p} x_{ik}\beta_k}$ para a estimativa do valor esperado de Y associado aos valores das variáveis preditoras na i-ésima observação, tem-se a forma mais compacta:

$$\frac{\partial^2 \ell}{\partial \beta_l \partial \beta_j}(\boldsymbol{\beta}) = -\sum_{i=1}^{n} x_{ij} x_{il} \lambda_i, \qquad \forall j, l = 0 : p \tag{4.36}$$

De novo, e uma vez que foi usada uma função de ligação canónica, estes elementos da matriz Hessiana da log-verosimilhança do Modelo Log-Linear não dependem da variável resposta Y, pelo que Hessiana e seu valor esperado são iguais, ou seja, o Método de Newton-Raphson e de Fisher coincidem.

Também de forma análoga ao que foi visto para o caso do Modelo Logit, defina-se a matriz $n \times n$ diagonal $\mathbf{W}$, cujos elementos diagonais são dados pelos n valores λ_i. A matriz Hessiana definida pelas equações (4.36), e a correspondente matriz de informação de Fisher podem escrever-se, em termos matriciais, como:

$$H = -\mathbf{X}^t \mathbf{W} \mathbf{X} \tag{4.37}$$

$$I = \mathbf{X}^t \mathbf{W} \mathbf{X} \tag{4.38}$$

A equação que define a iteração dos vectores $\boldsymbol{\beta}$ no algoritmo IRLS para a Regressão Logística é assim dada, na passagem da i-ésima para a $(i+1)$-ésima iteração, por:

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

Definindo o vector $\mathbf{z}^{[i]} = \mathbf{X}\boldsymbol{\beta}^{[i]} + \left(\mathbf{W}^{[i]} \right)^{-1} \left(\mathbf{y} - \boldsymbol{\lambda}^{[i]} \right)$, tem-se uma expressão de transição da iterada i para a iterada $i+1$ de aspecto idêntico à que já foi vista para o caso do Modelo Logit:

$$\boldsymbol{\beta}^{[i+1]} = \left(\mathbf{X}^t \mathbf{W}^{[i]} \mathbf{X} \right)^{-1} \mathbf{X}^t \mathbf{W}^{[i]} \mathbf{z}^{[i]} \tag{4.39}$$

A expressão indicada para o vector $\mathbf{z}^{[i]}$ pode ser entendida como uma aproximação linear da função de ligação do Modelo Log-Linear, em torno do ponto $\boldsymbol{\lambda}^{[i]}$. De facto,

$$g(\lambda) = \log(\lambda) \qquad \Longrightarrow \qquad g'(\lambda) = \frac{1}{\lambda}.$$

Logo, considerando:

$$z^{[i]} = g\left(\lambda^{[i]} \right) + g'\left(\lambda^{[i]} \right) \left(y - \lambda^{[i]} \right)$$

tem-se, em termos matriciais, recordando a definição da matriz $\mathbf{W}$ e a ligação entre valor esperado de Y e parte sistemática do Modelo:

$$\mathbf{z}^{[i]} = \mathbf{X}\boldsymbol{\beta}^{[i]} + \left(\mathbf{W}^{[i]} \right)^{-1} \left(\mathbf{y} - \boldsymbol{\lambda}^{[i]} \right) \tag{4.40}$$

4.5 Inferência sobre os parâmetros

Na disciplina de Complementos de Probabilidades e Estatística deste Mestrado foram dados alguns resultados gerais relativos a estimadores de máxima verosimilhança. Em particular, foi visto como esse tipo de estimadores são *consistentes* e, em condições gerais de regularidade são:

- assintoticamente Normais;
- assintoticamente centrados;
- assintoticamente de matriz de variâncias igual à inversa da matriz de informação de Fisher associada à estimação.

Aplicando estes resultados gerais aos estimadores $\hat{\boldsymbol{\beta}}$ obtidos pelo Método da Máxima Verosimilhança nos Modelos Lineares Generalizados, obtém-se directamente o seguinte resultado.

Teorema 4.1 *Dado um Modelo Linear Generalizado e admitindo as condições de regularidade necessárias, tem-se que os estimadores de Máxima Verosimilhança $\hat{\boldsymbol{\beta}}$ dos* p+1 *parâmetros da componente sistemática do Modelo verificam,* ***assintoticamente****:*

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}_{(p+1)}\left(\boldsymbol{\beta}, \mathbf{I}_{\boldsymbol{\beta}}^{-1}\right) \tag{4.41}$$

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

De forma igualmente directa, resultam as seguintes consequências do Teorema anterior, que serão de grande importância na inferência associada aos MLGs.

Teorema 4.2 *Dado um Modelo Linear Generalizado e admitindo as condições de regularidade necessárias, tem-se que os estimadores de Máxima Verosimilhança $\hat{\boldsymbol{\beta}}$ dos parâmetros da componente sistemática do Modelo verificam,* ***assintoticamente****:*

1. $\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^t \left(\mathbf{I}_{\boldsymbol{\beta}}\right) \left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \sim {\chi^2}_{(p+1)}$
2. *Se* $\mathbf{C}$ *fôr uma matriz não-aleatória, de tipo* $q \times (p+1)$*, tem-se* $\mathbf{C}\hat{\boldsymbol{\beta}} \sim \mathcal{N}_q\left(\mathbf{C}\boldsymbol{\beta}, \mathbf{C}\mathbf{I}_{\boldsymbol{\beta}}^{-1}\mathbf{C}^t\right)$.
3. *se* $\mathbf{a}$ *fôr um vector não-aleatório* $(p+1)$*-dimensional, tem-se:* $\frac{\mathbf{a}^t\hat{\boldsymbol{\beta}}-\mathbf{a}^t\boldsymbol{\beta}}{\sqrt{\mathbf{a}^t\mathbf{I}_{\boldsymbol{\beta}}^{-1}\mathbf{a}}} \sim \mathcal{N}(0,1)$.
4. *Se* $\mathbf{C}$ *fôr uma matriz real* $q \times (p+1)$*, tem-se:* $\left(\mathbf{C}\hat{\boldsymbol{\beta}}-\mathbf{C}\boldsymbol{\beta}\right)^t \left[\mathbf{C}\mathbf{I}_{\boldsymbol{\beta}}^{-1}\mathbf{C}^t\right]^{-1} \left(\mathbf{C}\hat{\boldsymbol{\beta}}-\mathbf{C}\boldsymbol{\beta}\right) \sim {\chi^2}_q$

Estes resultados são de demonstração imediata a partir do Teorema anterior e dos resultados gerais sobre vectores aleatórios Multinormais já estudados na disciplina de Complementos de Probabilidades e Estatística e revistos na disciplina de Modelação Estatística I. Os resultados fornecem a base para a construção de Intervalos de Confiança e Testes de Hipóteses a valores de combinações lineares dos parâmetros $\boldsymbol{\beta}$, assim como para testes de ajustamento de Modelos e de Submodelos, como se verá nas Subsecções seguintes.

4.5.1 Inferência sobre combinações lineares dos parâmetros

A Multinormalidade assintótica dos estimadores de Máxima Verosimilhança permite construir Intervalos de Confiança e Testes de Hipóteses assintóticos para qualquer combinação linear dos parâmetros $\boldsymbol{\beta}$. Tal possibilidade é útil, tendo em conta a já referida possibilidade de interpretar os valores dos parâmetros.

A derivação destes resultados é análoga ao que foi visto no caso do Modelo Linear. Assinale-se que na expressão surge a inversa da matriz de informação no ponto desconhecido $\boldsymbol{\beta}$. A fim de poder prosseguir no cálculo de Intervalos de Confiança e/ou na realização de Testes de Hipóteses, *essa matriz desconhecida é substituída por outra, conhecida, que consiste na matriz de informação calculada para a* estimativa $\hat{\boldsymbol{\beta}}$. Esta substituição reforça a necessidade de grandes amostras para que se possa confiar nos resultados.

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

$$\left] \ \mathbf{a}^t\hat{\boldsymbol{\beta}} - z_{\frac{\alpha}{2}} \cdot \sqrt{\mathbf{a}^t \mathbf{I}_{\hat{\beta}}^{-1} \mathbf{a}} \quad , \quad \mathbf{a}^t\hat{\boldsymbol{\beta}} + z_{\frac{\alpha}{2}} \cdot \sqrt{\mathbf{a}^t \mathbf{I}_{\hat{\beta}}^{-1} \mathbf{a}} \ \right[\tag{4.42}$$

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

Exemplo de Teste de Hipóteses (assintótico) a uma Combinação Linear dos β_j

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

Estatística do Teste: $Z = \dfrac{\mathbf{a}^t\hat{\boldsymbol{\beta}} - \mathbf{a}^t\boldsymbol{\beta}_{|H_0}}{\sqrt{\mathbf{a}^t \mathbf{I}_{\hat{\beta}}^{-1} \mathbf{a}}} \sim \mathcal{N}(0,1),$

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

A estatística de Wald

Para testar em simultâneo hipóteses sobre várias combinações lineares dos parâmetros $\boldsymbol{\beta}$, podemos usar a estatística de Wald, baseada no resultado do ponto 4 do Teorema 4.2, mas em que se substitui a matriz desconhecida $\mathbf{I}_\beta$ por $\mathbf{I}_{\hat{\beta}}$.

Teste de Hipóteses (assintótico) simultâneo a várias Combinações Lineares dos β_j

Seja $\mathbf{C}$ uma matriz não-aleatória $q \times (p+1)$, de característica q. Seja ξ um vector q-dimensional.

Hipóteses: $H_0 : \mathbf{C}\boldsymbol{\beta} = \xi \quad vs. \quad H_1 : \mathbf{C}\boldsymbol{\beta} \neq \xi$

Estatística do Teste: $\chi^2 = \left(\mathbf{C}\hat{\boldsymbol{\beta}} - \xi\right)^t \left[\mathbf{C}\mathbf{I}_{\hat{\beta}}^{-1}\mathbf{C}^t\right]^{-1} \left(\mathbf{C}\hat{\boldsymbol{\beta}} - \xi\right) \sim {\chi^2}_q,$

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

Uma das aplicações deste tipo de Testes permite testar se determinado Submodelo, em que apenas se considera um subconjunto das variáveis preditoras, é admissível. Esta aplicação é considerada na próxima Subsecção.

4.5.2 Testes a Submodelos

Dado um MLG com $p+1$ variáveis preditoras $X_0, X_1, ..., X_p$, é legítimo perguntar, tal como se fez no estudo do Modelo Linear, se é possível simplificar o Modelo através da exclusão de algumas das variáveis preditoras, sem com isso afectar de forma significativa o ajustamento do mesmo aos dados.

Seja S um subconjunto de $k+1$ índices das variáveis preditoras. O Submodelo obtido excluíndo as variáveis cujos índices não pertencem a S pode ser definido através das $p-k$ condições $\beta_j = 0, \forall j \notin S$. O conjunto destas restrições pode ser escrito, em forma matricial, como $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, onde $\mathbf{C}$ é uma matriz de tipo $(p-k)\times(p+1)$, cujas linhas são as $p-k$ linhas da matriz identidade $(p+1)\times(p+1)$ associadas às $p-k$ variáveis que não pertencem ao subconjunto S, isto é, cuja eventual exclusão do Modelo se deseja testar.

Teste de Wald

A estatística de Wald (introduzida no ponto anterior) pode ser usada para testar se o Submodelo com apenas $k+1$ variáveis preditoras é considerado admissível.

A fim de simplificar a notação, designemos por $\boldsymbol{\beta}_{\overline{S}}$ (e $\hat{\boldsymbol{\beta}}_{\overline{S}}$) o subvector de $\boldsymbol{\beta}$ (e $\hat{\boldsymbol{\beta}}$) associado às componentes de $\boldsymbol{\beta}$ cujos índices não pertencem a S (isto é, cuja exclusão se considera). Seja ainda $\mathbf{A}_{(\overline{S},\overline{S})}$ a submatriz principal duma matriz $\mathbf{A}$ associada a tomar as linhas e colunas de $\mathbf{A}$ cujos índices não pertencem ao conjunto S. Então, tem-se:

Teste de Hipóteses a Submodelo de um MLG

Hipóteses:

$$\begin{array}{lll} H_0 : \beta_j = 0, \quad \forall j \notin S & vs. & H_1 : \exists j \notin S, \text{ t.q. } \beta_j \neq 0 \\ \iff H_0 : \boldsymbol{\beta}_{\overline{S}} = \mathbf{0} & vs. & H_1 : \boldsymbol{\beta}_{\overline{S}} \neq \mathbf{0} \\ \iff \text{(Submodelo admissível)} & vs. & \text{(Submodelo não admissível)} \end{array}$$

Estatística do Teste:

$$\chi^2 = \hat{\boldsymbol{\beta}}_{\overline{S}}^t \left[\left(\mathbf{I}_{\hat{\beta}}^{-1} \right)_{(\overline{S},\overline{S})} \right]^{-1} \hat{\boldsymbol{\beta}}_{\overline{S}} \sim {\chi^2}_{p-k},$$

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

Note-se que a inversa duma submatriz não é igual à submatriz correspondente da inversa da matriz completa, pelo que não é possível simplificar ulteriormente a matriz que define a forma quadrática nesta estatística. A natureza unilateral direita da Região Crítica é fácil de compreender neste contexto: a ser verdade H_0, é de esperar que as componentes do vector estimado $\hat{\boldsymbol{\beta}}_{\overline{S}}$ sejam todas elas próximas de zero, contribuíndo assim para manter o valor de ${\chi^2}_{calc}$ também próximo de zero. Valores muito elevados deste valor calculado da estatística do Teste sugerem que pelo menos um dos β_j ($j \notin S$) seja diferente de zero.

Teste da razão de Verosimilhanças (Wilk)

Um outro teste à admissibilidade de um Submodelo pode ser obtido com base num resultado geral, dado na disciplina de Complementos de Probabilidades e Estatística deste Mestrado: o Teorema de Wilk. Recordem-se os resultados fundamentais a este respeito, extraídos das folhas da Prof. Manuela Neves de CPE (e **advirta-se que, neste contexto, θ indica um conjunto genérico de parâmetros, e não tem o significado específico do contexto duma família exponencial de distribuições**).

Definição 4.3 *Seja $(X_1, X_2, ..., X_n)$ uma amostra aleatória. Seja $L(\theta|\mathbf{x})$ a sua função verosimilhança. Designa-se* ***razão de verosimilhanças*** *a:*

$$L_n(\mathbf{x}) = \frac{\max\limits_{\theta\in\Theta_0} L(\theta|\mathbf{x})}{\max\limits_{\theta\in(\Theta_0\cup\Theta_1)} L(\theta|\mathbf{x})} \tag{4.43}$$

onde Θ_0 e Θ_1 designam dois conjuntos alternativos de condições sobre os valores dos parâmetros θ.

A razão de verosimilhanças, ou melhor, a sua transformação $\Lambda = -2\log(L_n)$ é utilizada como estatística de um teste às hipóteses:

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

O Teorema de Wilk garante que, sob certas condições de regularidade da função de verosimilhança, Λ tem distribuição ${\chi^2}_q$ sob H_0, onde q indica o número de restrições impostas aos parâmetros em H_0. A Região Crítica associada a este Teste é de tipo unilateral direito. Resumindo:

Teste de Razão de Verosimilhanças (contexto geral)

Hipóteses: $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$

Estatística do Teste: $\Lambda = -2\left(\max\limits_{\theta\in\Theta_0} \ell(\theta;\mathbf{x}) - \max\limits_{\theta\in(\Theta_0\cup\Theta_1)} \ell(\theta;\mathbf{x})\right) \sim {\chi^2}_q,$

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

Consideremos agora o contexto de um Modelo Linear Generalizado, onde os parâmetros θ são os $p+1$ coeficientes $\boldsymbol{\beta}$ da combinação linear que constitui a componente sistemática do Modelo. Designemos por Θ_0 os valores dos parâmetros resultantes de impôr a restrição $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, ou seja, $\boldsymbol{\beta}_{\overline{S}} = \mathbf{0}$[10]. Por outras palavras, Θ_0 fica associada ao Submodelo que exclui as $p-k$ variáveis cujos índices não pertencem ao conjunto S. Por Θ_1 indica-se a condição complementar de que pelo menos um desses parâmetros $\boldsymbol{\beta}_{\overline{S}}$ seja diferente de zero.

Os valores dos parâmetros que determinam o máximo da função de log-verosimilhanças para $\theta \in (\Theta_0 \cup \Theta_1)$ são as estimativas de Máxima Verosimilhança do Modelo Completo (sem restrições nos valores dos parâmetros). Por seu lado, os valores dos parâmetros que determinam o máximo da função de log-verosimilhanças para $\theta \in \Theta_0$ são as estimativas de Máxima Verosimilhança do Submodelo, $\hat{\boldsymbol{\beta}}_S$. Assim,

[10] Outras restrições aos valores de $\boldsymbol{\beta}$ são possíveis, mas consideram-se aqui as restrições associadas ao estudo da admissibilidade de um Submodelo.

para testar a admissibilidade de um Submodelo utilizando a Estatística de Wilk, e designando por ℓ_M a log-verosimilhança associada ao Modelo completo ($p+1$ preditores, estimados por $\hat{\boldsymbol{\beta}}_M$) e por ℓ_S a log-verosimilhança associada ao Submodelo ($k+1$ preditores, estimados por $\hat{\boldsymbol{\beta}}_S$) tem-se:

Teste de Wilk a Submodelos

Hipóteses:
$$\begin{array}{lll} H_0: \beta_j = 0, \quad \forall j \notin S & vs. & H_1: \exists j \notin S, \text{ t.q. } \beta_j \neq 0 \\ \iff H_0: \boldsymbol{\beta}_{\overline{S}} = \mathbf{0} & vs. & H_1: \boldsymbol{\beta}_{\overline{S}} \neq \mathbf{0} \\ \iff \text{(Submodelo admissível)} & vs. & \text{(Submodelo não admissível)} \end{array}$$

Estatística do Teste: $\Lambda = -2\left(\ell_S(\hat{\boldsymbol{\beta}}_S) - \ell_M(\hat{\boldsymbol{\beta}}_M)\right) \sim {\chi^2}_{p-k}$,

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

Mais adiante se verá que é possível escrever a estatística deste Teste numa forma alternativa. Mas para isso, será necesário introduzir o conceito de **Desvio** de um Modelo.

4.6 Desvio e Desvio Reduzido

No estudo de Modelos Lineares Generalizados surge um conceito que desempenha um papel análogo ao desempenhado pela Soma de Quadrados Residual nos Modelos Lineares, o conceito de Desvio[11] de um Modelo ajustado.

Comecemos por introduzir um conceito auxiliar. 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 aditiva e em que 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 chamado Modelo Saturado.

Um **Modelo com tantos parâmetros quantas as observações de Y disponíveis** diz-se um **Modelo Saturado**. Num Modelo Saturado, o ajustamento será "perfeito", mas de pouca utilidade. De facto, **num Modelo Linear Generalizado Saturado, a estimativa de cada valor esperado de Y coincide totalmente com o valor observado de Y correspondente, isto é, $\hat{E[Y_i]} = Y_i$**. Por exemplo, 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 função log-verosimilhança toma a forma $\mathbf{Xy} = \mathbf{X}\hat{\mu}$, onde $\hat{\mu}$ indica o vector estimado de valores esperados de Y para as n observações (ver equações 4.22 e 4.25, página 26). Se o Modelo é Saturado, existindo tantos parâmetros quantas observações, a matrix $\mathbf{X}$ é de tipo $n \times n$ e, em geral, invertível. Nesse caso, $\mathbf{y} = \hat{\mu}$. Assim, o Modelo Saturado ocupa o polo oposto em relação ao Modelo Nulo (já considerado no contexto do Modelo Linear): enquanto que neste último tudo é variação residual, não explicada pelo Modelo, no Modelo Saturado tudo é "explicado" pelo Modelo, não havendo lugar a variação residual.

Um tal ajustamento "total" dos dados ao Modelo é, na realidade, ilusório: o Modelo ajustado não terá

[11] *Deviance* em inglês.

robustez suficiente para se adaptar a outra amostra dos mesmo tipo de dados, e ao procurar acompanhar cada árvore, facilmente se perde a visão de conjunto da floresta. Neste sentido, está-se numa situação análoga à que existe quando se procura ajustar um polinómio de grau $n-1$ a n observações $\{(x_i, y_i)\}_{i=1}^n$.

Mas se um Modelo Saturado é uma situação indesejável na prática, é de utilidade como *termo de comparação* para medir o grau de ajustamento de um conjunto de dados a um MLG. É nessa ideia que se baseia a definição dos conceitos de *Desvio Reduzido* e *Desvio*, que a seguir se introduz.

Definição 4.4 *Considere-se um Modelo Linear Generalizado baseado em n observações independentes da variável resposta Y. Seja $\hat{\boldsymbol{\beta}}_M$ o vector estimado dos seus parâmetros e $\ell_M(\hat{\boldsymbol{\beta}}_M)$ a respectiva log-verosimilhança máxima. Considere-se um Modelo Saturado com n parâmetros. Designe-se por $\ell_T(\hat{\boldsymbol{\beta}}_T)$ a log-verosimilhança correspondente a tomar $E[\hat{Y}_i] = Y_i,\ \forall i = 1:n$. Define-se o* ***desvio reduzido*** *como sendo:*

$$D^* = -2\left(\ell_M(\hat{\boldsymbol{\beta}}_M) - \ell_T(\hat{\boldsymbol{\beta}}_T)\right) \tag{4.44}$$

Nesta definição, o Desvio Reduzido já não será, em geral, função do parâmetro θ da família exponencial de distribuições (já que a estimação de $\boldsymbol{\beta}$ fornece uma estimação desse parâmetro). Mas pode ainda ser função do outro parâmetro, ϕ, das distribuições na família exponencial bi-paramétrica de distribuições. Nesse caso, o Desvio Reduzido não é directamente calculável só com base nos dados, sendo necessária uma de duas coisas: (i) o conhecimento do valor de ϕ; ou (ii) a sua estimação.

Mas, em grande número de situações, existe uma terceira possibilidade. De facto, com frequência é possível admitir MLGs para os quais o parâmetro ϕ (muitas vezes associado à dispersão dos valores de Y) é constante nas n observações[12] e nos quais se verifica também outra condição adicional: $a(\phi) = \frac{\phi}{w_i}$ para algum conjunto conhecido de valores $\{w_i\}_{i=1}^n$. Por exemplo, nos Modelos considerados nesta disciplina para variáveis resposta Bernoulli tem-se (ver página 14): $\phi = 1$ e $a(\phi) = 1$, correspondendo a admitir que $w_i = 1, \forall i = 1:n$. Precisamente as mesmas opções correspondem à distribuição Poisson, e portanto à Regressão Log-linear. Neste caso, um conceito alternativo é de maior utilidade prática, o conceito de *Desvio*. Já no que respeita à distribuição "Binomial/n", tem-se (ver página 14): $\phi = 1$ e $a(\phi) = \frac{1}{n}$, correspondendo a admitir que $w_i = \frac{1}{n_i}, \forall i = 1:n$.

Definição 4.5 *Considere-se um Modelo Linear Generalizado com $a(\phi_i) = \frac{\phi}{w_i},\ \forall i = 1:n$, sendo ϕ e $\{w_i\}_{i=1}^n$ constantes. Utilizando as mesmas notações que na definição 4.4 e indicando as estimativas dos θ em cada Modelo através das letras M (Modelo) e T (Modelo Saturado), designa-se* ***Desvio*** *a:*

$$D \;=\; \phi \cdot D^* \;=\; -2\left(\ell_M(\hat{\boldsymbol{\beta}}_M) - \ell_T(\hat{\boldsymbol{\beta}}_T)\right) \cdot \phi \;=\; \sum_{i=1}^{n} 2w_i\left(y_i(\hat{\theta}_i^T - \hat{\theta}_i^M) - b(\hat{\theta}_i^T) + b(\hat{\theta}_i^M)\right) \tag{4.45}$$

Assinale-se que o Desvio assim definido já não depende de parâmetros desconhecidos, e pode ser calculado para qualquer conjunto de dados.

A expressão dos Desvios para as distribuições consideradas nos Modelos referidos nesta disciplina resultam ser os seguintes:

[12] Esta é, por exemplo, a hipótese de variâncias homogéneas feita no Modelo Linear, caso em que o parâmetro $\phi = \sigma^2$.

Distribuição Normal Na distribuição Normal, como já se viu (página 13) tem-se $\theta = \mu$, $b(\theta) = \frac{\mu^2}{2}$, $\phi = a(\phi) = \sigma^2$, pelo que $w_i = 1, \forall i = 1 : n$. Como para o Modelo Saturado se tem $\hat{\mu}_i^T = y_i$, a expressão do Desvio (equação 4.45) vem (escrevendo $\hat{\mu}_i^M$ sem a letra M, que deixa de ser necessária):

$$\begin{aligned} D &= 2\sum_{i=1}^{n} w_i \left[y_i \left(y_i - \hat{\mu}_i\right) - \frac{y_i^2}{2} + \frac{\left(\hat{\mu}_i\right)^2}{2} \right] \\ &= \sum_{i=1}^{n} \left[y_i^2 - 2 y_i \hat{\mu}_i + \hat{\mu}_i^2 \right] \\ &= \sum_{i=1}^{n} (y_i - \hat{u}_i)^2 \end{aligned} \tag{4.46}$$

Como no caso do Modelo Linear a função de ligação é a identidade, isto é, $\hat{\mu}_i = \mathbf{x}_i^t \hat{\boldsymbol{\beta}}$, tem-se que *o Desvio no caso do Modelo Linear não é mais que a Soma de Quadrados Residual, SQRE*. Nesse caso, *o Desvio Reduzido é* $\frac{SQRE}{\sigma^2}$, ou seja, a quantidade que, na disciplina de Modelação Estatística I, se verificou possuir distribuição ${\chi^2}_{n-(p+1)}$.

Distribuição de Bernoulli Para MLGs com variável resposta de distribuição Bernoulli, tem-se (ver página 14): $\theta = \log\left(\frac{p}{1-p}\right)$, $b(\theta) = -\log(1-p)$, $\phi = a(\phi) = 1$ e $w_i = 1, \forall i = 1 : n$. Como para o Modelo Saturado se tem $\hat{p}_i^T = y_i$, a expressão para $\hat{\theta}_i^T$ resulta numa divisão por zero quando $y_i = 1$. No entanto, esse parâmetro aparece apenas no contexto da expressão $y_i\hat{\theta}_i^T - b(\hat{\theta}_i^T) = y_i \log(y_i) + y_i \log(1-y_i) - \log(1-y_i) = y_i \log(y_i) - (1-y_i)\log(1-y_i)$. Ora, quer para $y_i = 0$, quer para $y_i = 1$, ambas as parcelas desta diferença se podem considerar nulas (verifique, lembrando que polinómios dominam logaritmos na velocidade de convergência). Assim, (e designando apenas por $\hat{p}_i$ a estimativa de p para o Modelo não Saturado) o Desvio vem:

$$\begin{aligned} D &= 2\sum_{i=1}^{n} \left[-y_i \log(\hat{p}_i) + y_i \log(1-\hat{p}_i) - \log(1-\hat{p}_i) \right] \\ &= -2\sum_{i=1}^{n} \left[y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i) \right] \end{aligned} \tag{4.47}$$

Note-se que *o Desvio Reduzido para a distribuição de Bernoulli é igual ao Desvio*.

Distribuição Binomial/n Para MLGs com variável resposta de distribuição Binomial/n, tem-se (ver página 14): $\theta = \log\left(\frac{p}{1-p}\right)$, $b(\theta) = -\log(1-p)$, $\phi = 1$ e $a(\phi_i) = \frac{\phi}{n_i} = \frac{1}{n_i}$, que é da forma $a(\phi_i) = \frac{\phi}{w_i}$ com $\phi = 1$ e $w_i = n_i, \forall i = 1 : m$. Tal como para o caso duma variável resposta Bernoulli, o Modelo Saturado gera $\hat{p}_i^T = y_i$. Tem-se $y_i\hat{\theta}_i^T - b(\hat{\theta}_i^T) = y_i \log(y_i) + y_i \log(1-y_i) - \log(1-y_i) = y_i \log(y_i) - (1-y_i)\log(1-y_i)$ e $y_i\hat{\theta}_i^M - b(\hat{\theta}_i^M) = y_i \log(\hat{p}_i) + y_i \log(1-\hat{p}_i) - \log(1-\hat{p}_i) = y_i \log(\hat{p}_i) - (1-y_i)\log(1-\hat{p}_i)$. Substituindo na expressão geral (4.45) da página 37, obtém-se o Desvio:

$$\begin{aligned} D &= 2\sum_{i=1}^{m} n_i \left[y_i \log\left(\frac{y_i}{1-y_i}\right) - \log(1-y_i) - y_i \log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) + \log(1-\hat{p}_i) \right] \\ &= 2\sum_{i=1}^{m} n_i \left[y_i \log(y_i) + (1-y_i)\log(1-y_i) - y_i \log(\hat{p}_i) - (1-y_i)\log(1-\hat{p}_i) \right] \end{aligned} \tag{4.48}$$

Recorde-se que neste contexto, os valores observados y_i $(i = 1 : m)$ são proporções de êxitos em n_i provas, e já não apenas valores do tipo 0 ou 1. No entanto, é possível que alguns dos valores y_i sejam apenas 0 ou 1. *Tal como para o caso duma distribuição Bernoulli, no caso de y_i tomar valores $y_i = 0, 1$, deve considerar-se que a soma $y_i \log(y_i) + (1 - y_i) \log(1 - y_i)$ na expressão acima referida é igual a zero na parcela correspondente a cada uma dessas observações.*

Note-se que *o Desvio Reduzido para a distribuição Binomial/n é igual ao Desvio*, uma vez que $\phi = 1$.

Distribuição Poisson Para MLGs com variável resposta com distribuição Poisson, tem-se (ver página 13): $\theta = \log(\lambda)$, $b(\theta) = \lambda$, $\phi = a(\phi) = 1$ e $w_i = 1, \forall i = 1 : n$. Como no Modelo Saturado $\hat{\lambda}_i^T = y_i$, a expressão do Desvio vem (escrevendo $\hat{\lambda}_i^M$ apenas como $\hat{\lambda}_i$):

$$\begin{aligned} D &= 2\sum_{i=1}^{n} \left[y_i \left(\log(y_i) - \log(\hat{\lambda}_i) \right) - y_i + \hat{\lambda}_i \right] \\ &= 2\sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{\hat{\lambda}_i} \right) - (y_i - \hat{\lambda}_i) \right] \end{aligned} \tag{4.49}$$

Note-se que *para a distribuição Poisson, Desvio Reduzido e Desvio coincidem*.

Distribuição Gama Para MLGs com variável resposta com distribuição Gama, tem-se (ver página 15): $\theta = \frac{1}{\mu} = \frac{1}{\alpha\beta}$, $b(\theta) = -\log(-\theta) = \log(\alpha\beta) = \log(\mu)$, $\phi = \frac{1}{\alpha}$ e $a(\phi) = \phi = \frac{1}{\alpha}$. Como indicado mais acima, *vamos admitir que para diferentes observações se tem $a(\phi_i) = \frac{\phi}{w_i}$, para algum conjunto de constantes w_i*. Como no Modelo Saturado $\hat{\mu}_i^T = \frac{1}{y_i}$, a expressão do Desvio vem (escrevendo $\hat{\mu}_i^M$ apenas como $\hat{\mu}_i$):

$$\begin{aligned} D &= 2\sum_{i=1}^{n} \left[y_i \cdot \left(\frac{-1}{y_i} \right) - \log(y_i) + y_i \cdot \frac{1}{\hat{\mu}_i} + \log(\hat{\mu}_i) \right] \\ &= 2\sum_{i=1}^{n} \left[\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} - \log\left(\frac{y_i}{\hat{\mu}_i} \right) \right] \end{aligned} \tag{4.50}$$

Note-se que *para a distribuição Gama, Desvio Reduzido e Desvio **não** coincidem*, uma vez que $\phi \neq 1$. Deixamos para depois o problema de como estimar ϕ.

4.6.1 Teste de Ajustamento do Modelo e Teste a Submodelos Encaixados

Os conceitos de Desvio e Desvio Reduzido desempenham um papel importante no estudo da qualidade do ajustamento de dados a um Modelo Linear Generalizado. A razão fundamental da sua utilidade reside no facto de a estatística de Teste utilizada no Teste de Wilk a Modelos Encaixados (página 4.5.2) ser apenas *a diferença dos Desvios Reduzidos de Modelo e Submodelo*. Assim, e embora *em geral a distribuição de um Desvio ou Desvio Reduzido não seja conhecida*[13], *a distribução da diferença de Desvios de dois Modelos Encaixados é assintoticamente $\chi^2{}_{p-k}$* onde $p - k$ indica a diferença no número de variáveis preditoras do Modelo e do seu Submodelo. Assim, o Teste de Wilk a Modelos Encaixados pode ser re-enunciado da seguinte forma.

[13] Para Modelos particulares pode ser, como foi visto no caso do Modelo Linear, para o qual o Desvio Reduzido tem distribuição (exactamente, não assintoticamente) χ^2.

Teste de Wilk a Submodelos Encaixados

Hipóteses: $\Leftrightarrow$ $H_0 : \beta_j = 0, \quad \forall j \notin S \quad vs. \quad H_1 : \exists j \notin S, \text{ t.q. } \beta_j \neq 0$
$H_0 : \boldsymbol{\beta}_{\overline{S}} = \mathbf{0} \quad vs. \quad H_1 : \boldsymbol{\beta}_{\overline{S}} \neq \mathbf{0}$
(Submodelo admissível) $vs.$ (Submodelo não admissível)

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

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

Embora os Desvios Reduzidos dependam, em geral do parâmetro desconhecido ϕ, **para vários dos MLGs considerados nesta disciplina, em que as variáveis resposta têm distribuições Bernoulli ou Poisson, a estatística resulta ser apenas a diferença dos Desvios**, que são totalmente conhecidos com base nos dados observados.

Teste de Ajustamento de um Modelo

No caso de Modelos Lineares Generalizados cuja componente sistemática inclui uma parcela aditiva constante, o conceito de ajustamento global do Modelo pode ser semelhante ao que já foi considerado aquando do 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 $\mathbf{x}_i^t \boldsymbol{\beta} = \beta_0$, $\forall i = 1 : n$. Nesse caso, 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 não se ajustar de forma significativamente diferente do Modelo sob estudo, é caso para concluir pela inutilidade do Modelo.

Teste de Wilk ao Ajustamento de um MLG

Hipóteses: $H_0 : \beta_j = 0, \quad \forall j = 1 : p \quad vs. \quad H_1 : \exists j = 1 : p, \text{ t.q. } \beta_j \neq 0$
(Modelo Nulo) $vs.$ (Modelo Nulo não admissível)

Estatística do Teste: $\Lambda = D^*_N - D^*_M \sim {\chi^2}_p,$

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

No Quadro anterior, D^*_N indica o Desvio Reduzido do Modelo Nulo. De novo, recorde-se que para MLGs com Componente Aleatória Poisson ou Bernoulli, Desvios Reduzidos e Desvios coincidem, pelo que a Estatística de Teste anterior é dada apenas pela diferença dos Desvios.

Selecção de Submodelos

Tal como sucede no estudo do Modelo Linear, a escolha de um Submodelo adequado, que simplifique um Modelo com um grande número de variáveis preditoras, pode ser determinado por considerações de diversa ordem. Caso considerações de natureza extra-estatística sugiram um determinado Submodelo, a questão da sua admissibilidade pode ser testada através do Teste de Wilk a Submodelos Encaixados, como já foi visto. No caso de não haver ideia prévia de qual Submodelo propôr, a pesquisa completa da admissibilidade dos $2^p - 2$ possíveis Submodelos (com $k+1$ variáveis preditoras, para qualquer $k = 1 : p-1$, e admitindo a existência de uma constante aditiva) coloca as mesmas dificuldades computacionais já consideradas no estudo do Modelo Linear, sempre que o valor de p não seja baixo. Nesses casos, **será possível utilizar algoritmos de escolha sequenciais semelhantes aos que foram considerados aquando do estudo do Modelo Linear. Métodos de exclusão ou inclusão sequenciais, ou métodos em que se alternam passos nos dois sentidos são possíveis, mas adoptando agora como critério para a inclusão/exclusão de variáveis num subconjunto preditor a maior/menor redução no Desvio que produzem, desde que esta seja significativa**. Por exemplo, um método de exclusão sequencial consistiria em (iniciando com o Modelo Completo de $p+1$ variáveis preditoras, incluíndo constante aditiva) verificar qual das variáveis preditoras do Submodelo corrente provocaria, pela sua exclusão do Modelo, um menor acréscimo do Desvio, e procedendo à sua exclusão, desde que esse menor acréscimo não fosse significativo, isto é, desde que o Submodelo assim resultante fosse considerado, no Teste de Wilk a Submodelos encaixados, como não diferindo significativamente do Submodelo corrente.

No programa estatístico R, 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. Para mais informações, consultar os testos de informação relativos a estes comandos, constantes no programa R (através do comando `help(drop1)`, por exemplo).

4.7 Algumas questões

Nesta Secção serão abordadas algumas questões importantes que aprofundam o conhecimento dos Modelos até aqui apresentados.

4.7.1 Esperança e Variância na família exponencial de distribuições

Embora não tenha sido necessário até este momento, é possível obter expressões gerais para o Valor Esperado e a Variância de variáveis aleatórias com distribuição na família exponencial de distribuições.

Para tal, recordem-se dois resultados gerais, dados na disciplina de Complementos de Probabilidades e Estatística, e relativos a propriedades de derivadas parciais da função log-verosimilhança. Embora estes resultados tenham sido dados para funções de verosimilhança com um único parâmetro, a sua generalização para funções de verosimilhança com mais do que um parâmetro é directa. Para mais pormenores, veja-se, por exemplo, o livro de M. Kendall e A. Stuart, *The Advanced Theory of Statistics*, Volume 2, 3a. edição, 1951 (página 54 e seguintes).

Teorema 4.3 *Seja $\ell(\theta; \mathbf{y})$ uma função de log-verosimilhança de observações com um vector de parâmetros $\theta = (\theta_1, \theta_2, ..., \theta_q)$. Então, sob condições de regularidade suficientes, verifica-se:*

1. $E\left[\frac{\partial \ell}{\partial \theta_r}\right] = 0, \quad \forall r$
2. $E\left[\left(\frac{\partial \ell}{\partial \theta_r}\right)^2\right] = -E\left[\frac{\partial^2 \ell}{\partial \theta_r^2}\right], \quad \forall r$

Neste enunciado, θ indica um vector genérico de parâmetros, e não tem o signficado mais específico de θ na distribuição genérica da família (bi-dimensional) exponencial de distribuições. Recorde-se que, nessa família, a log-verosimilhança é da forma (4.18):

$$\ell(\theta, \phi \ ; \ Y_1, Y_2, ..., Y_n) = \sum_{i=1}^{n} \left[\frac{Y_i \theta_i - b(\theta_i)}{a(\phi_i)} + c(Y_i, \phi_i)\right]$$

Tomando derivadas parciais em relação a θ_i, tem-se:

$$\frac{\partial \ell}{\partial \theta_i} = \frac{Y_i - b'(\theta_i)}{a(\phi_i)} \tag{4.51}$$

e

$$\frac{\partial^2 \ell}{\partial \theta_i^2} = \frac{-b''(\theta_i)}{a(\phi_i)} \tag{4.52}$$

Da primeira alínea do Teorema 4.3 resulta:

$$E[Y_i - b'(\theta_i)] = 0 \iff E[Y_i] = b'(\theta_i) \tag{4.53}$$

Da segunda alínea desse mesmo Teorema resulta:

$$\begin{aligned} E\left[\left(\frac{Y_i - b'(\theta_i)}{a(\phi_i)}\right)^2\right] = -E\left[-\frac{b''(\theta_i)}{a(\phi_i)}\right] &\iff \frac{V[Y_i]}{a^2(\phi_i)} = \frac{b''(\theta_i)}{a(\phi_i)} \\ &\iff V[Y_i] = b''(\theta_i) \cdot a(\phi_i) \end{aligned} \tag{4.54}$$

Recordando as expressões para as funções $a(\cdot)$ e $b(\cdot)$ que enquadravam as distribuições Normal, Bernoulli, Binomial/n, Poisson e Gama no contexto da família exponencial (ver Subsecção 4.1.1, a partir da página 12), podemos confirmar os valores conhecidos para as esperanças e variâncias destes cinco membros da família exponencial de distribuições:

Normal Como $b(\theta) = \frac{\theta^2}{2}$, tem-se $b'(\theta) = \theta$ e $b''(\theta) = 1$. Como $\theta = \mu$ e $a(\phi) = \sigma^2$, resulta que $E[Y_i] = \mu_i$ e $V[Y_i] = \sigma_i^2$, o que corresponde à hipótese de Normalidade dos Y_i (tendo-se, no Modelo Linear, ainda a igualdade de variâncias $\sigma_i^2 = \sigma^2, \ \forall i$).

Poisson Como $b(\theta) = \mathrm{e}^\theta$, tem-se $b'(\theta) = b''(\theta) = \mathrm{e}^\theta$. Como $\theta = \log(\lambda)$ e $a(\phi) = 1$, temos: $E[Y_i] = \lambda_i$ e $V[Y_i] = \lambda_i$, como seria de esperar se $Y_i \cap \mathrm{Po}(\lambda_i)$.

Bernoulli Como $b(\theta) = \log(1 + \mathrm{e}^\theta)$, tem-se $b'(\theta) = \frac{\mathrm{e}^\theta}{1+\mathrm{e}^\theta}$ e $b''(\theta) = \frac{\mathrm{e}^\theta}{(1+\mathrm{e}^\theta)^2}$. Como $\theta = \log\left(\frac{p}{1-p}\right)$ e $a(\phi) = 1$, tem-se: $E[Y_i] = p_i$ e $V[Y_i] = p_i(1 - p_i)$, como é próprio de uma Bernoulli de parâmetro p_i.

Binomial/n As expressões são idênticas às da Bernoulli, excepto a função $a(\phi) = \frac{1}{n}$. Assim, o valor esperado é igual ($E[Y_i] = p_i$), sendo a variância dada por $V[Y_i] = \frac{p_i(1-p_i)}{n_i}$. Se Y_i é da forma $\frac{X_i}{n_i}$, onde $X_i \cap B(n_i, p_i)$, os conhecimentos relativos à distribuição Binomial permitiam-nos afirmar que $E[Y_i] = \frac{E[X_i]}{n_i} = \frac{n_i p_i}{n_i} = p_i$, e $V[Y_i] = \frac{V[X_i]}{n_i^2} = \frac{n_i p_i (1-p_i)}{n_i^2}$, que simplifica para a expressão obtida acima.

Gama Como $b(\theta) = -\log(-\theta)$, tem-se $b'(\theta) = -\frac{1}{\theta}$ e $b''(\theta) = \frac{1}{\theta^2}$. Como $\theta = -\frac{1}{\mu} = -\frac{1}{\alpha\beta}$ e $a(\phi) = \frac{1}{\alpha}$, resulta que $E[Y_i] = -\frac{1}{\theta_i} = \mu_i = \alpha_i\beta_i$ e $V[Y_i] = b''(\theta_i) \cdot a(\phi_i) = \frac{1}{\theta_i^2} \cdot \phi = \alpha_i^2\beta_i^2 \cdot \frac{1}{\alpha_i} = \alpha_i\beta_i^2$.

Repare-se que a expressão genérica para a Variância de uma observação de Y, obtida na equação (4.54), é o produto de duas funções: uma é apenas função do chamado **parâmetro natural**, θ, das distribuições da família exponencial e a outra apenas função do **parâmetro de dispersão**, ϕ. A primeira destas funções, $b''(\theta)$, é muitas vezes designada a **função de variância** da distribuição de Y. As suas expressões para as distribuições estudadas nesta disciplina foram dadas mais acima. Por vezes escreve-se a função de variância como função do valor esperado de Y, ficando nesse caso as expressões acima indicadas com o seguinte aspecto:

Normal $f_V(\mu) = 1$;

Poisson $f_V(\lambda) = \lambda$;

Bernoulli $f_V(p) = p(1-p)$;

Binomial/n $f_V(p) = p(1-p)$ (não esquecer que a função de variância é apenas um dos factores de $V[Y]$);

Gama $f_V(\mu) = \mu^2$.

Uma abordagem alternativa à de Modelos Lineares Generalizados consiste em escolher outras expressões para estas funções de variância, procurando acompanhar eventuais afastamentos das expressões acima indicadas. Esta abordagem, proposta por Wedderburn, está ligada ao conceito de *quasi-verosimilhança* e extravasa o objectivo deste capítulo da disciplina.

4.7.2 Estimação do parâmetro de dispersão nos MLGs

A maioria dos casos particulares de distribuições da família exponencial considerados nesta disciplina tem o parâmetro de dispersão conhecido, com $\phi = 1$. No entanto, para as duas distribuições de variáveis aleatórias contínuas consideradas (Normal e Gama) isso não sucede, colocando-se o problema de como estimar o parâmtero de dispersão ϕ.

Uma primeira possibilidade consiste em estimar ϕ pelo Método da Máxima Verosimilhança. No entanto, é frequente utilizar uma abordagem alternativa para estimar ϕ, associada ao nome de *estatística de Pearson generalizada*. Assim, toma-se

$$\hat{\phi} = \frac{1}{n-(p+1)} \sum_{i=1}^{n} \frac{w_i(y_i - \hat{\mu}_i)^2}{f_v(\hat{\mu}_i)} , \tag{4.55}$$

onde $\hat{\mu}_i$ indica a estimativa do valor esperado de Y_i e $f_v(\hat{\mu})$ indica a função de variância associada a $\hat{\mu}_i$, referidas na página 43.

Considerem-se, por exemplo, modelos com variável resposta Normal, com os pesos usuais $w_i = 1$. Neste caso, o parâmetro ϕ é dado (ver página 13) pr $\phi = \sigma^2$. Tendo em conta que a função variância para esta distribuição é $f_v(\mu) = 1$, tem-se que

$$\hat{\phi} \;=\; \hat{\sigma}^2 \;=\; \frac{1}{n-(p+1)} \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 \;.$$

No caso de se utilizar a função de ligação identidade (ou seja, no caso do Modelo Linear), esta estimativa produz o habitual Quadrado Médio Residual utilizado para estimar σ^2. No caso de outras funções ligação, a estimativa corresponde ao que se utilizaria numa Regressão Não Linear com o valor esperado μ correspondente.

Para modelos com variável resposta Gama, em que $\phi = \frac{1}{\alpha}$ (ver página 15), e para os quais $f_V(\mu) = \mu^2 = (\alpha\beta)^2$, tem-se

$$\hat{\phi} \;=\; \frac{1}{\hat{\alpha}} \;=\; \frac{1}{n-(p+1)} \sum_{i=1}^{n} \frac{w_i (y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2} \;. \tag{4.56}$$

É possível mostrar que, em geral, se trata dum estimador assintoticamente centrado de ϕ e numericamente estável.

4.7.3 Outros critérios de avaliação do desempenho do Modelo

Na Secção 4.6 introduziram-se os conceitos de Desvio e Desvio Reduzido, que são uma ferramenta fundamental para avaliar o desempenho de um MLG. Mas têm sido sugeridas outras ferramentas para avaliar a qualidade do ajustamento de um dado Modelo.

A estatística de Pearson generalizada

O estimador do parâmetro de dispersão considerado na equação (4.55) é, a menos da constante $1/(n-(p+1))$, uma quantidade designada *estatística de Pearson generalizada*:

$$\chi^2 \;=\; \sum_{i=1}^{n} \frac{w_i (y_i - \hat{\mu}_i)^2}{f_v(\hat{\mu}_i)} \;. \tag{4.57}$$

Esta quantidade é por vezes usada como uma medida do grau de ajustamento de um dado Modelo, tendo por base o facto de que, quando aplicada ao Modelo Linear, produzir a usual Soma de Quadrados dos Resíduos ($SQRE$) - veja-se a discussão do estimador de ϕ para o caso desse Modelo. Valores baixos da estatística χ^2 indicam uma proximidade global entre os valores de Y_i e os valores médios estimados, $\hat{\mu}_i$, o que corresponde a uma boa correspondência entre os dados e o modelo ajustado.

O Critério de Informação de Akaike (AIC)

Outra forma frequente de avaliar a qualidade do ajustamento de um Modelo consiste em aplicar um conceito geral, o conceito de *critério de informação de Akaike (AIC)*. O AIC define-se a partir da

log-verosimilhança associada a algum contexto. No contexto de modelação, seja $\ell(\cdot)$ a função log-verosimilhança associada ao ajustamento de um dado modelo com k parâmetros $\underline{\theta}$ a um conjunto de dados $\underline{Y}$, estimando-se $\underline{\theta}$ por máxima verosimilhança. Então o valor do Critério de Informação de Akaike associado ao modelo é dado por -2 vezes a log-verosimilhança máxima (associada aos estimadores de máxima verosimilhança), mais duas vezes o número de parâmetros (k). Aplicado ao caso de MLGs, e dos seus $p+1$ parâmetros β_i, temos:

$$AIC \;=\; -2 \cdot \ell(\underline{\hat{\beta}}; \underline{Y}) + 2 \cdot (p+1) \; . \tag{4.58}$$

4.7.4 Modelos com Componente Aleatória Binomial/n

Foram já estudados vários modelos associados a Componentes Aleatórias com distribuição Bernoulli. No caso de se obterem várias observações dessa variável resposta para a mesma combinação de valores das variáveis preditoras (caso que é muito frequente quando as variáveis preditoras são indicatrizes de níveis de factores) obtem-se colecções de valores 0 e 1 de Y, para o mesmo valor do preditor linear. Nesses casos, uma forma natural de organizar a informação da amostra seria a de contar o número de respostas de tipo "êxito" para cada um desses valores do preditor linear que é comum a várias observações, tabelando assim as observações binárias obtidas em condições análogas. As contagens assim obtidas seriam valores X_j, com $j = 1 : m$, onde m indica o número de diferentes conjuntos de valores das variáveis preditoras. Caso as observações sejam independentes, as variáveis X_j terão Distribuição Binomial, com parâmetros n_j e p_j, indicando o número de observações e probabilidade de "êxito", respectivamente, em cada uma das m situações observadas.

Já foi visto na Subsecção 4.1.1 (página 12) que a distribuição Binomial não pertence à família exponencial de distribuições, mas que uma distribuição de $W = \frac{X}{n}$, sendo $X \cap B(n,p)$, pertence a essa família de distribuições. Assim, é concebível pensar em Modelos Lineares Generalizados em que a Componente Aleatória corresponda a *proporções* de "êxitos" num número fixo de provas de Bernoulli, para m situações diferentes, situações essas correspondentes a diferentes combinações de valores das variáveis preditoras. Assim, a Componente Aleatória de um tal Modelo seria dada por uma amostra $W_1, W_2, ..., W_m$ de variáveis indicando a proporção de êxitos em $n_1, n_2, ..., n_m$ provas, sendo as variáveis $n_j W_j$ Binomiais de parâmetros n_j e p_j.

Uma vez que m observações do tipo agora indicado correspondem a $n = \sum_{j=1}^{m} n_j$ observações de Bernoullis, surge naturalmente a pergunta de qual a relação entre Modelos em que se utilizem estas formas alternativas de organizar a informação. A resposta é que se trata, essencialmente, do mesmo Modelo.

Como ponto de partida para compreender esta última afirmação, repare-se que, como foi visto na Subsecção 4.2.2 (página 16), a log-verosimilhança para n observações independentes de Bernoullis $\{Y_i\}_{i=1}^n$ é da forma:

$$\ell(p) = \sum_{i=1}^{n} \left[\log(1-p_i) + y_i \log\left(\frac{p_i}{1-p_i}\right) \right] .$$

Ora, se há n_j observações efectuadas em condições idênticas ($j = 1 : m; \sum_{j=1}^{m} n_j = n$), é natural considerar que o parâmetro dessas n_j repetições seja equivalente, pelo que a expressão anterior da log-verosimilhança

se poderá escrever como:

$$\ell(p) = \sum_{j=1}^{m} \left[n_j \log(1-p_j) + x_j \log\left(\frac{p_j}{1-p_j}\right)\right] \tag{4.59}$$

onde x_j indica o número de êxitos nas n_j provas de Bernoulli associadas às condições j. Ora, a log-verosimilhança de m observações independentes de Binomiais $X_j \cap B(n_j, p_j)$, $j = 1 : m$, (com os valores de n_j fixos, isto é, não-aleatórios) é dada por:

$$\ell(p) = \sum_{j=1}^{m} \left[\log\binom{n_j}{x_j} + n_j \log(1-p_j) + x_j \log\left(\frac{p_j}{1-p_j}\right)\right] \tag{4.60}$$

(veja-se a Subsecção 4.1.1). Ora, as parcelas $\log\binom{n_j}{x_j}$ desta expressão da log-verosimilhança não dependem dos p_j, pelo que não intervêm na maximização de $\ell(p)$. As restantes parcelas são idênticas às da log-verosimilhança das n observações Bernoulli independentes. Assim, as duas funções log-verosimilhança são, do ponto de vista da estimação de parâmetros, iguais.

Uma vez que o parâmetro θ associado, quer à distribuição Bernoulli, quer à distribuição "Binomial/n", quando expressas na forma de distribuições da família exponencial, é idêntico – $\log\left(\frac{p}{1-p}\right)$ – isso significa que um modelo com Componente Aleatória "Binomial/n" terá função de ligação canónica idêntica à do caso Bernoulli. Associado à analogia das respectivas funções log-verosimilhança, temos assim que os dois tipos de Modelo podem considerar-se como essencialmente iguais, justificando assim a designação genérica de "modelos com componente aleatória Binomial".

Assinale-se que o Desvio do Modelo formulado com Componente Aleatória Binomial/n toma a expressão indicada na equação (4.48), da página 38.

4.8 MLGs no estudo de tabelas de contingência

Os Modelos Lineares Generalizados considerados até aqui admitem qualquer tipo de variáveis preditoras - quantitativas, qualitativas, ou de ambos os tipos. Esse facto ficou já expresso no exemplos considerados nas aulas e nos exercícios. No entanto, a importância dos MLGs - e, em particular, dos Modelos Log-lineares - no estudo de tabelas de contingência, merece uma referência especial. Trata-se de um contexto onde a variável resposta corresponde a contagens (uma variável discreta), que se pretendem relacionar com os níveis de uma ou mais variáveis qualitativas (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, e que podem frequentemente ser enquadradas no âmbito dos MLGs.

4.8.1 Tabelas de contingência com dois factores de classificação

Consideremos o caso frequente de tabelas de contingência com dois factores de classificação, ou seja, a contabilização do número de ocorrências de um dado fenómeno ou acontecimento, para cada possível combinação de níveis de dois factores. Como exemplo, considere-se uma tabela de contagens do número de observações de cada uma de várias espécies (primeiro factor) em cada um de um conjunto de locais (segundo factor). As observações correspondem assim a tabelas com o seguinte aspecto.

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

Quando não há restrições sobre o número total (ou numa das margens) de observações, como será o caso no exemplo das tabelas de tipo locais × espécies, *poderá ser admissível a hipótese de que as contagens são 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 $Y_{ij} = n_{ij}$ ($i = 1 : a$, $j = 1 : b$), tenha distribuição Poisson.

Consideremos o que significa ajustar um modelo tipo ANOVA factorial, prevendo, além de efeitos principais de cada factor, efeitos de interacção entre os dois factores. Como já sabemos da disciplina de Modelação Estatística I, nesse caso haverá uma parcela constante, correspondente a uma combinação de níveis de referência em cada factor, $a-1$ parcelas associadas aos efeitos principais de nível para cada um dos níveis do Factor A, excepto o nível de referência, $b-1$ parcelas análogas para o Factor B, e ainda $(a-1)(b-1)$ parcelas de interacção entre níveis (excluíndo os de referência) de cada factor. Todas essas parcelas totalizam $1 + (a-1) + (b-1) + (a-1)(b-1) = ab$ efeitos previstos num modelo. E a dificuldade reside no facto de esse ser também o número de observações da variável resposta (havendo *uma única contagem n_{ij} para cada célula*). Assim, um modelo deste tipo será um *modelo saturado*[14], cujo ajustamento não deixa margem para variação residual, com todos os inconvenientes associados aos modelos saturados, referidos na Secção 4.6 (página 36).

Mais útil será o ajustamento de modelos associados a hipóteses mais específicas sobre a natureza de fundo da relação entre os factores associados à tabela. Em particular a **hipótese de independência** entre os factores é uma hipótese interessante.

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

$$\begin{aligned} & E[Y_{ij}] &=& \lambda_{ij} &=& n\, p_{ij} \\ \iff \quad & E[Y_{ij}] &=& \lambda_{ij} &=& n\, p_{i\cdot}\, p_{\cdot j} \end{aligned}$$

onde $p_{i\cdot}$ e $p_{\cdot j}$ indicam as probabilidade marginais associadas aos níveis i do Factor A e j do Factor B, respectivamente. Nesse caso, surge de forma natural a ideia de logaritmizar a relação acima referida, gerando então a relação de base:

$$\log\left(E[Y_{ij}]\right) \quad = \quad \log(n) + \log(p_{i\cdot}) + \log(p_{\cdot j}) \tag{4.61}$$

que é uma relação do *tipo da ANOVA a dois factores, sem interacção*:

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

[14] Este é um contexto onde o conceito de *modelo saturado* surge de forma natural.

Estamos assim perante um Modelo com variável resposta Poisson, cujo valor esperado, após uma transformação logarítmica, será dado por uma combinação linear de variáveis indicatrizes de níveis de cada factor. Ou seja, estamos perante um **Modelo Log-linear**, com variáveis preditoras associadas aos níveis de dois factores. Tal como no caso do Modelo Linear aplicado a delineamentos de tipo ANOVA, várias convenções serão possíveis para associar as parcelas α_i e β_j aos níveis do factor. A convenção mais natural (mas com dificuldades de generalização) seria a de associar, para qualquer $i = 1:a$ e $j = 1:b$,

$$\begin{aligned} \alpha_i = \log(p_{i.}) &\iff p_{i.} = \mathrm{e}^{\alpha_i} \\ \beta_j = \log(p_{.j}) &\iff p_{.j} = \mathrm{e}^{\beta_j} \end{aligned}$$

Como sabemos, uma opção desse tipo (sem qualquer restrição adicional) geraria uma matriz do delineamento $\mathbf{X}$ associada com dependências lineares nas suas colunas (veja-se a discussão desta questão nos apontamentos da disciplina de Modelação Estatística I). Por uma questão de coerência com as opções que então foram tomadas, e que são também utilizadas no programa R, poderemos considerar antes que a célula associada à combinação do primeiro nível de cada factor é uma célula de referência[15], sendo a situação nas restantes células analisada comparativamente com essa célula de referência. Nesse caso, teremos

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

Logaritmizando, temos as relações

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

Assim, temos a relação característica de um modelo tipo ANOVA a dois factores, sem interacção,

$$\log(\lambda_{ij}) = \log\left(E[Y_{ij}]\right) = \mu + \alpha_i + \beta_j \quad , \quad i = 2:a\ ,\ j = 2:b\ ,$$

onde (repare que, por definição $\alpha_1 = \beta_1 = 0$):

$$\begin{aligned} \mu &= \log(\lambda_{11}) \iff \mathrm{e}^{\mu} = \lambda_{11} = n \cdot p_{1.} \cdot p_{.1} \\ \alpha_i &= \log\left(\tfrac{p_{i.}}{p_{1.}}\right) \iff \mathrm{e}^{\alpha_i} = \tfrac{p_{i.}}{p_{1.}} \quad (i = 2:a) \\ \beta_j &= \log\left(\tfrac{p_{.j}}{p_{.1}}\right) \iff \mathrm{e}^{\beta_j} = \tfrac{p_{.j}}{p_{.1}} \quad (j = 2:b) \end{aligned}$$

O valor de n, o número total de observações é conhecido. As estimativas de máxima verosimilhança dos parâmetros μ, α_i e β_j serão produzidas pelo ajustamento do Modelo Log-linear, e resultam ser dadas de forma directa a partir das frequências relativas marginais:

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

[15] A célula de referência pode ser qualquer célula. Esta opção serve apenas para fixar ideias e simplificar a notação. No programa R a célula de referência corresponde aos níveis de cada factor cujos nomes apareçam em primeiro lugar, por ordem alfabética.

pelo que

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

O Desvio associado ao modelo é uma medida do grau de afastamento da hipótese de independência. Uma vez que o modelo log-linear saturado – que não prevê qualquer relação especial na tabela de contingências – corresponde a ajustar um modelo de tipo factorial a 2 factores com interacção, como foi visto acima, é possível encarar o teste de independência como um teste aos modelos encaixados. O Desvio do modelo sem interacção corresponde ao valor da estatística de Wilks para uma comparação do submodelo sem interacção (isto é, a hipótese de independência) face ao modelo completo, com interacção (em que não se admite qualquer relação especial). A rejeição da hipótese nula corresponde a rejeitar a hipótese de independência entre os factores.

A abordagem ao estudo das tabelas de contingência de dupla entrada abordada nesta Subsecção é uma abordagem paralela ao estudo de independência através do conhecido teste do Qui-quadrado. A tradicional estatística do teste χ^2 para testes de independência, dada por

$$\chi^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} ,$$

onde $O_{ij} = n_{ij}$ indica o número de observações na célula (i, j) e $\hat{E}_{ij} = n_{..}\hat{p}_{i.}\hat{p}_{.j}$ indica o número esperado estimado de observações nessa mesma célula, resulta ser a *estatística de Pearson generalizada* (equação 4.57, página 44), uma vez que para distribuições de Poisson, a função variância é $f_v(\lambda) = \lambda$. Ao utilizar-se o teste de Wilks para testar a hipótese de independência, serão de esperar resultados análogos (embora não exactamente iguais) aos que se obtêm com o teste do qui-quadrado. O estudo de tabelas de contingência através de modelos log-lineares ganha mais interesse quando se considera o caso de tabelas com três (ou mais) factores de classificação. Esta questão será retomada na Subsecção 4.8.3 (página 52).

Naturalmente que a validade desta abordagem depende da validade dos pressupostos que lhe serviram de base, entre os quais o da distribuição de Poisson para as contagens. Na subsecção 4.8.2 veremos que uma abordagem análoga é felizmente possível, mesmo quando se fixam previamente o número total de observações, n, ou ainda os totais marginais para uma das margens. Este facto dá ao estudo de tabelas de contingência através de Modelos Log-lineares uma grande generalidade.

4.8.2 Poissons e Multinomiais

Os MLGs em que se admite que a componente aleatória tem distribuição Poisson aplicam-se quando as observações da variável resposta são independentes. Assim, as contagens associadas a cada observação Y_i são independentes das restantes. Existe uma situação frequente em que esta hipótese não se verifica, e que no entanto pode ser estudada através de modelos Log-Lineares.

Considere-se que se tem *um número previamente fixado de sujeitos*, e se observa apenas quantos desses sujeitos recaem em categorias definidas por um ou mais factores. Como exemplo, imagine-se que se

dispõem de $n = 40$ indivíduos de uma dada espécie animal, pretendendo classificá-los de acordo com dois critérios de classificação: o sexo e a idade, sendo esta última considerada uma variável categórica com três níveis. O número de indivíduos que recai em cada uma das 6 células (isto é, das seis combinações de sexo e idade) constitui a observação Y_{ij}, a variável resposta que se deseja modelar. Admita-se que, sem a restrição adicional $\sum_{i,j} Y_{ij} = 40 = n$, seria concebível imaginar que o número de observações que recai na célula (i, j) fosse dada por uma Poisson, de parâmetro λ_{ij}.

A restrição resultante de fixar previamente o número total de contagens (que pode, naturalmente, surgir em tabelas de classificação com um único factor, ou em tabelas de contingência com três ou mais factores de classificação) tem como consequência a necessidade de considerar uma outra distribuição para o conjunto das contagens, a distribuição Multinomial, que generaliza a distribuição Binomial para situações onde existem mais do que dois possíveis resultados passíveis de contabilização.

Definição 4.6 *Seja $(Y_1, Y_2, ..., Y_m)$ um vector aleatório cujas componentes tomam valores inteiros não-negativos, sujeitos à restrição $\sum_{i=1}^{m} Y_i = n$, e cuja função de massa probabilística é dada por:*

$$f(y_1, ..., y_m; p_1, ..., p_m) = \frac{n!}{y_1! y_2! \cdots y_m!} p_1^{y_1} p_2^{y_2} \cdots p_m^{y_m}$$

*sendo os m parâmetros $p_1, p_2, ..., p_m$ sujeitos às restrições $p_i \geq 0, \forall i = 1 : m$ e $\sum_{i=1}^{m} p_i = 1$. Então diz-se que o vector $(Y_1, .., Y_m)$ tem **distribuição Multinomial** com parâmetros n e $p_1, p_2, ..., p_m$.*

Numa distribuição Multinomial com $m = 2$, a componente Y_1 é uma distribuição Binomial com parâmetros n e p_1, sendo Y_2 uma Binomial $B(n, p_2)$. Em geral, **a distribuição marginal de qualquer das m componentes Y_i do vector $(Y_1, ..., Y_m)$ é uma Binomial, de parâmetros n e p_i.**

Considere-se agora de novo o problema de m variáveis aleatórias independentes $\{Y_i\}_{i=1}^{m}$, com distribuição de Poisson de parâmetro $\lambda_i, (i = 1 : m)$. Determine-se qual a distribuição das variáveis Y_i, *condicionais à restrição* $\sum_{i=1}^{m} Y_i = n$. Como a soma de Poissons independentes é ainda Poisson, com parâmetro dado pela soma dos respectivos parâmetros (resultado conhecido desde a disciplina de Fundamentos de Probabilidades e Estatística, do Curso Preparatório para este Mestrado), tem-se a seguinte probabilidade conjunta condicional:

$$\begin{aligned}
P\left[Y_1 = y_1, Y_2 = y_2, ..., Y_m = y_m \middle| \sum_{i=1}^{m} Y_i = n\right] &= \frac{P\left[Y_1 = y_1, Y_2 = y_2, ..., Y_m = y_m, \sum_{i=1}^{m} Y_i = n\right]}{P\left[\sum_{i=1}^{m} Y_i = n\right]} \\
&= \frac{P\left[Y_1 = y_1, Y_2 = y_2, ..., Y_m = n - \sum_{i=1}^{m-1} y_i\right]}{P\left[\sum_{i=1}^{m} Y_i = n\right]} \\
&= \frac{\mathrm{e}^{-\lambda_1}\frac{\lambda_1^{y_1}}{y_1!} \cdot \mathrm{e}^{-\lambda_2}\frac{\lambda_2^{y_2}}{y_2!} \cdot \ldots \cdot \mathrm{e}^{-\lambda_{m-1}}\frac{\lambda_{m-1}^{y_{m-1}}}{y_{m-1}!} \cdot \mathrm{e}^{-\lambda_m}\frac{\lambda_m^{y_m}}{y_m!}}{\mathrm{e}^{-\sum_{i=1}^{m}\lambda_i} \cdot \frac{(\sum_{i=1}^{m}\lambda_i)^{\sum_{i=1}^{m} y_i}}{\sum_{i=1}^{m} y_i!}}
\end{aligned}$$

escrevendo $y_m = n - \sum_{i \neq m} y_i$. Os factores envolvendo as exponenciais cancelam todos, e tem-se $n = \sum_{i=1}^{m} y_i$. Re-arrumando a expressão, escrevendo

$$\tau = \sum_{i=1}^{m} \lambda_i$$

e factorizando convenientemente a potência que surge no denominador, tem-se:

$$P\left[Y_1 = y_1, ..., Y_m = y_m \middle| \sum_{i=1}^{m} Y_i = n\right] = \frac{n!}{y_1! y_2! \cdots y_m!} \cdot \left(\frac{\lambda_1}{\tau}\right)^{y_1} \cdot \left(\frac{\lambda_2}{\tau}\right)^{y_2} \cdots \left(\frac{\lambda_m}{\tau}\right)^{y_m} \tag{4.62}$$

que é a função de massa probabilística duma Multinomial com parâmetros n e $p_i = \frac{\lambda_i}{\tau}$ $(i = 1 : m)$.

Assim, m **contagens de Poisson, independentes, sujeitas à restrição adicional de a soma das contagens ser n, têm uma distribuição conjunta Multinomial** com os parâmetros indicados. O facto de se ter usado um único índice i não invalida o raciocínio para tabelas de contingência com mais de um factor de classificação, mantendo-se fixo o número total de observações em todas as células dessa classificação.

À partida, parece tratar-se de um tipo de Modelo totalmente novo, que exige um estudo separado. Mas como veremos de seguida, trata-se de uma situação que pode ser estudada com um Modelo Log-Linear.

De facto, considere-se o modelo Multinomial acabado de introduzir. A verosimilhança duma amostra Multinomial deste tipo é dada pela equação (4.62), e a respectiva log-verosimilhança por:

$$\begin{aligned} \ell(\boldsymbol{\lambda}) &= \log\left(\frac{n!}{y_1! y_2! \cdots y_m!}\right) + \sum_{i=1}^{m} y_i \log\left(\frac{\lambda_i}{\tau}\right) \\ &= \log\left(\frac{n!}{y_1! y_2! \cdots y_m!}\right) + \sum_{i=1}^{m} y_i \log(\lambda_i) - \sum_{i=1}^{m} y_i \log(\tau) \end{aligned}$$

A primeira parcela desta expressão não depende dos parâmetros desconhecidos λ_i e pode ser ignorada para efeitos de estimação de máxima verosimilhança dos λ_is. Temos assim a expressão seguinte para o *núcleo* da log-verosimilhança da Multinomial indicada:

$$\ell^N(\boldsymbol{\lambda}) = \sum_{i=1}^{m} y_i \log(\lambda_i) - n \log(\tau) \tag{4.63}$$

Somando e subtraíndo $\tau = \sum_{i=1}^{m} \lambda_i$, obtém-se a seguinte expressão para este núcleo da log-verosimilhança da Multinomial:

$$\ell^N(\boldsymbol{\lambda}) = \sum_{i=1}^{m} y_i \log(\lambda_i) - \sum_{i=1}^{m} \lambda_i + \tau - n \log(\tau) \tag{4.64}$$

Considerando agora, como no Modelo Log-Linear, que se tem $\log(\lambda_i)$ dado por uma combinação linear de variáveis preditoras $\mathbf{x}_i^t \boldsymbol{\beta}$, obtem-se a seguinte expressão para o núcleo da log-verosimilhança do modelo Multinomial:

$$\ell^N(\boldsymbol{\lambda}) = \underbrace{\sum_{i=1}^{m} y_i \mathbf{x}_i^t \boldsymbol{\beta} - \sum_{i=1}^{m} \mathrm{e}^{\mathbf{x}_i^t \boldsymbol{\beta}}}_{\ell_1(\boldsymbol{\beta})} - \underbrace{(n \log(\tau) - \tau)}_{\ell_2(\tau)} \tag{4.65}$$

Ora a primeira parte desta expressão, $\ell_1(\boldsymbol{\beta})$, é a log-verosimilhança para m Poissons independentes de parâmetros $\lambda_i = \mathrm{e}^{\mathbf{x}_i^t\boldsymbol{\beta}}$ (veja-se a equação (4.23), na página 27). A segunda parte, $\ell_2(\tau)$, é o núcleo da log-verosimilhança da observação n para uma Poisson de parâmetro τ. Assim, a log-verosimilhança de um Modelo Log-Linear sem total de contagens pré-especificado é a soma da log-verosimilhança da Multinomial (modelo com n fixo) e uma parcela que corresponde à log-verosimilhança do valor fixado de n. Como função dos parâmetros $\boldsymbol{\beta}$ do preditor linear, a log-verosimilhança da Multinomial é maximizada pelos mesmos valores de $\hat{\boldsymbol{\beta}}$ que maximizam a log-verosimilhança do Modelo Log-Linear. Assim, o estudo dum Modelo Multinomial, associado a uma tabela de contingências em que o número total de observações está fixado, e a probabilidade de recair na célula i é dada por $p_i = \frac{\lambda_i}{\sum_{i=1}^{m} \lambda_i}$, pode ser feito com base no estudo do Modelo Log-Linear associado a m Poissons independentes de parâmetro λ_i.

Este resultado estende em muito a aplicabilidade das ferramentas de estudo de Modelos Log-Lineares. Raciocínios análogos podem ser aduzidos para outro tipo de determinação prévia de números de contagens, quando numa tabela de contingências de dupla entrada se fixa previamente o número de observações em cada linha ou em cada coluna. Também será possível associar Modelos Log-Lineares a essas situações.

4.8.3 Tabelas de contingência com três factores de classificação

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, existindo uma contagem $Y_{ijk} = n_{ijk}$ 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 com três dimensões.

No que se segue, iremos admitir que as contagens em cada célula duma tavbela som três factores de classificação são observações independentes com distribuição de Poisson, de parâmetros λ_{ijk}, que correspondem às contagens esperadas em cada célula. Como se viu para o caso de tabelas com duas entradas, é possível generalizar a utilização dos modelos log-lineares resultantes, mesmo que as observações não sejam independentes, para o caso de haver um número total, ou totais marginais, fixos.

A tentativa de ajustar o modelo mais geral possível, isto é, um modelo log-linear do tipo ANOVA factorial, a três factores, com todas as possíveis interacções (tripla e os três tipos de interacção dupla) iria produzir um modelo do tipo:

$$log(E[Y_{ijk}]) \;=\; \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} \;. \tag{4.66}$$

Trata-se (como vimos na disciplina de Modelação Estatística I) de um modelo com abc parâmetros. Mas neste contexto, esse é também o número de observações da variável resposta $Y_{ijk} = n_{ijk}$. Assim, o modelo mais geral é um modelo saturado. De novo, os modelos úteis correspondem a modelos com algum tipo de estrutura associada à tabela.

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

Definição 4.7 *Sejam A, B e C três factores com, respectivamente, a, b e c níveis. Seja p_{ijk} a probabilidade de, numa observação aleatória à qual se associam os níveis de cada factor, se verificar o nível i do factor A, j do factor B e k do factor C ($i = 1 : a$, $j = 1 : b$ e $k = 1 : c$), sendo as respectivas probabilidades marginais indicadas de forma análoga, mas substituindo ponto(s) pelo(s) índice(s) ao longo do qual(is)*

se considerou a margem[16]*. Então,*

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

$$p_{ijk} \quad = \quad p_{i..} \cdot p_{.j.} \cdot p_{..k} \qquad \forall\, i,j,k\ ; \tag{4.67}$$

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

$$p_{ijk} \quad = \quad p_{i..} \cdot p_{.jk} \qquad \forall\, i,j,k \tag{4.68}$$

   (com definições análogas para as duas outras formas de isolar um dos três factores);

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

$$p_{ij|k} \quad = \quad p_{i.|k} \cdot p_{.j|k} \qquad \forall\, i,j,k \tag{4.69}$$

   (com definições análogas para as duas outras formas de isolar um dos três factores);

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

$$p_{ij.} \quad = \quad p_{i..} \cdot p_{.j.} \qquad \forall\, i,j \tag{4.70}$$

   (com definições análogas para os outros dois pares dos três factores);

Existem relações de implicação entre vários destes tipos de independência.

Teorema 4.4 *Sejam A, B e C três factores, sendo feitas observações aleatórias que se associam aos níveis de cada factor. Então,*

1. *Se A, B e C são mutuamente independentes, cada factor é conjuntamente independente dos outros dois;*
2. *Se A é conjuntamente independente de (B,C), então (A,C) são condicionalmente independentes de B e (A,B) são condicionalmente independentes de C. (Resultados análogos se verificam começando por qualquer das duas restantes independências conjuntas);*
3. *A independência conjunta de (A,B) com C implica, quer a independência marginal de A e C, quer a independência marginal de B e C. (Resultados análogos verificam-se começando por qualquer das duas restantes independências conjuntas).*

[16] *Assim, por exemplo, $p_{ij.}$ indica a probabilidade de a observação recair no nível i do factor A e j do factor B, qualquer que seja o nível do factor C associado. De forma análoga, $p_{.j.}$ indica a probabilidade da observação recair no nível j do factor B, qualquer que sejam os níveis dos outros dois factores.*

Demonstração.

1. Vamos considerar o caso da independência conjunta de (A,B) com C. Queremos provar que, *para qualquer* i, j, k se tem:

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

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

$$\begin{aligned} & p_{ij.} = \sum_{k=1}^{c} p_{ijk} = \sum_{k=1}^{c} p_{i..} \cdot p_{.j.} \cdot p_{..k} \\ \Longleftrightarrow \quad & p_{ij.} = p_{i..} \cdot p_{.j.} \sum_{k=1}^{c} p_{..k} = p_{i..} \cdot p_{.j.} \ , \end{aligned} \tag{4.71}$$

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

2. Vamos provar que se A é conjuntamente independente de (B,C), então A e B são condicionalmente independentes de C (sendo os outros casos análogos). Temos de mostrar que, para qualquer i, j, k, se tem

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

Ora,

$$\begin{aligned} & p_{ij|k} = \frac{p_{ijk}}{p_{..k}} = \frac{p_{i..} \cdot p_{.jk}}{p_{..k}} \\ \Longleftrightarrow \quad & p_{ij|k} = p_{i..} \cdot p_{.j|k} \ , \end{aligned} \tag{4.72}$$

donde, somando ao longo do índice j se tem

$$\begin{aligned} & \sum_{j=1} p_{ij|k} = p_{i..} \cdot \sum_{j=1}^{b} \frac{p_{.jk}}{p_{..k}} \\ \Longleftrightarrow \quad & p_{i.|k} = p_{i..} \ . \end{aligned}$$

Substituindo esta última expressão para $p_{i..}$ na equação 4.72, obtem-se o resultado desejado:

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

3. Queremos agora provar que a independência conjunta de, digamos, o par (A,B) com C implica a independência marginal de A e C, ou seja, queremos provar que, para quaisquer i, j, k se verifica

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

Ora, o resultado é evidente somando 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.

∇

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 (equação (4.69), página 53), tem-se a seguinte expressão alternativa para a definição de A e B serem independentes condicionalmente a C:

$$p_{ijk} \quad = \quad \frac{p_{i.k} \cdot p_{.jk}}{p_{..k}} \tag{4.73}$$

Vejamos agora 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.

O modelo para a independência mútua dos três factores

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

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

Daqui decorre que, logaritmizando, se tem

$$\log(\lambda_{ijk}) \quad = \quad \log\left(E[Y_{ijk}]\right) \quad = \quad \log(n) + \log(p_{i..}) + \log(p_{.j.}) + \log(p_{..k}) \ , \tag{4.74}$$

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

$$\log\left(E[Y_{ijkj}]\right) \quad = \quad \mu + \alpha_i + \beta_j + \gamma_k \ . \tag{4.75}$$

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

$$\lambda_{111} \quad = \quad E[Y_{111}] \quad = \quad n \cdot p_{1..} \cdot p_{.1.} \cdot p_{..1} \tag{4.76}$$

$$\begin{aligned} \lambda_{ijk} \quad = \quad E[Y_{ijk}] \quad &= \quad n \cdot p_{i..} \cdot p_{.j.} \cdot p_{..k} \\ &= \quad \lambda_{111} \cdot \frac{p_{i..}}{p_{1..}} \cdot \frac{p_{.j.}}{p_{.1.}} \cdot \frac{p_{..k}}{p_{..1}} \quad , \qquad \forall\ i = 2:a\ ,\ j = 2:b\ ,\ k = 2:c \end{aligned} \tag{4.77}$$

Logaritmizando, temos as relações

$$\log(\lambda_{111}) \quad = \quad \log(E[Y_{111}]) \quad = \quad \log(n) + \log(p_{1..}) + \log(p_{.1.}) + \log(p_{..1}) \tag{4.78}$$

$$\log(\lambda_{ijk}) \quad = \quad \log(E[Y_{ijk}]) \quad = \quad \log(\lambda_{111}) + \log\left(\frac{p_{i..}}{p_{1..}}\right) + \log\left(\frac{p_{.j.}}{p_{.1.}}\right) + \log\left(\frac{p_{..k}}{p_{..1}}\right) \tag{4.79}$$

$$, \quad \forall\ i = 2:a\ ,\ j = 2:b\ ,\ k = 2:c$$

Assim, **associado à independência mútua dos três factores temos a relação característica de um modelo tipo ANOVA a três factores, sem qualquer tipo de interacção**,

$$\boxed{\log(\lambda_{ijk}) \quad = \quad \mu + \alpha_i + \beta_j + \gamma_k \qquad , \qquad i = 2:a \ , \ j = 2:b \ , \ k = 2:c \ ,} \tag{4.80}$$

onde

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

Cabe salientar que, neste modelo, os três tipos de efeitos principais, α_i, β_j e γ_k são log-razões de probabilidades. Em particular, cada α_i é o logaritmo da razão entre a probabilidade marginal do nível i do factor A e a probabilidade marginal do nível de referência (o nível $i = 1$) desse mesmo factor[17]. Assim, podemos dizer que a uma "transição" de uma observação do nível de referência do factor A para o nível i desse mesmo factor corresponde (mantendo tudo o resto igual) uma multiplicação por e^{α_i} do valor esperado para a contagem da célula . Interpretações análogas aplicam-se aos outros dois tipos de efeitos de factor.

As **estimativas de máxima verosimilhança** de cada um destes efeitos são as que resultam de substituir cada uma das probabilidades marginais pela frequência relativa correspondente. Assim, por exemplo, e para qualquer i, a probabilidade marginal $p_{i..}$ é estimada por $\hat{p}_{i..} = \frac{n_{i..}}{n_{...}}$, sendo as restantes probabilidades estimadas de forma análoga.

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

Modelos para a independência conjunta de um factor face ao par restante

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. Ora, esse facto significa que o número esperado de observações na célula (i, j, k) será dada por

$$\lambda_{ijk} = E[Y_{ijk}] = n \cdot p_{ijk} = n \cdot p_{i..} \cdot p_{.jk} \ . \tag{4.81}$$

Para modelar esta relação, iremos admitir que o logaritmo deste valor esperado é uma soma tipo ANOVA, com uma parcela constante para todas as observações, parcelas de efeitos principais de cada factor e ainda parcelas de interacção entre os factores B e C:

$$\log\left(\lambda_{ijk}\right) \quad = \quad \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} \ . \tag{4.82}$$

A justificação para este tipo de modelo pode ser vista sob vários ângulos:

[17] Assinale-se que, caso $i = 1$, a expressão para α_i vem igual a zero. O mesmo sucede com os outros factores. Assim, até é possível levantar a restrição sobre os valores de cada índice i, j, k.

- a inexistência de qualquer relação especial na tabela corresponderia a admitir a relação típica dos modelos ANOVA factoriais a três factores, com todas as interacções possíveis. Mas a existência de independência conjunta de (B,C) com A significa, por um lado, que não é necessária a parcela da tripla interacção, uma vez que as parcelas dos efeitos principais de A e as parcelas das interacções B-C cobrem também esse efeito; e por outro lado, tendo em conta que essa independência conjunta implica a independência marginal, quer de A e B, quer de A e C (veja-se o ponto 3 do Teorema 4.4, na página 53), também as parcelas das duplas interacções agora referidas são dispensáveis. Assim sobram apenas os efeitos principais de cada factor e a dupla interacção B-C, como indicado na equação (4.82).

- um modelo para a relação B-C sem qualquer tipo especial de estrutura seria (como foi visto aquando da discussão das tabelas com 2 factores de classificação) um modelo com parcelas de efeitosa principais dos factores B e C, e ainda de interacção B-C. Estas parcelas estariam associadas a $p_{.jk}$ na expressão (4.81) para p_{ijk}. Falta ainda cobrir $p_{i..}$, tornando-se assim necessário acrescentar parcelas de efitos principais do factor A, obtendo-se então assim o modelo indicado em (4.82).

- pode ainda construir-se o modelo a partir da ideia-base que $\lambda_{ijk} = E[Y_{ijk}] = np_{i..} \cdot p_{.jk}$. Considerando que a célula de referência é a célula de cruzamento dos níveis $i = j = k = 1$, tem-se:

$$\begin{array}{rcccccc} \lambda_{111} & = & E[Y_{111}] & = & np_{111} & = & np_{1..} \cdot p_{.11} \\ \iff & & \log(\lambda_{111}) & = & \log(np_{1..} \cdot p_{.11}) & = & \mu \end{array}$$

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

$$\begin{array}{rcccccc} \lambda_{i11} & = & E[Y_{i11}] & = & np_{i11} & = & np_{i..} \cdot p_{.11} \\ \iff & & \lambda_{i11} & = & n \cdot p_{i11} & = & np_{i..} \cdot p_{.11} \quad = \quad (np_{1..} \cdot p_{.11}) \cdot \dfrac{p_{i..}}{p_{1..}} \\ \iff & & \log(\lambda_{i11}) & = & \log(n \cdot p_{1..} \cdot p_{.11}) + \log\left(\dfrac{p_{i..}}{p_{1..}}\right) & = & \mu + \underbrace{\log\left(\dfrac{p_{i..}}{p_{1..}}\right)}_{=\,\alpha_i} \end{array}$$

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

$$\begin{array}{rcccccc} \lambda_{1j1} & = & E[Y_{1j1}] & = & n \cdot p_{1j1} & = & np_{1..} \cdot p_{.j1} \\ \iff & & \lambda_{1j1} & = & n \cdot p_{1j1} & = & np_{1..} \cdot p_{.j1} \quad = \quad (n \cdot p_{1..} \cdot p_{.j1}) \cdot \dfrac{p_{.j1}}{p_{.11}} \\ \iff & & \log(\lambda_{1j1}) & = & \log(np_{1..} \cdot p_{.11}) + \log\left(\dfrac{p_{.j1}}{p_{.11}}\right) & = & \mu + \underbrace{\log\left(\dfrac{p_{.j1}}{p_{.11}}\right)}_{=\,\beta_j} \end{array}$$

  Consideremos ainda as células em que a μ apenas se acrescenta um dos efeitos principais do factor C, ou seja, uma parcela do tipo γ_k $(k = 2 : c)$, que deverá corresponder a uma célula em que

$i = j = 1$, mas $k > 1$. Teremos então, para $k = 2 : c$,

$$\begin{aligned}
\lambda_{11k} &= E[Y_{11k}] = np_{11k} = np_{1..} \cdot p_{.1k} \\
\iff \lambda_{11k} &= np_{11k} = np_{1..} \cdot p_{.1k} = (np_{1..} \cdot p_{.11}) \cdot \frac{p_{.1k}}{p_{.11}} \\
\iff \log(\lambda_{11k}) &= \log(np_{1..} \cdot p_{.11}) + \log\left(\frac{p_{.1k}}{p_{.11}}\right) = \mu + \underbrace{\log\left(\frac{p_{.1k}}{p_{.11}}\right)}_{=\gamma_k}
\end{aligned}$$

Falta apenas obter um tipo de parcelas do modelo: as parcelas de interacção B-C, $(\beta\gamma)_{jk}$. Para as obter, consideremos uma célula em que $i = 1$, mas $j, k \neq 1$: Nesse caso, temos, para $j = 2 : b$ e $k = 2 : c$,

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

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. Resumindo, **o modelo associado à independência conjunta de (B,C) com A** é,

$$\log(\lambda_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} \ , \qquad i = 2 : a, \ j = 2 : b, \ k = 2 : c \ . \tag{4.83}$$

sendo

$$\mu = \log(n \cdot p_{1..} \cdot p_{.11}) \tag{4.84}$$

$$\alpha_i = \log\left(\frac{p_{i..}}{p_{1..}}\right) \qquad \forall i = 2 : a \tag{4.85}$$

$$\beta_j = \log\left(\frac{p_{.j1}}{p_{.11}}\right) \qquad \forall j = 2 : b \tag{4.86}$$

$$\gamma_k = \log\left(\frac{p_{.1k}}{p_{.11}}\right) \qquad \forall k = 2 : c \tag{4.87}$$

$$(\beta\gamma)_{jk} = \log\left(\frac{p_{.jk} \cdot p_{.11}}{p_{.j1} \cdot p_{.1k}}\right) \qquad \forall j = 2 : b \ , k = 2 : c \tag{4.88}$$

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.

Registe-se que, para este tipo de modelos, os efeitos principais de cada factor mantêm a sua natureza de log-razões de probabilidades. Mas, enquanto que os efeitos principais do factor que aparece isolado (no exemplo acima, o factor A) permanecem log-razões da probabilidade marginal do nível i sobre a

probabilidade marginal do nível de referência, $i = 1$, já nos caso dos efeitos associados aos dois factores que são conjuntamente independentes (B e C no exemplo anterior), trata-se da log-razão de probabilidades conjuntas. Assim, por exemplo, β_j é a log-razão da probabilidade conjunta de se estar no nível j do factor B *e no nível de referência do factor C*, sobre a probabilidade conjunta de se estar no nível de referência do factor B *e no nível de referência do factor C*[18]. A interpretação do efeito de interacção, $(\beta\gamma)_{jk}$ é mais complexa. Mas trata-se ainda duma log-razão de probabilidades.

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, $\hat{p}_{.jk} = \frac{n_{.jk}}{n_{...}}$.

Como se viu anteriormente, a independência mútua implica a independência conjunta de cada factor com o par restante. Isto é, a independência mútua de A,B e C implica a independência conjunta de, digamos, A com (B,C), 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. Por outras palavras, podemos comparar o desvio do modelo (4.83) com o desvio do submodelo encaixado (4.80). Se os modelos diferirem significativamente, a hipótese de independência mútua deverá ser rejeitada a favor da independência conjunta indicada.

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 (4.82) da independência de (B,C) com A pode ser referenciado de forma compacta como modelo (B:C), sendo essa efeito de interacção dupla que é previsto no modelo.

Modelos para a independência de um par de factores condicional ao factor restante

Consideremos agora a independência de um par de factores, condicional ao terceiro factor. Por exemplo, consideremos a independência de (A,B), condicional a C. Como foi salientado na terceira nota após o Teorema 4.4 (equação (4.73), página 55), esta independência condicional pode escrever-se apenas em termos das probabilidades conjuntas e marginais:

$$p_{ijk} \quad = \quad \frac{p_{i.k} \cdot p_{.jk}}{p_{..k}} \tag{4.89}$$

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. 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

$$\lambda_{ijk} = E[Y_{ijk}] = n \cdot p_{ijk} = n \cdot \frac{p_{i.k} \cdot p_{.jk}}{p_{..k}} \ . \tag{4.90}$$

Para modelar esta relação, iremos admitir que o logaritmo deste valor esperado é uma soma tipo ANOVA, com uma parcela constante para todas as observações, parcelas de efeitos principais de cada factor e ainda

[18] Pode ainda pensar-se, dividindo estas duas probabilidades por $p_{..1}$, que estamos a falar da log-razão entre a probabilidade de estar no nível j de B, a dividir pela probabilidade de estar no nível de referência do factor B, sendo ambas estas probabilidades *condicionais a estar-se no nível de referência do factor C.*

parcelas de interacção entre os factores A-C e B-C, obtendo-se o seguinte **modelo da independência de (A,B) condicional a C**:

$$\log(\lambda_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} \ , \qquad i = 2:a,\ j = 2:b,\ k = 2:c\ . \tag{4.91}$$

sendo

$$\mu = \log\left(n \cdot \frac{p_{11.} \cdot p_{1.1}}{p_{1..}}\right) \tag{4.92}$$

$$\alpha_i = \log\left(\frac{p_{i.1}}{p_{1.1}}\right) \qquad \forall i = 2:a \tag{4.93}$$

$$\beta_j = \log\left(\frac{p_{.j1}}{p_{.11}}\right) \qquad \forall j = 2:b \tag{4.94}$$

$$\gamma_k = \log\left(\frac{p_{1.k} \cdot p_{.1k} \cdot p_{..1}}{p_{1.1} \cdot p_{.11} \cdot p_{..k}}\right) \qquad \forall i = 2:a \tag{4.95}$$

$$(\alpha\gamma)_{ik} = \log\left(\frac{p_{i.k} \cdot p_{1.1}}{p_{1.k} \cdot p_{i.1}}\right) \qquad \forall i = 2:a\ , k = 2:c \tag{4.96}$$

$$(\beta\gamma)_{jk} = \log\left(\frac{p_{.jk} \cdot p_{.11}}{p_{.1k} \cdot p_{.j1}}\right) \qquad \forall j = 2:b\ , k = 2:c \tag{4.97}$$

A justificação para esta opção de modelo está, como já se indicou, na possibilidade de dispensar a tripla interacção, desde que se mantenham as duas interacções duplas indicadas. E 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 se obtêm substituíndo cada probabilidade p pela correspondente probabilidade estimada $\hat{p}$, dada pela frequência relativa correspondente na tabela.

Tal como no caso anterior, verifica-se que o *modelo agora discutido contém como submodelos, além do modelo de independência mútua, também os modelos de independência conjunta de (B,C) com A (que surge se* $(\alpha\gamma)_{ik} = 0$*, para todo o* i *e* k*) e de (A,C) com B (que surge se* $(\beta\gamma)_{jk} = 0$*, para todo o* j *e* k*).* Pode-se, portanto, testar se a independência condicional é melhor como opção de modelo em relação às duas independências conjuntas que surgem como casos particulares deste modelo anulando, ou uma ou outra, das duplas interacções presentes. Mais uma vez, o facto de os modelos surgirem como modelos encaixados está associado às implicações entre os tipos de independência considerados no Teorema 4.4 (página 53).

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:B,A:C).

A tabela 4.1 indica as designações mnemónicas para os vários tipos de modelos considerados até aqui.

Notação	Tipo de Modelo	Equação do Modelo para $\log(\lambda_{ijk})$	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	(eq. 4.66, pg. 52)	–

Tabela 4.1: Notações mnemónicas para os modelos log-lineares associados aos vários tipos de independência numa tabela de contingências a três entradas.

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. Por um lado registou-se a sentença do réu (condenação à morte, ou não). Depois registou-se a raça do réu (branco ou negro). E finalmente a raça da vítima (branco ou negro). As contagens em cada uma das 8 células são indicadas na Tabela 4.2.

		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 4.2: Dados de 326 julgamentos por homicídio nos EUA de Radelet, M. *Racial characteristics and the inposition 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
```

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
```

Uma vez que a tabela de contingências que está sendo analisada é do tipo mais simples, $2 \times 2 \times 2$, cada linha dos resultados está associada a um *tipo* de efeitos. 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
```

Repare-se com a exclusão de uma das parcelas de dupla interacção teve um efeito muito pouco relevante na qualidade do ajustamento: com 2 graus de liberdade, o desvio reduzido residual continua a ser quase nulo. Embora não de forma tão clara, a interacção "sentença:raça da vítima" também parece ser pouco importante (e não significativa a um nível como 1%). A sua exclusão dá origem ao seguinte ajustamento.

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

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                     2.8139     0.1770  15.897  < 2e-16 ***
sentencaOutra                   2.0864     0.1767  11.807  < 2e-16 ***
raca.reunegro                  -0.8741     0.1500  -5.828  5.6e-09 ***
raca.vitimanegro               -2.8201     0.3431  -8.219  < 2e-16 ***
raca.reunegro:raca.vitimanegro  3.3116     0.3786   8.748  < 2e-16 ***
---
    Null deviance: 395.9153  on 7  degrees of freedom
Residual deviance:   8.1316  on 3  degrees of freedom
AIC: 53.813
```

É agora evidente que não é possível simplificar ulteriormente o modelo, sem perder muito significativamente na qualidade de ajustamento. O modelo final ajustado corresponde a um modelo de independência conjunta dos factores (*Raça.réu*,*Raça.vítima*), face ao factor *Sentença*. Trata-se de uma relação que, como se viu no Teorema 4.4 (pg. 53), implica que a raça do réu é independente da sentença, condicionada à raça da vítima, tal como a raça da vítima é independente da sentença, condicionada à raça do réu. Mas repare-se como existe uma fortíssima interacção entre os dois factores que surgem associados: *Raça.réu* e *Raça.vítima*. Adiante retomaremos esta discussão.

Os valores estimados dos parâmetros do modelo têm a interpretação indicada aquando do estudo geral de modelos de independência conjunta, simplificada pelo facto de haver apenas dois níveis em todos os factores. O valor estimado da constante μ, que é $\hat{\mu} = 2.8139$ significa que o valor esperado na célula de referência (a célula de condenação, para réus brancos e vítmias brancas) é $e^2.8139 = 16.67482$, próximo do valor observado (19). No caso do primeiro factor (*Sentença*), o valor $\hat{\alpha}_2 = 2.0864$ significa que, em relação a esse valor esperado para a célula de referência, o valor esperado na célula resultante de transitar para "Outra sentença" (mantendo réu e vítima brancos) é $e^{2.0864} = 8.055862$ vezes maior, ou seja, é $e^{2.0864} * 16.67482 = 134.3300$, muito próximo do valor observado (132). Outros parâmetros de efeitos principais têm interpretações análogas.

Tabelas parciais e o paradoxo de Simpson

Acabámos de ver como as sentenças atribuídas eram independentes da raça do réu, dada a raça da vítima, e da raça da vítima, dada a raça do réu. E no entanto, havia uma forte interacção entre raça da vítima

e raça do réu. Olhando para a tabela compreende-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. Neste tipo de situações, surge uma situação conhecida por *paradoxo de Simpson*.

Comecemos por introduzir um conceito auxiliar. Designa-se por **tabela parcial** a 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 do réu" é a seguinte:

		Sentença	
Raça Réu	Raça Vítima	Pena de Morte	Outra Pena
Branco	Branco	19	132
	Negro	0	9

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 ao mesmo par de factores indicado na tabela parcial, 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

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.

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%, enquantro 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. 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**.

Vimos que o paradoxo de Simpson ocorreu nas asociações entre factores A-B, num contexto em que havia independência conjunta de (B,C) com A. Para que não surja este paradoxo, teria de haver independência de (A,C) *condicional a B*, ou de (B,C) *condicional* a A, isto é, teria de ser aplicável o modelo (A:B,B:C) ou o modelo (A:B,A:C). Numa situação desse tipo, é possível estudar a associação entre A e B a partir da tabela marginal, seguros de que as correspondentes associações nas tabelas parciais são de tipo análogo.

4.9 Resíduos e Validação do Modelo

O conceito de resíduos, $e_i = y_i - \hat{y}_i$, usado no Modelo Linear como ferramenta para a validação das hipóteses subjacentes ao Modelo, tem de ser adaptado nos MLGs, onde, diversamente do que acontecia nos Modelos Lineares, não se contempla a existência de erros aleatórios aditivos, como hipóteses distribucionais associadas.

Em Modelos Lineares Generalizados utilizam-se diversos conceitos de resíduos, sendo os principais os **resíduos de Pearson** e os **resíduos do desvio**.

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

Definição 4.8 *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 à* ***raíz quadrada da contribuição de cada observação para a estatística de Pearson generalizada*** *(equação 4.57, página 44):*

$$r_i{}^P = \frac{(Y_i - \hat{\mu}_i) \cdot \sqrt{w_i}}{\sqrt{f_v(\hat{\mu}_i)}} \tag{4.98}$$

No denominador desta definição encontramos a raíz quadrada da função de variância associada à observação Y_i correspondente. A expressão para esta função de variância é, naturalmente, diferente para cada distribuição de Y (veja-se a página 43). Assim,

No caso de Y ter distribuição Normal: Tem-se $f_v(\hat{\mu}_i) = 1$, pelo que o resíduo de Pearson será o habitual resíduo do Modelo Linear:

$$r_i^P = Y_i - \hat{\mu}_i \tag{4.99}$$

No caso de Y ter distribuição Bernoulli: Tem-se $f_v(\hat{p}_i) = \hat{p}_i(1 - \hat{p}_i)$, e o resíduo de Pearson é:

$$r_i^P = \frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}} \tag{4.100}$$

No caso de Y ter distribuição Binomial/n: Tem-se de novo $f_v(\hat{p}_i) = \hat{p}_i(1 - \hat{p}_i)$, mas agora existem ponderações $w_i = n_i$, pelo que o resíduo de Pearson é dado por:

$$r_i^P = \frac{Y_i - \hat{p}_i}{\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n_i}}} \tag{4.101}$$

No caso de Y ter distribuição Poisson: Tem-se $f_v(\hat{\lambda}_i) = \hat{\lambda}_i$, e o resíduo de Pearson é:

$$r_i^P = \frac{Y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}} \tag{4.102}$$

No caso de Y ter distribuição Gama: Tem-se $f_v(\hat{\mu}_i) = \hat{\mu}_i^2 = (\hat{\alpha}_i\hat{\beta}_i)^2$, e o resíduo de Pearson é:

$$r_i^P = \frac{Y_i - \hat{\alpha}_i\hat{\beta}_i}{\hat{\alpha}_i\hat{\beta}_i} \tag{4.103}$$

Assinale-se que, **para cada uma destas distribuições, a expressão para os resíduos de Pearson traduz-se numa expressão em termos dos preditores lineares ajustados, com base na função de ligação específica que tenha sido utilizada no Modelo**. Assim, por exemplo,

uma Regressão Logística ajustada a dados dicotómicos gera resíduos de Pearson da forma:

$$r_i^P = \frac{Y_i - \frac{1}{1+\mathrm{e}^{-\mathbf{x}_i^t\hat{\beta}}}}{\sqrt{\frac{1}{1+\mathrm{e}^{-\mathbf{x}_i^t\hat{\beta}}} \cdot \frac{\mathrm{e}^{-\mathbf{x}_i^t\hat{\beta}}}{1+\mathrm{e}^{-\mathbf{x}_i^t\hat{\beta}}}}} = \frac{Y_i \cdot (1+\mathrm{e}^{-\mathbf{x}_i^t\hat{\beta}}) - 1}{\sqrt{\mathrm{e}^{-\mathbf{x}_i^t\hat{\beta}}}} \tag{4.104}$$

uma Regressão Probit ajustada a dados dicotómicos gera resíduos de Pearson da forma:

$$r_i^P = \frac{Y_i - \Phi(\mathbf{x}_i^t\beta)}{\sqrt{\Phi(\mathbf{x}_i^t\beta) \cdot (1 - \Phi(\mathbf{x}_i^t\beta))}} \tag{4.105}$$

um Modelo Log-log do complementar ajustado a dados dicotómicos gera resíduos de Pearson da forma:

$$r_i^P = \frac{Y_i - \left(1 - \mathrm{e}^{-\mathrm{e}^{\mathbf{x}_i^t\beta}}\right)}{\sqrt{\left(1 - \mathrm{e}^{-\mathrm{e}^{\mathbf{x}_i^t\hat{\beta}}}\right) \cdot \left(\mathrm{e}^{-\mathrm{e}^{\mathbf{x}_i^t\hat{\beta}}}\right)}} \tag{4.106}$$

De forma análoga, as fórmulas (4.102) ou (4.103) para os resíduos de Pearson poderão ser concretizadas, para cada modelo específico (isto é, para cada escolha específica de função de ligação a associar à variável-resposta com a distribuição indicada), produzindo resíduos directamente calculáveis a partir dos valores observados (para as variáveis preditoras e para a variável resposta) e dos coeficientes da combinação linear ajustados, $\hat{\beta}_j$.

4.9.2 Resíduos do Desvio

Um conceito alternativo de resíduos aproveita a analogia no papel desempenhado pelo Desvio no estudo dum MLG, e da Soma de Quadrados dos Resíduos no Modelo Linear. Uma vez que o Desvio é uma soma

de n parcelas, uma para cada observação, podemos escrever o Desvio sob a forma:

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

Toma-se como definição deste novo tipo de resíduo a raíz quadrada de cada parcela d_i, afectada de um sinal correspondente ao facto de a correspondente observação y_i ser inferior, ou superior, ao seu valor esperado estimado.

Definição 4.9 *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 (dado na Definição 4.5). Designa-se* ***resíduos do Desvio*** *de cada observação a:*

$$r_i{}^D \quad = \quad sinal(y_i - \hat{\mu}_i) \cdot \sqrt{d_i} \tag{4.107}$$

Vejamos agora qual a expressão dos resíduos do Desvio para as distribuições da Componente Aleatória estudadas nesta disciplina.

No caso de Y ter distribuição Normal , sabemos (equação 4.46, página 38) que $d_i = (y_i - \hat{\mu}_i)^2$. Assim, a expressão dos resíduos do Desvio para Y Normal é:

$$r_i^D = y_i - \hat{\mu}_i \tag{4.108}$$

pelo que, **no caso do Modelo Linear, os resíduos do Desvio não são mais que os habituais resíduos subjacentes à Soma dos Quadrados Residual**;

No caso de Y ter distribuição Bernoulli ,

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

(equação 4.47, página 38). Assim, a expressão dos resíduos do Desvio para Y Bernoulli é:

$$r_i^D \quad = \quad sinal(y_i - \hat{p}_i) \cdot \sqrt{d_i} \quad = \quad \begin{cases} -\sqrt{-2 \log(1 - \hat{p}_i)} & \text{se} \quad y_i = 0 \\ \sqrt{-2 \log(\hat{p}_i)} & \text{se} \quad y_i = 1 \end{cases} \tag{4.109}$$

No caso de Y ter distribuição Binomial/n ,

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

(equação 4.48, página 38). Assim, a expressão dos resíduos do Desvio para Y Binomial/n é:

$$r_i^D = \begin{cases} \sqrt{-2n_i \left[y_i \log(y_i) + (1 - y_i) \log(1 - y_i) - y_i \log(\hat{p}_i) - (1 - y_i) \log(1 - \hat{p}_i) \right]} & \text{se} \quad y_i \neq 0, 1 \\ \sqrt{2n_i \left[y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i) \right]} & \text{se} \quad y_i \in \{0, 1\} \; . \end{cases} \tag{4.110}$$

No caso de Y ter distribuição Poisson: Neste caso,

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

(equação 4.49, página 39). Assim, a expressão dos resíduos do Desvio para Y Poisson é:

$$r_i^D \quad = \quad sinal(y_i - \hat{\lambda}_i) \cdot \sqrt{d_i} \quad = \quad sinal(y_i - \hat{\lambda}_i) \cdot \sqrt{2 \left[y_i \log\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i) \right]} \qquad (4.111)$$

No caso de Y ter distribuição Gama: Neste caso,

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

(equação 4.50, página 39). Assim, a expressão dos resíduos do Desvio para Y Gama é:

$$r_i^D \quad = \quad sinal(y_i - \hat{\mu}_i) \cdot \sqrt{d_i} \quad = \quad sinal(y_i - \hat{\mu}_i) \cdot \sqrt{2 \cdot \left[\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} - \log\left(\frac{y_i}{\hat{\mu}_i}\right) \right]} \qquad (4.112)$$

4.9.3 Os Resíduos na Validação de um MLG

Os resíduos definidos nas Subsecções anteriores podem ser utilizados para validar as hipóteses subjacentes a um Modelo Linear Generalizado, seguindo ideias gerais já consideradas aquando do estudo do Modelo Linear. Em geral, prefere-se a utilização de resíduos do Desvio aos resíduos de Pearson. É frequentemente sugerida uma transformação dos resíduos, designados *resíduos padronizados*, cuja discussão, no entanto, se omite por falta de tempo e dadas as dificuldades adicionais que o seu cálculo envolve. Nesta, como noutras questões associadas à validação de Modelos Lineares Generalizados sugere-se a consulta de McCullagh & Nelder (1989) ou Turkman & Silva (2000) para uma discussão mais aprofundada.

Como ideias de fundo, refira-se que subjacente à validade de um MLG está, quer a validade da hipótese distribucional associada à sua componente aleatória, quer a adequabilidade da componente sistemática como preditor linear, quer ainda a adequabilidade da função de ligação escolhida para relacionar essas duas componentes do MLG.

Os afastamentos a estas hipóteses podem ser *sistemáticos*, ou seja, resultado de um desajustamento global dos dados às hipóteses referidas, ou *isolados*, isto é, respeitantes apenas a uma, ou poucas, observações atípicas.

Sugerem-se as seguintes inspecções gráficas:

resíduos contra esperanças estimadas: este é o gráfico correspondente ao gráfico de resíduos *vs.* valores ajustados no Modelo Linear. No contexto de MLGs, sugere-se que seja um gráfico de resíduos (ou até dos resíduos padronizados, acima referidos) sobre transformações das esperanças estimadas, transformações essas que diferem consoante a distribuição dos Y_i e que visam fazer com que o gráfico tenha 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 parâmetro λ;
- $2\arcsin(\hat{p})$ para Y Bernoulli de parâmetro p.
- $-\frac{2}{\sqrt{\hat{\mu}}}$ para Y Gama de parâmetros α e β com $\mu = \alpha\beta$.

Assim, **um bom ajustamento do Modelo Linear Generalizado deve produzir gráficos de resíduos contra estas transformações das Esperanças estimadas, em que os resíduos se dispersam em torno do valor 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.

resíduos contra uma variável preditora: um tipo de gráfico também análogo ao que foi considerado no caso do Modelo Linear, e de leitura semelhante; a sua utilidade é tanto maior quanto menor fôr o número de variáveis preditoras.

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

4.10 Exercícios

1. No livro de W.N. Venables e B.D. Ripley, *Modern Applied Statistics with S-Plus* (1994, Springer-Verlag), refere-se uma experiência que 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, e registou-se o número de individuos de cada lote que morria, ou ficava inactivo, no fim desse período de exposição. Os resultados (isto é, o número de mortes) são sintetizados na seguinte tabela, sendo as doses expressas em μg.

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

   (a) Construa uma nuvem de pontos com, no eixo horizontal, a variável Dose, e no eixo vertical, a proporção de Mortes em cada lote de 20 individuos. Repita, mas agora utilizando cores diferentes para representar os lotes associados a individuos de cada Sexo. Comente.

   (b) Repita a alínea anterior, mas agora associando o eixo horizontal à variável $\log_2$(Dose). Esta transformação pode justificar-se, tendo em conta que as doses utilizadas duplicam em cada nova situação experimental. Comente.

   (c) Ajuste uma Regressão Logística aos dados, ignorando as diferenças de sexo, e utilizando como variável preditora $\log_2$(Dose). Comente os resultados obtidos. Trace, por cima da nuvem de pontos obtida na alínea anterior, a curva estimada para a probabilidade de morte, $p(x)$, onde x indica valores de $\log_2$(Dose).

   (d) Repita a alínea anterior, mas agora cruzando a variável preditora $\log_2$(Dose) com o factor Sexo. Ajuste, por cima da nuvem de pontos, as curvas estimadas para a probabilidade de morte, associadas a cada Sexo: $p_F(x)$ e $p_M(x)$. Comente.

   (e) Repita as duas alíneas anteriores, mas utlizando agora um Modelo *Probit*. Comente os resultados.

   (f) Repita de novo os ajustamentos, mas utilizando um Modelo Linear Generalizado com Componente aleatória adequada e utilizando uma função de ligação Log-log do Complementar. Comente os resultados.

2. A fim de estudar os efeitos cancerígenos de um produto tóxico em ratos, foram administradas três diferentes doses da substância tóxica (0, 0,45 e 0,75 partes por 10 000) a algumas centenas de ratos, durante um de dois períodos de exposição (16 ou 24 meses). No final do período de exposição verificava-se a existência de tumores nos ratos. Os resultados da experiência foram os seguintes:

		Dose		
Exposição		0	0,45	0,75
16 meses	Ratos com tumores	1	3	7
	Ratos sem tumores	204	301	186
24 meses	Ratos com tumores	20	98	118
	Ratos sem tumores	742	790	469

Ajustou-se um Modelo Linear Generalizado adequado para uma componente aleatória dicotómica, **com função de ligação *probit***, e sem prever efeitos de interacção entre as variáveis preditoras **Dose** e tempo de **Exposição**. Obtiveram-se os seguintes resultados:

```
> summary(mice2.glm2)
Call: glm(formula = cbind(Mortes, Sobreviventes) ~ Dose + Exposicao,
    family = binomial(probit))
Deviance Residuals:
[1]   0.78561  -0.45537   0.05861  -0.43065   0.50701  -0.25993
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.84737    0.39457 -12.285  < 2e-16 ***
Dose         1.43436    0.13967  10.270  < 2e-16 ***
Exposicao    0.12287    0.01629   7.542 4.64e-14 ***

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 198.5347  on 5  degrees of freedom
Residual deviance:   1.3381  on 3  degrees of freedom
AIC: 33.594
Number of Fisher Scoring iterations: 3
Correlation of Coefficients:
          (Intercept) Dose
Dose           -0.265
Exposicao      -0.980 0.09
```

(a) Descreva em maior pormenor o tipo de modelo ajustado, indicando a relação considerada entre o surgimento de tumores e as variáveis preditoras.

(b) Comente a qualidade do ajustamento do modelo aos dados.

(c) Considera possível simplificar ulteriormente o modelo sem prejuízo significativo na qualidade do ajustamento? Justifique formalmente.

(d) Com base no modelo ajustado, responda às seguintes questões:

i. Para uma dose de 0,75 partes por 10 000 da substância tóxica, qual a proporção esperada de ratos com tumores ao fim de 36 meses de exposição?

ii. Qual a dose associada a 50% de ratos com tumor ao fim de 24 meses de exposição?

3. Defina os seguintes conceitos, no contexto de Modelos Lineares Generalizados:

(a) função de ligação

(b) resíduo do desvio

4. Descreva a utilização do algoritmo de Newton-Raphson na obtenção dos estimadores de máxima verosimilhança dos parâmetros dum modelo linear generalizado.

5. O livro de P. McCullagh e J.A. Nelder, *Generalized Linear Models* (2a. edição, 1989, Chapman & Hall) é, indiscutivelmente, o livro de referência para Modelos Lineares Generalizados. Nesse livro, nas páginas 300-302, é discutido um conjunto de dados onde se mediram tempos de coagulação

(em segundos) de sangue, para plasma normal diluído com nove diferentes concentrações de plasma sem a proteína protrombina (do tipo protease serina, produzida no fígado e que, quando activada - gerando a trombina - está associada à coagulação do sangue). Dois diferentes lotes de um agente activador da coagulação foram utilizados. Os dados observados foram os seguintes.

	Tempo de coagulação	
Concentração	Lote 1	Lote 2
5	118	69
10	58	35
15	42	26
20	35	21
30	27	18
40	25	16
60	21	13
80	19	12
100	18	12

Deseja-se estudar o efeito das diferentes concentrações de plasma sem protrombina sobre os tempos de coagulação, eventualmente admitindo diferenças associadas aos agente de coagulação utilizados.

(a) Represente graficamente *tempo* de coagulação (eixo vertical) contra *concentrações* de plasma (eixo horizontal), utilizando símbolos e/ou cores diferentes para representar as observações de cada lote. Comente.

(b) É sugerido que a relação entre as variáveis *tempo* e concentração de plasma sem protrombina (variável *conc*) é de tipo hiperbólico, ou seja da forma $tempo = \frac{1}{\beta_0+\beta_1 \cdot conc}$. Produza uma representação gráfica adequada para validar visualmente esta proposta. Comente.

(c) Após um estudo gráfico adequado, conclui-se que a relação mais adequada parece ser do tipo hiperbólico mas sobre os logaritmos das concentrações de plasma, ou seja, da forma $tempo = \frac{1}{\beta_0+\beta_1 \log(conc)}$. Confirme, produzindo a representação gráfica adequada.

(d) Para ajustar a relação indicada na alínea anterior, a função de ligação indicada é a função recíproco, $g(\mu) = \frac{1}{\mu}$, utilizando como preditor a variável das log-concentrações. Mas permanece de pé a escolha de qual a distribuição a associar à variável-resposta *tempo*. Ajuste dois diferentes MLGs, admitindo:

   i. que *tempo* tem distribuição Normal (Nota: No *R*, este ajustamento corresponde a dar o argumento `family=gaussian(link=''inverse'')` no comando `glm`);

   ii. que *tempo* tem distribuição Gama (Nota: No *R*, este ajustamento corresponde a dar o argumento `family=Gamma`, não sendo necessário especificar a função de ligação, uma vez que a função recíproco é a função de ligação canónica para a distribuição Gama).

   Trace as curvas correspondentes a cada ajustamento por cima da nuvem de pontos de *tempo* (eixo vertical) contra log-concentrações de plasma (eixo horizontal). Comente.

(e) Compare os ajustamentos obtidos na alínea anterior. Comente, e diga qual a escolha mais adequada para distribuição de *tempo*, tendo em conta a natureza e valores dessa variável resposta, e o conjunto da informação disponível.

6. No livro *Categorical Data Analysis*, de Alan Agresti (John Wiley & Sons, 1990), refere-se um estudo sobre a associação entre os níveis de tensão arterial e de colesterol e a existência de problemas cardio-vasculares. Foram observados 1329 pessoas, tendo sido registado, para cada pessoa, se existiam problemas cardio-vasculares e ainda os correspondentes valores de tensão arterial e nível de colesterol (em *mg/100 ml*). As observações foram agrupadas em classes de tensão arterial e de níveis de colesterol, como se indica na tabela seguinte, onde, entre parenteses se indica o número de pessoas observadas em cada grupo, e fora dos parenteses se indica o número dessas pessoas com problemas cardio-vasculares.

Tensão	Colesterol						
Arterial	< 200	$200-209$	$210-219$	$220-244$	$245-259$	$260-284$	> 284
< 117	2 (53)	0 (21)	0 (15)	0 (20)	0 (14)	1 (22)	0 (11)
$117-126$	0 (66)	2 (27)	1 (25)	8 (69)	0 (24)	5 (22)	1 (19)
$127-136$	2 (59)	0 (34)	2 (21)	2 (83)	0 (33)	2 (26)	4 (28)
$137-146$	1 (65)	0 (19)	0 (26)	6 (81)	3 (23)	2 (34)	4 (23)
$147-156$	2 (37)	0 (16)	0 (6)	3 (29)	2 (19)	4 (16)	1 (16)
$157-166$	1 (13)	0 (10)	0 (11)	1 (15)	0 (11)	2 (13)	4 (12)
$167-186$	3 (21)	0 (5)	0 (11)	2 (27)	2 (5)	6 (16)	3 (14)
> 186	1 (5)	0 (1)	3 (6)	1 (10)	1 (7)	1 (7)	1 (7)

Considerando que o número total de observações em cada célula desta tabela foi previamente fixado, responda às seguintes alíneas.

(a) Ajuste um Modelo de Regressão Logística que procure relacionar a existência de problemas cardio-vasculares apenas com o preditor quantitativo *tensão arterial*. Comente os resultados. **Nota:** Considere que todas as observações de uma dada classe de valores de tensão são dados pelos seguintes representantes das classes: 111.5, 121.5, 131.5, 141.5, 151.5, 161.5, 176.5 e 191.5.

(b) Ajuste um Modelo de Regressão Logística que procure relacionar a existência de problemas cardio-vasculares com dois preditores quantitativos: *tensão arterial* e *nível de colesterol*. Comente os resultados. **Nota:** Além de considerar a Nota à alínea anterior, considere também que os níveis de colesterol são dados pelos seguintes representantes de cada classe: 190, 205, 215, 232, 252, 272 e 290.

(c) Considere de novo o caso de haver um único preditor para a existência de doenças cardio-vasculares, o preditor *tensão arterial*, mas pense agora neste preditor como um *factor*. Está-se assim numa situação análoga às consideradas nos Modelos ANOVA, aquando do estudo do Modelo Linear. O ajustamento de um Modelo de Regressão Logística neste caso produz, para observações do primeiro nível do Factor (o nível de tensão arterial inferior a 117, usado como nível de referência no ajustamento), uma relação do tipo

$$\log\left(\frac{p_1}{1-p_1}\right) \;=\; \alpha_1$$

onde p_1 indica a probabilidade de haver doença cardio-vascular, dado que se está no primeiro nível do factor, e α_1 indica a constante aditiva do preditor linear. Para observações nos

restantes níveis do Factor, tem-se uma relação do tipo

$$\log\left(\frac{p_i}{1-p_i}\right) \;=\; \alpha_1 + \alpha_i, \qquad i = 2:k,$$

sendo p_i a probabilidade de haver doença cardio-vascular, dado que se está no i-ésimo nível do factor, e α_i o acréscimo a α_1 resultante do facto de se ter uma observação do nível i do Factor. Interprete o significado da hipótese

$$\alpha_2 \;=\; \alpha_3 \;=\; ... \;=\; \alpha_k \;=\; 0,$$

isto é, da hipótese de o Modelo sob estudo ser equivalente ao Modelo Nulo.
Sugestão: Defina Y como a variável dicotómica que toma valor 1 caso uma pessoa tenha problemas cardio-vasculares, e valor 0 caso contrário. Considere a variável aleatória X que indica o nível (a classe) da tensão arterial duma pessoa. Considere as probabilidades condicionais $P[Y = 1 \mid X = x]$. Qual a relação entre as variáveis aleatórias Y e X no caso de ser verdadeira a hipótese $\alpha_i = 0,\ \forall\ i > 1$?

(d) Ajuste o modelo considerado na alínea anterior e discuta os resultados.

7. Entre os conjuntos de dados disponibilizados no programa `R` encontra-se o conjunto designado `HairEyeColor`. Trata-se duma tabela de contingências em que 592 alunos foram classificados de acordo com a sua côr de cabelo (4 níveis), côr dos olhos (4 níveis) e sexo. Estude essa tabela através de modelos log-lineares, para ver se é admissível algum tipo de independência entre os factores.

Apêndice A

Funções de $\mathbb{R}^n$ – revisão

Vejamos algumas noções básicas sobre funções em $\mathbb{R}^n$, e nomeadamente o cálculo de extremos locais de tais funções. Para uma discussão mais pormenorizada deste tema, consulte-se, por exemplo, *Calculus in Vector Spaces* (2a. edição), Corbin, L.J. & Szczaba, R.H., Marcel Dekker, 1995.

Como em muitas outras situações, a utilização de notação matricial simplifica notavelmente a discussão.

Seja:

$$\begin{array}{rl} f: & \mathbb{R}^n \longrightarrow \mathbb{R} \\ & \mathbf{x} \longrightarrow f(\mathbf{x}) \end{array}$$

Admita-se que as funções às quais se faz referência são sempre diferenciáveis.

Represente-se o **operador** que a cada função (do tipo acima referido) faz corresponder o vector das suas derivadas parciais por $\frac{\partial}{\partial \mathbf{x}}$, isto é:

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

e seja $\left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\right)_{(\mathbf{x}=\mathbf{x}_0)} \in \mathbb{R}^n$ o valor desse vector de derivadas parciais no ponto $\mathbf{x} = \mathbf{x}_0$, ou **gradiente** de f no ponto $\mathbf{x} = \mathbf{x}_0$. Os gradientes de alguns tipos de funções são particularmente simples de calcular. De facto, é fácil verificar que:

1. $f(\mathbf{x}) = c \quad , \quad \forall \mathbf{x} \in \mathbb{R}^n \quad \Longrightarrow \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{0} \quad , \forall \mathbf{x} \in \mathbb{R}^n$
2. $f(\mathbf{x}) = \mathbf{a}^t\mathbf{x} = \sum_{i=1}^n a_i x_i \quad \Longrightarrow \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{a}$, para qualquer vector de coeficientes $\mathbf{a} \in \mathbb{R}^n$.
3. **(Forma quadrática)**. $f(\mathbf{x}) = \mathbf{x}^t\mathbf{A}\mathbf{x} = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j \Longrightarrow \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$, para qualquer matriz simétrica $\mathbf{A}_{nxn}$.
4. **(Forma quadrática generalizada)**. $f(\mathbf{x}) = (\mathbf{a} - \mathbf{C}\mathbf{x})^t\mathbf{A}(\mathbf{a} - \mathbf{C}\mathbf{x}) \Longrightarrow \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = -2\mathbf{C}^t\mathbf{A}(\mathbf{a} - \mathbf{C}\mathbf{x})$, onde $\mathbf{A}_{pxp}$ é qualquer matriz simétrica, $\mathbf{C}_{pxn}$ é uma matriz de coeficientes e $\mathbf{a}_{px1}$ é um vector de coeficientes.

Por sua vez, a matriz das segundas derivadas parciais, isto é, a matriz $\mathbf{H}$ cujo elemento genérico da linha i, coluna j é dado por:

$$(\mathbf{H})_{(i,j)} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}$$

designa-se a **matriz Hessiana** de f. Tem-se:

1. $f(\mathbf{x}) = \mathbf{x}^t\mathbf{A}\mathbf{x} \quad \Longrightarrow \quad \mathbf{H} = 2\mathbf{A}$
2. $f(\mathbf{x}) = (\mathbf{a} - \mathbf{C}\mathbf{x})^t\mathbf{A}(\mathbf{a} - \mathbf{C}\mathbf{x}) \quad \Longrightarrow \quad \mathbf{H} = 2\mathbf{C}^t\mathbf{A}\mathbf{C}$

Para calcular extremos locais de funções em $\mathbb{R}^n$ (que admitem segunda derivada), tem-se:

1. **Condição necessária (pontos de estacionaridade):**

$$\left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\right)_{(\mathbf{x}=\mathbf{x}_0)} = \mathbf{0}$$

2. **Condição suficiente (se $\mathbf{x}_0$ é ponto de estacionaridade)**:

$\mathbf{H}_{(\mathbf{x}=\mathbf{x}_0)}$	definida positiva	$\Longrightarrow \mathbf{x}_0$ é mínimo.
$\mathbf{H}_{(\mathbf{x}=\mathbf{x}_0)}$	definida negativa	$\Longrightarrow \mathbf{x}_0$ é máximo.
$\mathbf{H}_{(\mathbf{x}=\mathbf{x}_0)}$	indefinida	$\Longrightarrow \mathbf{x}_0$ não é máximo nem mínimo

Se $\mathbf{H}_{(\mathbf{x}=\mathbf{x}_0)}$ fôr semi-definida (positiva ou negativa), $\mathbf{x}_0$ poderá ou não ser extremo (mínimo ou máximo, respectivamente), mas não é possível garantir sem uma análise ulterior. Não sendo essencial para os nossos propósitos, não abordaremos ulteriormente esta questão.