Modelos Lineares Generalizados

- da teoria à prática -

M. Antónia Amaral Turkman
DEIO/FC e CEAUL, Universidade de Lisboa

e

Giovani Loiola Silva
DM/IST e CMA, Universidade Técnica de Lisboa

[0]Este trabalho foi parcialmente financiado por FCT - PRAXIS XXI - FEDER

Prefácio

Os *Modelos Lineares Generalizados* foram introduzidos no início dos anos 70, tendo um impacto muito grande no desenvolvimento da estatística aplicada. No princípio o seu uso esteve confinado a um grupo restrito de investigadores, dada a falta de bibliografia acessível e à complexidade inicial do GLIM, primeiro *software* totalmente dirigido à aplicação da metodologia. Foram precisos cerca de 20 anos para que a teoria e prática dos Modelos Lineares Generalizados passasse da "cadeira" do investigador para o domínio público. Este avanço só foi possível a partir do momento em que o *software* existente passou a ser um pouco mais "amigável". Actualmente, a maioria dos pacotes estatísticos de maior expansão já contém módulos adequados ao estudo destes modelos. Pode pois dizer-se que, neste momento, o conhecimento adequado da metodologia dos Modelos Lineares Generalizados é imprescindível para qualquer indivíduo que utilize métodos estatísticos. Daí a escolha deste tema para um novo *mini-curso*, a ser ministrado por ocasião do VIII Congresso Anual da Sociedade Portuguesa de Estatística.

A importância dos Modelos Lineares Generalizados não é apenas de índole prática. Do ponto de vista teórico a sua importância advém, essencialmente, do facto de a metodologia destes modelos constituir uma abordagem unificada de muitos procedimentos es-

tatísticos correntemente usados nas aplicações e promover o papel central da verosimilhança na teoria da inferência.

Nos três primeiros capítulos desta monografia fazemos um desenvolvimento da teoria dos modelos lineares generalizados. Para o efeito admitimos que o leitor domina conceitos básicos de estatística e que está familiarizado com todas as ideias subjacentes ao modelo de regressão normal. Nos dois últimos capítulos aplicamos a teoria exposta a exemplos práticos retirados da literatura, fazendo uso de *software* apropriado. Alguns desses programas *input/output* encontram-se nos apêndices.

O tempo disponível para a concretização desta monografia não nos permitiu atingir todos os objectivos a que no início nos propusemos. Com efeito, gostaríamos de ter abordado, mesmo que de passagem, vários temas importantes, como seja, modelos mistos, generalização a modelos multivariados, abordagem bayesiana e métodos não paramétricos. A ambição era desmedida... Assim ficámo-nos pelo estritamente essencial. Esperamos, no entanto, que este pequeno trabalho seja útil e que num futuro próximo o possamos melhorar substancialmente, nomeadamente com o auxílio de sugestões e críticas que nos queiram fazer chegar às mãos.

Não queremos deixar de terminar este prefácio sem agradecer a ajuda que tivemos na fase final extremamente crítica da escrita desta monografia. Devemos muito à infinita paciência da comissão organizadora do Congresso, que nos deixou trabalhar até "às últimas" e aos conselhos do Paulo Soares.

Lisboa, 16 de Setembro de 2000

Antónia Turkman e Giovani Silva

Índice

Capítulo 1

Introdução aos Modelos Lineares Generalizados

Em muitos estudos estatísticos, quer sejam de natureza experimental ou observacional, somos confrontados com problemas em que o objectivo principal é o de estudar a relação entre variáveis, ou mais particularmente, analisar a influência que uma ou mais variáveis (*explicativas*), medidas em indivíduos ou objectos, têm sobre uma variável de interesse a que damos o nome de *variável resposta*. O modo como, em geral, o estatístico aborda tal problema é através do estudo de um modelo de regressão que relacione essa variável de interesse com as variáveis ditas explicativas.

O modelo linear normal, "criado" no início do século XIX por Legendre e Gauss, dominou a modelação estatística até meados do século XX, embora vários modelos não lineares ou não normais tenham entretanto sido desenvolvidos para fazer face a situações que não eram adequadamente explicadas pelo modelo linear normal. São exemplo disso, tal como referem McCullagh and Nelder

(1989) e Lindsey (1997), o modelo *complementar log-log* para ensaios de diluição (Fisher, 1922), os modelos *probit* (Bliss, 1935) e *logit* (Berkson, 1944; Dyke and Patterson, 1952; Rasch, 1960) para proporções, os modelos *log-lineares* para dados de contagens (Birch, 1963), os modelos de regressão para análise de sobrevivência (Feigl and Zelen, 1965; Zippin and Armitage, 1966; Glasser, 1967).

Todos os modelos anteriormente descritos apresentam uma estrutura de regressão linear e têm em comum, o facto da variável resposta seguir uma distribuição dentro de uma família de distribuições com propriedades muito específicas: a *família exponencial*. Os Modelos Lineares Generalizados introduzidos por Nelder e Wedderburn (1972) correspondem a uma síntese destes e de outros modelos, vindo assim unificar, tanto do ponto de vista teórico como conceptual, a teoria da modelação estatística até então desenvolvida. São pois casos particulares dos modelos lineares generalizados, doravante referido como MLG, os seguintes modelos:

- modelo de regressão linear clássico,
- modelos de análise de variância e covariância,
- modelo de regressão logística,
- modelo de regressão de Poisson,
- modelos log-lineares para tabelas de contingência multidimensionais,
- modelo *probit* para estudos de proporções, etc.

Devido ao grande número de modelos que englobam e à facilidade de análise associada ao rápido desenvolvimento computacional que

se tem verificado nas últimas décadas, os MLG têm vindo a desempenhar um papel cada vez mais importante na análise estatística, apesar das limitações ainda impostas, nomeadamente por manterem a estrutura de linearidade, pelo facto das distribuições se restringirem à família exponencial e por exigirem a independência das respostas. Há já actualmente, na literatura, muitos desenvolvimentos da teoria da modelação estatística onde estes pressupostos são relaxados mas, o não acompanhamento dos modelos propostos com *software* adequado à sua fácil implementação, faz com que se anteveja ainda, por algum tempo, um domínio dos MLG em aplicações de natureza prática.

1.1 Notação e Terminologia; Tipo de Dados

Ao longo de todo este texto iremos estar interessados em situações experimentais em que há uma variável aleatória Y de interesse primário, a que damos o nome de *variável resposta* ou *variável dependente*, e um vector $\mathbf{x} = (x_1, \ldots, x_k)^T$ de k variáveis explicativas, também designadas por *covariáveis* ou *variáveis independentes*, que acreditamos explicar parte da variabilidade inerente a Y. A variável resposta Y pode ser contínua, discreta ou dicotómica. As covariáveis, determinísticas ou estocásticas, podem ser também de qualquer natureza: contínuas, discretas, qualitativas de natureza ordinal ou dicotómicas.

Assumimos que temos dados da forma

$$(y_i, \mathbf{x}_i), \quad i = 1, \ldots, n, \tag{1.1}$$

resultantes da realização de $(Y, \mathbf{x})$ em n indivíduos ou unidades

experimentais, sendo as componentes Y_i do vector aleatório $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ independentes. Irá ser útil, no desenrolar da teoria, a representação dos dados em (1.1) na forma matricial

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} x_{11} & x_{12} & ... & x_{1k} \\ x_{21} & x_{22} & ... & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & ... & x_{nk} \end{pmatrix}. \tag{1.2}$$

Em muitas situações práticas, principalmente quando as variáveis explicativas são de natureza qualitativa, há muitos indivíduos na amostra que partilham do mesmo vector de covariáveis, significando isto que a matriz X tem vários grupos de linhas idênticas. Assim pode ter interesse em apresentar os dados, não desagrupados como em (1.2), mas de uma forma agrupada.

Suponhamos então que podemos associar os indivíduos em g grupos distintos, de tal modo que os n_j indivíduos do grupo j ($j = 1, ..., g$ com $\sum_{j=1}^{g} n_j = n$) partilhem do mesmo vector de covariáveis, digamos $\mathbf{x}_j = (x_{j1}, ..., x_{jk})^T$. Os dados passarão a ser então representados por

$$\overline{\mathbf{y}} = \begin{pmatrix} \overline{y}_1 \\ \overline{y}_2 \\ \vdots \\ \overline{y}_g \end{pmatrix} \qquad X_g = \begin{pmatrix} x_{11} & x_{12} & ... & x_{1k} \\ x_{21} & x_{22} & ... & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{g1} & x_{g2} & ... & x_{gk} \end{pmatrix}, \tag{1.3}$$

onde $\overline{y}_j, j = 1, ..., g$, representa a média das variáveis respostas dos indivíduos que pertencem ao j-ésimo grupo e não existem linhas idênticas em X_g.

O agrupamento dos dados é particularmente importante, e tem significado especial, em situações em que as covariáveis são todas de natureza qualitativa.

1.2 A Família Exponencial

Como já se disse anteriormente, os modelos lineares generalizados pressupõem que a variável resposta tenha uma distribuição pertencente a uma família particular, a família exponencial. A definição que vamos aqui apresentar é a adequada para os modelos para a variável resposta que interessa considerar no âmbito dos MLG. Veja-se, *e.g.*, Cox and Hinkley (1974), para uma definição mais geral de família exponencial k-paramétrica e suas propriedades.

Definição 1 (Família exponencial)

Diz-se que uma variável aleatória Y tem distribuição pertencente à família exponencial de dispersão (ou simplesmente família exponencial) se a sua função densidade de probabilidade (f.d.p.) ou função massa de probabilidade (f.m.p.) se puder escrever na forma

$$f(y|\theta,\phi) = \exp\Big\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\Big\}, \tag{1.4}$$

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

Na definição anterior, θ é a forma canónica do parâmetro de localização e ϕ é um parâmetro de dispersão suposto, em geral, conhecido. Neste caso a distribuição descrita em (1.4) faz parte da família exponencial uniparamétrica. Quando o parâmetro ϕ é desconhecido a distribuição pode ou não fazer parte da família exponencial bi-paramétrica, tal como é geralmente definida (veja, *e.g.*, Cox and Hinkley, 1974). Admite-se, ainda, que a função $b(\cdot)$ é diferenciável e que o suporte da distribuição não depende dos parâmetros. Neste caso prova-se que a família em consideração obedece às condições

habituais de regularidade[1].

1.2.1 Valor médio e variância

Seja $\ell(\theta;\phi,y)=\ln(f(y|\theta,\phi))$. Defina-se a função *score*

$$S(\theta) \;=\; \frac{\partial \ell(\theta;\phi,Y)}{\partial\theta}. \tag{1.5}$$

Sabe-se que para famílias regulares se tem

$$\begin{array}{rcl} E(S(\theta)) & = & 0 \\ E(S^2(\theta)) & = & E\Big[\Big(\frac{\partial \ell(\theta;\phi,Y)}{\partial\theta}\Big)^2\Big] \;=\; -E\Big[\frac{\partial^2 \ell(\theta;\phi,Y)}{\partial\theta^2}\Big] \end{array} \tag{1.6}$$

e portanto como, no caso em que $f(y|\theta,\phi)$ é dado por (1.4),

$$\ell(\theta;\phi,y)=\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi),$$

obtém-se

$$S(\theta) \;=\; \frac{Y-b'(\theta)}{a(\phi)} \qquad \frac{\partial S(\theta)}{\partial\theta} \;=\; -\frac{b''(\theta)}{a(\phi)}, \tag{1.7}$$

onde $b'(\theta)=\frac{\partial b(\theta)}{\partial\theta}$ e $b''(\theta)=\frac{\partial^2 b(\theta)}{\partial\theta^2}$.

Assim de (1.6) e (1.7)

$$E(Y) \;=\; \mu \;=\; a(\phi)E(S(\theta))+b'(\theta) \;=\; b'(\theta) \tag{1.8}$$

$$var(Y) \;=\; a^2(\phi)var(S(\theta)) \;=\; a^2(\phi)\frac{b''(\theta)}{a(\phi)} \;=\; a(\phi)b''(\theta). \tag{1.9}$$

Vê-se assim que a variância de Y é o produto de duas funções; uma, $b''(\theta)$, que depende apenas do parâmetro canónico θ (e portanto do valor médio μ), a que se dá o nome de *função de variância*

[1]Para um estudo de condições de regularidade necessárias no desenvolvimento do estudo que se vai fazer, deve consultar-se um livro avançado de Estatística. Aconselha-se, por exemplo, Sen and Singer (1993).

e que se costuma representar por $V(\mu)$ e outra, $a(\phi)$, que depende apenas do parâmetro de dispersão ϕ.

Em muitas situações de interesse, observa-se que a função $a(\phi)$ toma a forma

$$a(\phi) = \frac{\phi}{\omega},$$

onde ω é uma constante conhecida, obtendo-se portanto a variância de Y como o produto do parâmetro de dispersão por uma função apenas do valor médio.

Neste caso a função definida em (1.4) escreve-se na forma

$$f(y|\theta, \phi, \omega) = \exp\Big\{\frac{\omega}{\phi}(y\theta - b(\theta)) + c(y, \phi, \omega)\Big\}. \tag{1.10}$$

1.2.2 Exemplos

Vejamos alguns exemplos de distribuições conhecidas que pertencem à família em estudo.

Exemplo 1.1 Normal

Se Y segue uma distribuição normal com valor médio μ e variância σ^2 ($Y \sim N(\mu, \sigma^2)$), a f.d.p. de Y é dada por

$$\begin{aligned} f(y|\mu, \sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(y-\mu)}{\sigma^2}^2} \\ &= \exp\Big\{\frac{1}{\sigma^2}\Big(y\mu - \frac{\mu^2}{2}\Big) - \frac{1}{2}\Big(\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\Big)\Big\} \end{aligned}$$

para $y \in I\!R$. Tem-se então que esta função é do tipo (1.10) com

$$\begin{aligned} \theta &= \mu, \\ b(\theta) &= \frac{\mu^2}{2}, \qquad & c(y, \phi) &= -\frac{1}{2}\Big(\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\Big), \\ b'(\theta) &= \mu, & b''(\theta) &= V(\mu) = 1, \end{aligned}$$

$$\begin{aligned} a(\phi) &= \frac{\phi}{\omega}, \qquad \phi = \sigma^2, \qquad \omega = 1, \\ var(Y) &= b''(\theta)a(\phi) = \sigma^2. \end{aligned}$$

De (1.8) e (1.9) temos, aliás como é bem sabido, que

$$E(Y) = \mu, \quad var(Y) = \sigma^2.$$

A função de variância é, neste caso, $V(\mu) = 1$; o parâmetro canónico é o valor médio μ e σ^2 é o parâmetro de dispersão.

Exemplo 1.2 Binomial

Se Y for tal que mY tem uma distribuição binomialdistribuição normal com parâmetros m e π ($Y \sim B(m,\pi)/m$), a sua f.m.p. é dada por

$$\begin{aligned} f(y|\pi) &= \binom{m}{ym}\pi^{ym}(1-\pi)^{m-ym} \\ &= \exp\left\{ym\ln\pi + m(1-y)\ln(1-\pi) + \ln\binom{m}{ym}\right\} \\ &= \exp\left\{m(y\theta - \ln(1+e^{\theta})) + \ln\binom{m}{ym}\right\} \end{aligned}$$

com $y \in \{0, \frac{1}{m}, \frac{2}{m}, ..., 1\}$ e $\theta = \ln\left(\frac{\pi}{1-\pi}\right)$.

Vê-se assim que esta f.m.p. é da forma (1.10) com

$$\begin{aligned} \theta &= \ln\left(\frac{\pi}{1-\pi}\right), \\ b(\theta) &= \ln(1+e^{\theta}), \qquad c(y,\phi) = \ln\binom{m}{ym} \\ b'(\theta) &= \frac{e^{\theta}}{1+e^{\theta}} = \pi, \qquad b''(\theta) = V(\mu) = \frac{e^{\theta}}{(1+e^{\theta})^2} = \pi(1-\pi), \\ a(\phi) &= \frac{\phi}{\omega}, \qquad \phi = 1, \qquad \omega = m, \end{aligned}$$

De (1.8) e (1.9) obtém-se directamente

$$E(Y) = b'(\theta) = \pi, \quad var(Y) = b''(\theta)a(\phi) = \frac{\pi(1-\pi)}{m}.$$

O parâmetro canónico é a função *logit*, $\ln\left(\frac{\pi}{1-\pi}\right)$.

Exemplo 1.3 Gama

Se Y tem distribuição gama com parâmetros de forma ν e de escala ν/μ ($Y \sim Ga(\nu, \nu/\mu)$), a sua f.d.p. é

$$\begin{aligned} f(y|\nu,\mu) &= \frac{1}{\Gamma(\nu)}\Big(\frac{\nu}{\mu}\Big)^{\nu} y^{\nu-1} \exp\Big(-\frac{\nu}{\mu}y\Big) \\ &= \exp\Big\{\nu\Big(-\frac{y}{\mu} - \ln\mu\Big) + (\nu-1)\ln y - \ln\Gamma(\nu) + \nu\ln\nu\Big\} \\ &= \exp\{\nu(\theta y + \ln(-\theta)) + (\nu-1)\ln y - \ln\Gamma(\nu) + \nu\ln\nu\} \end{aligned}$$

com $y > 0$ e $\theta = -\frac{1}{\mu}$.

Temos novamente uma f.d.p. da forma (1.10) com

$$\begin{aligned} \theta &= -\frac{1}{\mu}, \\ b(\theta) &= -\ln(-\theta), \qquad c(y,\phi) = (\nu-1)\ln y + \nu\ln\nu - \ln\Gamma(\nu), \\ b'(\theta) &= -\frac{1}{\theta}, \qquad\qquad b''(\theta) = V(\mu) = \frac{1}{\theta^2} = \mu^2, \\ a(\phi) &= \frac{\phi}{\omega}, \quad \phi = \frac{1}{\nu}, \quad \omega = 1. \end{aligned}$$

Novamente de (1.8) e (1.9) tem-se que

$$E(Y) = -\frac{1}{\theta} = \mu, \quad var(Y) = \frac{\mu^2}{\nu}.$$

A função de variância é, neste caso, $V(\mu) = \mu^2$ e o parâmetro de dispersão é $\frac{1}{\nu}$.

Na tabela 1.1 apresentamos uma lista das principais distribuições que pertencem à família exponencial com a respectiva caracterização.

Tabela 1.1: Algumas distribuições da família exponencial.

distribuição	normal	binomial	Poisson	gama	gaussiana inversa
Notação	$N(\mu, \sigma^2)$	$B(m, \pi)/m$	$P(\lambda)$	$Ga(\nu, \frac{\nu}{\mu})$	$IG(\mu, \sigma^2)$
Suporte	$(-\infty, +\infty)$	$\{0, \frac{1}{m}, ..., 1\}$	$\{0, 1, ...\}$	$(0, +\infty)$	$(0, +\infty)$
θ	μ	$\ln\left(\frac{\pi}{1-\pi}\right)$	$\ln \lambda$	$-\frac{1}{\mu}$	$-\frac{1}{2\mu^2}$
$a(\phi)$	σ^2	$\frac{1}{m}$	1	$\frac{1}{\nu}$	σ^2
ϕ	σ^2	1	1	$\frac{1}{\nu}$	σ^2
ω	1	m	1	1	1
$c(y, \phi)$	$-\frac{1}{2}(\frac{y^2}{\phi} + \ln(2\pi\phi))$	$\ln\binom{m}{my}$	$-\ln y!$	$\nu \ln \nu - \ln \Gamma(\nu)$ $+(\nu - 1) \ln y$	$-\frac{1}{2}\{\ln(2\pi\phi y^3)$ $+\frac{1}{y\phi}\}$
$b(\theta)$	$\frac{\theta^2}{2}$	$\ln(1 + e^\theta)$	e^θ	$-\ln(-\theta)$	$-(-2\theta)^{1/2}$
$b'(\theta) = E(Y)$	θ	$\pi = \frac{e^\theta}{1+e^\theta}$	$\lambda = e^\theta$	$\mu = -\frac{1}{\theta}$	$\mu = (-2\theta)^{-1/2}$
$b''(\theta) = V(\mu)$	1	$\pi(1 - \pi)$	λ	μ^2	μ^3
$var(Y)$	σ^2	$\frac{\pi(1-\pi)}{m}$	λ	$\frac{\mu^2}{\nu}$	$\mu^3\sigma^2$

1.3 Descrição do Modelo Linear Generalizado

Os modelos lineares generalizados são uma extensão do modelo linear clássico

$$\mathbf{Y} = Z\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

onde Z é uma matriz de dimensão $n \times p$ de especificação do modelo (em geral a matriz de covariáveis X com um primeiro vector unitário), associada a um vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ de parâmetros, e $\boldsymbol{\varepsilon}$ é um vector de erros aleatórios com distribuição que se supõe $N_n(\mathbf{0}, \sigma^2\mathbf{I})$.

Estas hipóteses implicam obviamente que $E(\mathbf{Y}|Z) = \boldsymbol{\mu}$ com $\boldsymbol{\mu} = Z\boldsymbol{\beta}$, ou seja, o valor esperado da variável resposta é uma função linear das covariáveis.

A extensão mencionada é feita em duas direcções. Por um lado, a distribuição considerada não tem de ser normal, podendo ser qualquer distribuição da família exponencial; por outro lado, embora se mantenha a estrutura de linearidade, a função que relaciona o valor esperado e o vector de covariáveis pode ser qualquer função diferenciável.

Assim os MLG são caracterizados pela seguinte estrutura:

1. *Componente aleatória*

   Dado o vector de covariáveis $\mathbf{x}_i$ as variáveis Y_i são (condicionalmente) independentes com distribuição pertencente à *família exponencial* da forma (1.4) ou (1.10), com $E(Y_i|\mathbf{x}_i) = \mu_i = b'(\theta_i)$ para $i = 1, .., n$ e, possivelmente, um parâmetro de dispersão ϕ não dependente de i.

2. *Componente estrutural ou sistemática*

   O valor esperado μ_i está relacionado com o *preditor linear* $\eta_i = \mathbf{z}_i^T \boldsymbol{\beta}$ através da relação

   $$\mu_i = h(\eta_i) = h(\mathbf{z}_i^T \boldsymbol{\beta}), \qquad \eta_i = g(\mu_i),$$

   onde

   - h é uma função monótona e diferenciável;
   - $g = h^{-1}$ é a *função de ligação*;
   - $\boldsymbol{\beta}$ é um vector de parâmetros de dimensão p;
   - $\mathbf{z}_i$ é um vector de especificação de dimensão p, função do vector de covariáveis $\mathbf{x}_i$.

Em geral $\mathbf{z}_i = (1, x_{i1}, ..., x_{ik})^T$ com $k = p - 1$. Contudo, quando existem covariáveis qualitativas elas têm de ser, em geral, convenientemente codificadas à custa de variáveis binárias mudas (*dummy*); por exemplo, se uma variável qualitativa (ou factor) tem q categorias, então são necessárias $q - 1$ variáveis binárias para a representar. Essas variáveis têm então de ser incluídas no vector $\mathbf{z}$.

A escolha da função de ligação depende do tipo de resposta e do estudo particular que se está a fazer. Na tabela 1.2 apresentamos uma lista das principais funções de ligação que se costumam considerar.

Tem especial interesse a situação em que o preditor linear coincide com o parâmetro canónico, isto é, $\theta_i = \eta_i$, o que obviamente implica $\theta_i = \mathbf{z}_i^T \boldsymbol{\beta}$. A função de ligação correspondente diz-se então *função de ligação canónica*.

Uma vantagem em usar a função de ligação canónica é que, nesse caso, desde que o parâmetro de escala seja conhecido, o vector

Tabela 1.2: Funções de ligação.

identidade	recíproca	quadrática inversa
μ	$\frac{1}{\mu}$	$\frac{1}{\mu^2}$
raiz quadrada	expoente	logarítmica
$\sqrt{\mu}$	$(\mu + c_1)^{c_2}$	$\ln(\mu)$
logit	complementar log-log	*probit*
$\ln(\frac{\mu}{1-\mu})$	$\ln[-\ln(1-\mu)]$	$\Phi^{-1}(\mu)$

parâmetro desconhecido da estrutura linear admite uma estatística suficiente mínima de dimensão fixa. Esta questão irá ser abordada mais tarde na secção 2.1.2.

Por fim, note-se que os modelos lineares generalizados englobam uma boa parte dos modelos mais populares na análise estatística de dados como se ilustra na tabela 1.3.

Tabela 1.3: Alguns Modelos Lineares Generalizados.

componente	componente estrutural		modelo
aleatória	f. ligação	covariáveis	
normal	identidade	contínuas	regressão linear
normal	identidade	categorizadas	análise de variância
normal	identidade	mistas	análise de covariância
binomial	*logit*	mistas	regressão logística
Poisson	logarítmica	mistas	log-linear

1.4 Exemplos de Modelos Lineares Generalizados

Nesta secção remos apresentar alguns dos modelos lineares generalizados mais comuns nas aplicações. Convém distinguir modelos para três tipos de respostas: (i) de natureza contínua, (ii) de natureza dicotómica, ou na forma de proporções e (iii) na forma de contagens. Por essa razão apresentamos os exemplos agrupados de acordo com essa divisão.

1.4.1 Modelos para respostas contínuas

Modelo normal

Já se referiu anteriormente que os MLG correspondem a uma generalização do modelo de regressão linear. Com efeito, se tivermos n respostas independentes $Y_i \sim N(\mu_i, \sigma^2)$, $i = 1..., n$ onde

$$\mu_i = \mathbf{z}_i^T \boldsymbol{\beta} = \sum_{j=1}^{p} z_{ij}\beta_j,$$

o modelo considerado é um modelo linear generalizado, dado que:

- as variáveis resposta são independentes,
- a distribuição é da forma (1.10), com $\theta_i = \mathbf{z}_i^T\boldsymbol{\beta}$, $\phi = \sigma^2$ e $\omega_i = 1$.
- o valor esperado μ_i está relacionado com o *preditor linear* $\eta_i = \mathbf{z}_i^T\boldsymbol{\beta}$ através da relação $\mu_i = \eta_i$,
- a *função de ligação* é a função identidade, que é, neste caso a função de ligação canónica.

Para este modelo podemos escrever $Y_i = \mathbf{z}_i^T\boldsymbol{\beta} + \varepsilon_i,\ i = 1, ..., n$ onde os ε_i são i.i.d. $N(0, \sigma^2)$.

Note-se ainda que a formulação apresentada inclui facilmente o caso especial em que $Y_i \sim N(\mu_i, \sigma_i^2)$, com $\sigma_i^2 = \frac{\sigma^2}{\omega_i}$, onde ω_i é um peso conhecido associado à i-ésima observação.

O modelo normal (modelo linear clássico), introduzido atrás, pressupõe, como se sabe, que a variância das resposta seja constante. Contudo, na prática, surgem por vezes situações de variáveis resposta de natureza contínua, em que a variância não é constante. Uma transformação que se usa com frequência para estabilizar a variância, é a transformação logarítmica, a qual é possível se as respostas forem positivas. Admitindo então que o logaritmo das respostas Y_i segue uma distribuição normal, pode considerar-se um modelo de regressão linear clássico para o logaritmo das respostas. Neste caso, ter-se-á a relação $\eta_i = E\{\ln(Y_i)\} = \mathbf{z}_i^T\boldsymbol{\beta}$. Por várias razões e, em particular, se há necessidade das conclusões serem apresentadas na escala original das respostas, então é mais conveniente assumir que $\ln E(Y_i) = \mathbf{z}_i^T\boldsymbol{\beta}$, ou seja, que $E(Y_i) = \exp(\mathbf{z}_i^T\boldsymbol{\beta})$. Se, por outro lado, assumirmos que a variância aumenta com o valor médio de modo que o coeficiente de variação se mantém constante, o modelo gama passa a ser um modelo adequado para as respostas (McCullagh and Nelder, 1989).

Modelo gama

Admitindo então que as respostas são variáveis aleatórias $Y_i \sim Ga(\nu, \frac{\nu}{\mu_i})$ independentes, com $\mu_i = \exp(\mathbf{z}_i^T\boldsymbol{\beta})$ obtém-se um modelo linear generalizado (o modelo de regressão gama), visto que:

- as variáveis resposta são independentes,

- a distribuição é da forma (1.10), com $\theta_i = -\frac{1}{\exp(\mathbf{z}_i^T\boldsymbol{\beta})}$, $\phi = \frac{1}{\nu}$ e $\omega_i = 1$,
- o valor esperado μ_i está relacionado com o *preditor linear* $\eta_i = \mathbf{z}_i^T\boldsymbol{\beta}$ através da relação $\mu_i = \exp(\eta_i)$,
- a *função de ligação* é a função logarítmica.

Neste caso podemos escrever o modelo na forma $Y_i = \exp(\mathbf{z}_i^T\boldsymbol{\beta})\,\epsilon_i$, $i = 1, ..., n$ com ϵ_i i.i.d. $Ga(\nu, \nu)$.

Um exemplo de aplicação deste modelo é feito no capítulo de aplicações práticas.

Note-se que a função de ligação considerada não é a função de ligação canónica. A função de ligação canónica obtém-se quando $\eta_i = \theta_i$, o que neste caso corresponde a ter $-\frac{1}{\mu_i} = \mathbf{z}_i^T\boldsymbol{\beta}$. A função de ligação canónica é então a função recíproca.

Dado que $\mu_i > 0$, a utilização do modelo gama com função de ligação canónica implica que se têm de impor restrições aos valores possíveis para os parâmetros β_j da estrutura linear. Nelder (1966) apresenta um exemplo de aplicação de regressão gama com função de ligação canónica.

Modelo gaussiano inverso

Suponhamos que $Y_i \sim IG(\mu_i, \sigma^2)$, isto é

$$f(y_i|\mu_i, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2 y_i^3}} \exp\Big\{ - \frac{(y_i - \mu_i)^2}{2\mu_i^2\sigma^2 y_i}\Big\}, \qquad y_i > 0$$

e $\mu_i = (\exp\{\mathbf{z}_i^T\boldsymbol{\beta}\})^{\frac{1}{2}}$, para $i = 1, ..., n$.

Neste caso obtém-se, como facilmente se verifica, um modelo linear generalizado com função de ligação canónica.

Este modelo é útil para estudos de análise de regressão com dados consideravelmente assimétricos e, em particular, no caso em que as respostas representam tempos de vida.

1.4.2 Modelos para dados binários ou na forma de proporções

Suponhamos que temos n variáveis resposta independentes $Y_i \sim B(1, \pi_i)$, *i.e.*,

$$f(y_i|\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, \qquad y_i = 0, 1,$$

e que a cada indivíduo i ou unidade experimental, está associado um vector de especificação $\mathbf{z}_i$, resultante do vector de covariáveis $\mathbf{x}_i$, $i = 1, \ldots, n$.

Como $E(Y_i) = \pi_i$ e, de acordo com a tabela 1.1, se tem para este modelo, $\theta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$, ao fazer

$$\theta_i = \eta_i = \mathbf{z}_i^T\boldsymbol{\beta},$$

concluímos que a função de ligação canónica é a função *logit*. Assim a probabilidade de sucesso, ou seja $P(Y_i = 1) = \pi_i$, está relacionada com o vector $\mathbf{z}_i$ através de

$$\pi_i = \frac{\exp(\mathbf{z}_i^T\boldsymbol{\beta})}{1+\exp(\mathbf{z}_i^T\boldsymbol{\beta})}. \tag{1.11}$$

É fácil de ver que a função $F : I\!R \to [0, 1]$, definida por

$$F(x) = \frac{\exp(x)}{1+\exp(x)},$$

é uma função de distribuição. Ela é, com efeito, a *função de distribuição logística*. Por esse motivo, o MLG definido pelo modelo

binomial com função de ligação canónica (*logit*) é conhecido por *modelo de regressão logística*.

Repare-se que devido ao facto de, neste modelo, se ter $E(Y_i) = \mu_i \in [0,1]$, em princípio, não só a função de distribuição logística, como qualquer outra função de distribuição, pode ser candidata a função inversa da função de ligação. Nomeadamente podemos supor que a relação existente entre as probabilidades de sucesso π_i e o vector de covariáveis é da forma

$$\pi_i = \Phi(\eta_i) = \Phi(\mathbf{z}_i^T \boldsymbol{\beta}), \tag{1.12}$$

onde $\Phi(\cdot)$ é a função de distribuição de uma variável aleatória $N(0,1)$. Obtemos assim uma função de ligação $g(\mu_i) = \Phi^{-1}(\mu_i)$ designada por função de ligação *probit*.

O modelo linear generalizado, obtido pela associação do modelo binomial para as respostas, com a função de ligação *probit* conduz ao *modelo de regressão probit*.

Outra função de distribuição que se costuma considerar para candidata a função inversa da função de ligação, é a função de distribuição de Gumbel, ou função de distribuição de extremos,

$$F(x) = 1 - \exp(-\exp(x)), \qquad x \in I\!R.$$

Considerando então

$$h(\mathbf{z}_i^T \boldsymbol{\beta}) = 1 - \exp(-\exp(\mathbf{z}_i^T \boldsymbol{\beta})) = \pi_i,$$

obtém-se a função *complementar log-log*

$$\ln(-\ln(1-\pi_i)) = \mathbf{z}_i^T \boldsymbol{\beta} \tag{1.13}$$

para função de ligação.

O modelo linear generalizado, obtido pela associação do modelo binomial para as respostas, com a função de ligação *complementar log-log* conduz ao *modelo de regressão complementar log-log*.

A utilização de uma ou outra função de ligação, e consequentemente, a escolha do modelo de regressão a utilizar, depende da situação em causa. Em geral, a adaptabilidade dos modelos *probit* e logístico é bastante semelhante, já que as funções correspondentes não se afastam muito uma da outra após um ajustamento adequado dos correspondentes preditores lineares. O modelo *complementar log-log* pode dar respostas diferentes já que a função *complementar log-log*, mesmo após o ajustamento do preditor linear η, se distancia das anteriores, tendo um crescimento mais abrupto (ver, Fahrmeir and Tutz, 1994, pg. 27). A função de ligação *complementar log-log* é mais utilizada para análise de dados sobre incidência de doenças.

Nos capítulos dedicados a aplicações práticas apresentaremos exemplos de modelos de regressão logístico e *probit*.

Relação com modelos lineares latentes

Variáveis aleatórias binárias podem ser consideradas como resultantes da dicotomização de variáveis aleatórias contínuas. Com efeito, se $\mathcal{Z}$ for uma variável aleatória contínua com função de distribuição $F_{\mathcal{Z}}(\cdot)$, e se, em vez de se observar $\mathcal{Z}$, se observar apenas se $\mathcal{Z}$ está acima ou abaixo um determinado valor crítico r, a variável aleatória

$$Y = \begin{cases} 1, & \text{se } \mathcal{Z} \leq r \\ 0, & \text{se } \mathcal{Z} > r, \end{cases}$$

é uma variável aleatória de Bernoulli com probabilidade de sucesso $\pi = P(Y = 1) = F_{\mathcal{Z}}(r)$. $\mathcal{Z}$ será assim uma variável aleatória latente não observada.

Os modelos para dados binários apresentados podem, deste modo, ser explicados como resultantes de modelos lineares latentes. Com efeito, se

$$\mathcal{Z} = \mathbf{z}^T\boldsymbol{\alpha} + \sigma\varepsilon,$$

onde σ é um parâmetro de escala desconhecido, $\boldsymbol{\alpha}^T = (\alpha_1, ..., \alpha_p)$, $\mathbf{z} = (1, z_2, ..., z_p)^T$ é um vector de especificação e ε tiver distribuição $F(\cdot)$ (*e.g.*, logística, normal reduzida, ou de extremos) então

$$\begin{aligned}
\pi = P(Y = 1) &= P(\mathcal{Z} \leq r) \\
&= P(\mathbf{z}^T\boldsymbol{\alpha} + \sigma\varepsilon \leq r) \\
&= P(\varepsilon \leq \frac{r - \alpha_1}{\sigma} - \sum_{j=2}^{p} z_j \frac{\alpha_j}{\sigma}) \\
&= F(\mathbf{z}^T\boldsymbol{\beta}),
\end{aligned}$$

com $\beta_1 = \frac{r-\alpha_1}{\sigma}$ e $\beta_j = -\frac{\alpha_j}{\sigma}$, $j = 2, ..., p$.

A abordagem de um modelo de dados binários, na perspectiva de um modelo linear latente, permite a interpretação dos parâmetros β's em função desse modelo; no entanto, se σ for desconhecido os efeitos das covariáveis no modelo linear ($\alpha_j, j = 2, .., p$) só são conhecidos a menos do factor $\frac{1}{\sigma}$ e portanto só se pode atribuir significado aos valores relativos dos parâmetros (*e.g.*, β_2/β_3) e não aos seus valores absolutos.

Dados agrupados

Até aqui estivemos a supor que os dados se encontram numa forma não agrupada. Em muitas situações de interesse acontece, como já foi referido na secção 1.1, que vários indivíduos, ou unidades experimentais, partilham do mesmo vector de covariáveis, podendo então os indivíduos ser agrupados de acordo com os diferentes *padrões*

de covariáveis. Neste caso consideramos como variável resposta a frequência relativa de sucessos do grupo, *i.e.*, $\overline{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$, onde n_i é o número de indivíduos no i-ésimo grupo. Como as frequências absolutas têm distribuição $B(n_i, \pi_i)$, as frequências relativas são distribuídas de acordo com $B(n_i, \pi_i)/n_i$, *i.e.*

$$P(\overline{Y}_i = \overline{y}_i|\pi_i) = \binom{n_i}{n_i\overline{y}_i}\pi_i^{n_i\overline{y}_i}(1-\pi_i)^{n_i - n_i\overline{y}_i} \qquad \overline{y}_i = 0, \frac{1}{n_i}, ..., 1.$$

Repare-se que, como ainda se tem $E(\overline{Y}_i) = \pi_i$, podem ainda considerar-se as mesmas funções de ligação que se consideraram para o caso em que as respostas são binárias. A mesma metodologia pode também ser aplicada no caso em que as respostas individuais Y_i (não agrupadas) são $B(n_i, \pi_i)$ bastando para tal considerar como resposta a variável $\frac{Y_i}{n_i}$.

Sobredispersão ou Extra variação binomial

Um fenómeno que ocorre com frequência nas aplicações é as respostas apresentarem uma variância superior à variância explicada pelo modelo binomial. Este fenómeno, denominado de *sobredispersão* ou *extra variação binomial*, pode ser devido ao facto de existir heterogeneidade entre os indivíduos não explicada pelas covariáveis, ou pelo facto de haver correlação entre as respostas. Esta última situação acontece quando, por exemplo, as respostas correspondem a indivíduos da mesma família, ou a indivíduos que comungam dos mesmos factores ambientais, formando assim grupos naturais, embora a heterogeneidade não explicada também produza correlação entre as respostas. Este problema pode ser resolvido se se introduzir um parâmetro $\phi > 1$ de *sobredispersão* de tal modo

que

$$var(Y_i|\mathbf{x}_i) = \phi\frac{\pi_i(1-\pi_i)}{n_i},$$

onde $n_i > 1$ é a dimensão do grupo. Note-se, no entanto, que já não é possível escrever a distribuição de Y_i na forma da família exponencial (1.4). O modelo fica apenas determinado pelo valor médio e variância.

Mais detalhes para tratar do problema de sobredispersão na família binomial podem ser encontrados, *e.g.*, em Collet (1991) e Fahrmeir and Tutz (1994).

1.4.3 Modelos para respostas na forma de contagens

Dados na forma de contagens aparecem com muita frequência nas aplicações. São exemplo disso o número de acidentes, o número de chamadas telefónicas, o número de elementos numa fila de espera, etc. Também são dados deste tipo as frequências em cada célula de uma tabela de contingência. O modelo de Poisson, como se sabe, desempenha um papel fundamental na análise deste tipo de dados. Como também se já referiu na secção 1.2.2 este é um modelo na família exponencial que tem a particularidade de o valor médio ser igual à variância.

Se se considerar que as respostas Y_i são independentes e bem modeladas por uma distribuição de Poisson de valor médio μ_i e que $\ln(\mu_i) = \mathbf{z}_i^T\boldsymbol{\beta}$ com $i = 1, ..., n$, *i.e.*

$$\begin{aligned} f(y_i|\mathbf{x}_i) &= e^{-\mu_i}\frac{\mu_i^{y_i}}{y_i!} \\ &= \exp\{-e^{\mathbf{z}_i^T\boldsymbol{\beta}} + y_i\mathbf{z}_i^T\boldsymbol{\beta} - \ln y_i!\}, \quad y_i = 0, 1, ..., \end{aligned}$$

obtém-se um modelo linear generalizado com função de ligação canónica, conhecido por *modelo de regressão de Poisson*, ou *modelo log-linear*.

Para o caso do modelo de Poisson, a função logarítmica é a função de ligação que geralmente se utiliza.

Sob condições bastante fracas, pode mostrar-se que a análise de uma tabela de contingência sob amostragem de Poisson, é a mesma que a análise sob amostragem multinomial ou produto-multinomial (e.g., Christensen, 1997). Assim, o modelo de regressão de Poisson é também útil na modelação e estudo de tabelas de contingência multidimensionais, apesar de em tabelas de contingência as observações não serem independentes.

A imposição pelo modelo Poisson da variância ser igual ao valor médio, produz, também com frequência, problemas de sobredispersão idênticos ao referido anteriormente para dados de natureza binária. O modo mais simples de resolver o problema é, novamente, o de considerar um parâmetro de sobredispersão ϕ de tal modo que $var(Y_i|\mathbf{x}_i) = \phi\mu_i$ para $i = 1, ..., n$. Há, no entanto, modelos mais complexos que entram em consideração com variação extra nos dados. Veja-se, *e.g.*, Breslow (1984) e Fahrmeir and Tutz (1994).

1.5 Metodologia dos Modelos Lineares Generalizados

Há três etapas essenciais que devemos seguir ao tentar modelar dados através de um MLG:

- Formulação dos modelos,

- Ajustamento dos modelos,
- Selecção e validação dos modelos.

Na **formulação do modelo** há que entrar em consideração com

(i) escolha da distribuição para a variável resposta. Para isso há necessidade de examinar cuidadosamente os dados; por exemplo, a distribuição gama e normal inversa são apropriadas para modelar dados de natureza contínua e que mostram assimetrias; por vezes pode haver necessidade de transformar previamente os dados, etc.

Assim, uma análise preliminar dos dados, é fundamental para que se possa fazer uma escolha adequada da família de distribuições a considerar.

(ii) escolha das covariáveis e formulação apropriada da matriz de especificação. Aqui há que entrar em linha de conta com o problema específico em estudo e, muito particularmente, ter em atenção a codificação apropriada das variáveis de natureza qualitativa, criando nomeadamente, caso se revele necessário, variáveis mudas, para correctamente definir variáveis de natureza policotómica.

(iii) escolha da função de ligação. A escolha de uma função de ligação compatível com a distribuição do erro proposto para os dados deve resultar de uma combinação de considerações a priori sobre o problema em causa, exame intensivo dos dados, facilidade de interpretação do modelo, etc. Existem funções de ligação que produzem propriedades estatísticas desejadas para o modelo, como iremos ver na secção 2.1.2, mas a conveniência matemática por si só não deve determinar a escolha da função de ligação.

A fase do **ajustamento do modelo** (ou modelos) passa pela estimação dos parâmetros do modelo, isto é, pela estimação dos coeficientes β's associados às covariáveis, e do parâmetro de dispersão

ϕ caso ele esteja presente. É importante também nesta fase estimar parâmetros que representam medidas da adequabilidade dos valores estimados, obter intervalos de confiança e realizar testes de bondade de ajustamento. O problema da inferência em modelos lineares generalizados será tratado no capítulo 2.

Nos problemas em que a metodologia dos MLG tem cabimento, ou seja em problemas de regressão, há em geral um número considerável de possíveis variáveis explicativas. A fase de **selecção e validação dos modelos** tem por objectivo encontrar submodelos com um número moderado de parâmetros que ainda sejam adequados aos dados, detectar discrepâncias entre os dados e os valores preditos, averiguar da existência de *outliers* ou/e observações influentes, etc. Na selecção do melhor modelo para explicar o problema em estudo, devem ainda ser ponderados três factores: adequabilidade, parcimónia e interpretação. Um bom modelo é aquele que consegue atingir um equilíbrio entre esses três factores.

O problema da selecção e validação dos modelos será tratado no capítulo 3.

Capítulo 2

Inferência

De modo a poder aplicar a metodologia dos modelos lineares generalizados a um conjunto de dados há necessidade, após a formulação do modelo que se pensa adequado, de proceder à realização de inferências sobre esse modelo. A inferência com MLG é, essencialmente, baseada na verosimilhança. Com efeito, não só o método da máxima verosimilhança é o método de eleição para estimar os parâmetros de regressão, como também testes de hipóteses sobre os parâmetros do modelo e de qualidade de ajustamento são, em geral, métodos baseados na verosimilhança.

Os métodos inferenciais que vamos discutir neste capítulo pressupõem que o modelo está completamente e correctamente especificado, de acordo com a formulação apresentada na secção 1.3. Contudo, por vezes, essa suposição não é realista. É o caso, por exemplo, em que se verifica que há sobredispersão num modelo de Poisson ou binomial e que portanto há necessidade de alterar a variância através da introdução de um parâmetro de sobredispersão, como se viu na secção 1.4. Nessa situação já não é possível especi-

ficar completamente o modelo, uma vez que não existe uma distribuição apropriada dentro da família exponencial que tenha aqueles valor médio e variância. O modelo fica assim apenas especificado pelo valor médio e pela variância. Este problema pode ser ultrapassado, e ainda é possível realizar inferências, utilizando a ideia de *modelos de quasi-verosimilhança*. Esta questão será brevemente abordada no fim do capítulo.

2.1 Estimação

Consideremos que temos dados na forma $(y_i, \mathbf{x}_i), i = 1, ..., n$, como em (1.1), onde y_i é o valor observado da variável resposta para a i-ésima unidade experimental e $\mathbf{x}_i$ é o correspondente vector de covariáveis. Designemos ainda por $\mathbf{z}_i = \mathbf{z}_i(\mathbf{x})$ o vector de especificação de dimensão p, associado ao vector de covariáveis e já previamente definido no capítulo anterior. Para evitar complexidades no desenrolar da teoria, admitimos ainda que a matriz de especificação Z dada por:

$$Z = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n)^T = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{pmatrix} \tag{2.1}$$

é de característica completa ("*full rank*"), isto é, tem característica igual à ordem p (mínimo entre n e p já que se assume $n > p$). Isto implica que a matriz $Z^T Z$ tem característica p.

A metodologia de estimação que vamos apresentar neste capítulo não é alterada, quer se tenham os dados desagrupados (amostra de dimensão n), ou os dados agrupados (amostra de dimensão g). Por

facilidade de exposição, e sem perda de generalidade, suporemos sempre que a dimensão é n.

Num modelo linear generalizado o parâmetro $\boldsymbol{\beta}$ é o parâmetro de interesse, o qual é estimado pelo método da máxima verosimilhança. O parâmetro de dispersão ϕ, quando existe, é considerado um parâmetro perturbador, sendo a sua estimação feita pelo método dos momentos.

2.1.1 Verosimilhança e matriz de informação de Fisher

Consideremos então um modelo linear generalizado definido, tal como na secção 1.3 por:

$$f(y_i|\theta_i,\phi,\omega_i) = \exp\Big\{\frac{\omega_i}{\phi}(y_i\theta_i - b(\theta_i)) + c(y_i,\phi,\omega_i)\Big\}, \qquad (2.2)$$

com função de ligação $g(\mu_i) = \eta_i = \mathbf{z}_i^T\boldsymbol{\beta}$, sendo as variáveis aleatórias Y_i independentes.

No que se segue convém recordar ainda que θ_i é função de μ_i, sendo $b'(\theta_i) = \mu_i = h(\eta_i)$ onde $h(\cdot)$ é a função inversa da função de ligação $g(\cdot)$ e que $var(Y_i) = \frac{\phi}{\omega_i}b''(\theta_i)$.

A função de verosimilhança, como função de $\boldsymbol{\beta}$, é

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} f(y_i|\theta_i,\phi,\omega_i) \\ &= \prod_{i=1}^{n} \exp\Big\{\frac{\omega_i}{\phi}(y_i\theta_i - b(\theta_i)) + c(y_i,\phi,\omega_i)\Big\} \\ &= \exp\Big\{\frac{1}{\phi}\sum_{i=1}^{n}\omega_i(y_i\theta_i - b(\theta_i)) + \sum_{i=1}^{n} c(y_i,\phi,\omega_i)\Big\} \qquad (2.3) \end{aligned}$$

e portanto o logaritmo da função de verosimilhança (que passaremos

a chamar de *log-verosimilhança*) é dado por

$$\begin{aligned} \ln L(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{\omega_i(y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi, \omega_i) \\ &= \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}), \end{aligned} \tag{2.4}$$

onde

$$\ell_i(\boldsymbol{\beta}) = \frac{\omega_i(y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi, \omega_i) \tag{2.5}$$

é a contribuição de cada observação y_i para a verosimilhança.

Admitindo que se verificam certas condições de regularidade (ver, *e.g.*, Sen and Singer, 1993) os estimadores de máxima verosimilhança para $\boldsymbol{\beta}$ são obtidos como solução do sistema de equações de verosimilhança

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_j} = 0, \quad j = 1, ..., p.$$

Para obter estas equações escrevemos:

$$\frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_j} = \frac{\partial \ell_i(\theta_i)}{\partial \theta_i} \frac{\partial \theta_i(\mu_i)}{\partial \mu_i} \frac{\partial \mu_i(\eta_i)}{\partial \eta_i} \frac{\partial \eta_i(\boldsymbol{\beta})}{\partial \beta_j}$$

sendo

$$\begin{aligned} \frac{\partial \ell_i(\theta_i)}{\partial \theta_i} &= \frac{\omega_i(y_i - b'(\theta_i))}{\phi} = \frac{\omega_i(y_i - \mu_i)}{\phi}, \\ \frac{\partial \mu_i}{\partial \theta_i} &= b''(\theta_i) = \frac{\omega_i var(Y_i)}{\phi}, \\ \frac{\partial \eta_i(\boldsymbol{\beta})}{\partial \beta_j} &= z_{ij}. \end{aligned}$$

Assim

$$\frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_j} = \frac{\omega_i(y_i - \mu_i)}{\phi} \frac{\phi}{\omega_i var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} z_{ij} \tag{2.6}$$

e as equações de verosimilhança para $\boldsymbol{\beta}$ são

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) z_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \qquad j = 1, ..., p. \tag{2.7}$$

A função *score*, tal como já foi definida em (1.5), é o vector p-dimensional

$$s(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} s_i(\boldsymbol{\beta}),$$

onde $s_i(\boldsymbol{\beta})$ é o vector de componentes $\frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_j}$ obtidas em (2.6).

O elemento genérico de ordem j da função *score* é então

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) z_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}. \tag{2.8}$$

A matriz de covariância da função *score*, $\mathcal{I}(\boldsymbol{\beta}) = E[-\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}]$ é conhecida por *matriz de informação de Fisher*. Para obter a matriz de informação de Fisher temos de considerar o valor esperado das segundas derivadas de $\ell_i(\boldsymbol{\beta})$.

Tem-se, para famílias regulares, que

$$\begin{aligned}
-E\Big(\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k}\Big) &= E\Big(\frac{\partial \ell_i}{\partial \beta_j} \frac{\partial \ell_i}{\partial \beta_k}\Big) \\
&= E\Big[\Big(\frac{(Y_i - \mu_i) z_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}\Big)\Big(\frac{(Y_i - \mu_i) z_{ik}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}\Big)\Big] \\
&= E\Big[\frac{(Y_i - \mu_i)^2 z_{ij} z_{ik}}{(var(Y_i))^2} \Big(\frac{\partial \mu_i}{\partial \eta_i}\Big)^2\Big] \\
&= \frac{z_{ij} z_{ik}}{var(Y_i)} \Big(\frac{\partial \mu_i}{\partial \eta_i}\Big)^2
\end{aligned}$$

e, portanto, o elemento genérico de ordem (j, k) da matriz de informação de Fisher é

$$-\sum_{i=1}^{n} E\Big(\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k}\Big) = \sum_{i=1}^{n} \frac{z_{ij} z_{ik}}{var(Y_i)} \Big(\frac{\partial \mu_i}{\partial \eta_i}\Big)^2. \tag{2.9}$$

Na forma matricial temos

$$\mathcal{I}(\boldsymbol{\beta}) = Z^T W Z \ , \tag{2.10}$$

onde W é a matriz diagonal de ordem n cujo i-ésimo elemento é

$$\varpi_i = \frac{\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2}{var(Y_i)} = \frac{\omega_i \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2}{\phi V(\mu_i)}. \tag{2.11}$$

A última igualdade em (2.11) aparece devido à relação entre a função de variância $V(\mu) = b''(\theta)$ e a variância de Y, nomeadamente $var(Y) = \frac{\phi}{\omega} V(\mu)$, referida em (1.9), na secção 1.2.1.

2.1.2 Função de ligação canónica e estatísticas suficientes

Quando a função de ligação é a canónica, isto é, quando

$$\theta_i = \eta_i = \mathbf{z}_i^T \boldsymbol{\beta},$$

a verosimilhança pode escrever-se na forma

$$\begin{aligned} L(\boldsymbol{\beta}) &= \exp\Big\{ \sum_{i=1}^{n} \frac{\omega_i (\mathbf{z}_i^T \boldsymbol{\beta} y_i - b(\theta_i))}{\phi} + \sum_{i=1}^{n} c(y_i, \phi, \omega_i) \Big\} \\ &= \exp\Big\{ \sum_{j=1}^{p} \Big(\sum_{i=1}^{n} \frac{\omega_i y_i z_{ij}}{\phi} \Big) \beta_j - \sum_{i=1}^{n} \frac{\omega_i b(\theta_i)}{\phi} + \sum_{i=1}^{n} c(y_i, \phi, \omega_i) \Big\} \end{aligned}$$

o que mostra, pelo teorema da factorização, que se ϕ for conhecido, a estatística suficiente mínima para o modelo (para o vector parâmetro $\boldsymbol{\beta}$) tem dimensão p e é dada pelo vector $\sum_{i=1}^{n} \omega_i y_i \mathbf{z}_i = (\sum_{i=1}^{n} \omega_i y_i z_{i1}, \ldots, \sum_{i=1}^{n} \omega_i y_i z_{ip})^T$. Se ϕ for desconhecido e a família ainda se puder escrever na forma da família exponencial, então aquele vector é uma componente da estatística suficiente mínima.

Vejamos alguns exemplos.

Exemplo 2.1 Modelo Poisson com ligação canónica

Suponhamos que temos respostas $Y_i \sim P(\lambda_i), i = 1, ..., n$, *i.e.*

$$f(y_i|\lambda_i) = \exp\{y_i \ln(\lambda_i) - \lambda_i - \ln(y_i!)\}, \qquad y_i = 0, 1, \ldots$$

Como $\theta_i = \ln \lambda_i$, a função de ligação canónica é a função logarítmica, *i.e.*, $\ln \lambda_i = \mathbf{z}_i^T \boldsymbol{\beta}$.

A verosimilhança em função de $\boldsymbol{\beta}$ é

$$L(\boldsymbol{\beta}) = \exp\Big\{ \sum_{i=1}^{n} y_i \mathbf{z}_i^T \boldsymbol{\beta} - \sum_{i=1}^{n} e^{\mathbf{z}_i^T \boldsymbol{\beta}} - \sum_{i=1}^{n} \ln(y_i!) \Big\}$$

e, portanto, a estatística suficiente mínima é

$$\sum_{i=1}^{n} y_i \mathbf{z}_i = (\sum_{i=1}^{n} y_i z_{i1}, ..., \sum_{i=1}^{n} y_i z_{ip})^T.$$

Exemplo 2.2 Modelo Gama com função de ligação canónica

Suponhamos que $Y_i \sim Ga(\nu, \frac{\nu}{\mu_i})$.

Tem-se, como já se viu no exemplo 1.3 da secção 1.2.2, que

$$f(y_i|\nu, \mu_i) = \exp\{\nu(\theta_i y_i + \ln(-\theta_i)) + (\nu - 1) \ln y_i + \nu \ln \nu - \ln \Gamma(y_i)\},$$

com $\theta_i = -\frac{1}{\mu_i}$ e $\phi = \frac{1}{\nu}$.

Assim a função de ligação canónica é dada por, $-\frac{1}{\mu_i} = \mathbf{z}_i^T \boldsymbol{\beta}$.

A verosimilhança, como função de $\boldsymbol{\beta}$ é então

$$\begin{aligned} L(\boldsymbol{\beta}) = \exp\Big\{ & \nu \textstyle\sum_{i=1}^{n} y_i \mathbf{z}_i^T \boldsymbol{\beta} + \nu \sum_{i=1}^{n} \ln(\mathbf{z}_i^T \boldsymbol{\beta}) + \\ & (\nu - 1) \textstyle\sum_{i=1}^{n} \ln y_i + n\nu \ln \nu - \sum_{i=1}^{n} \ln(\Gamma(y_i)) \Big\}. \end{aligned}$$

Se ν for conhecido a estatística suficiente mínima é então o vector

$$\sum_{i=1}^{n} y_i \mathbf{z}_i = \Big(\sum_{i=1}^{n} y_i z_{i1}, ..., \sum_{i=1}^{n} y_i z_{ip} \Big)^T.$$

Se ν for desconhecido a estatística suficiente mínima é

$$\Big(\Big(\sum_{i=1}^{n} y_i \mathbf{z}_i\Big)^T, \sum_{i=1}^{n} \ln y_i\Big).$$

Outra característica interessante dos MLG com função de ligação canónica, diz respeito à facilidade computacional associada às estimativas de máxima verosimilhança e à relação existente entre a matriz de informação de Fisher

$$\mathcal{I}(\boldsymbol{\beta}) = E\Big[-\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\Big]$$

e a matriz Hessiana

$$\mathcal{H}(\boldsymbol{\beta}) = \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}.$$

Com efeito, se a função de ligação for a canónica, $\theta_i = \eta_i$ e portanto

$$\begin{aligned} \frac{\partial \ell_i}{\partial \beta_j} &= \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \eta_i}{\partial \beta_j} \\ &= \frac{\omega_i(y_i - \mu_i)}{\phi} z_{ij} \end{aligned}$$

e as equações de verosimilhança resumem-se a

$$\sum_{i=1}^{n} \frac{\omega_i(y_i - \mu_i)}{\phi} z_{ij} = 0, \qquad j = 1, ..., p,$$

o que implica

$$\sum_{i=1}^{n} \omega_i y_i z_{ij} = \sum_{i=1}^{n} \omega_i \hat{\mu}_i z_{ij},$$

onde $\hat{\mu}_i$ representam as estimativas de máxima verosimilhança de μ_i, $i = 1, .., n$.

Quando os $\omega_i = 1$ para todo o i aquelas equações resumem-se a

$$\sum_{i=1}^{n} y_i z_{ij} = \sum_{i=1}^{n} \hat{\mu}_i z_{ij},$$

ou na forma matricial a

$$Z^T\mathbf{y} = Z^T\widehat{\boldsymbol{\mu}}.$$

Se, como é habitual, a primeira coluna da matriz Z for um vector unitário, *i.e.*, $z_{i1} = 1, i = 1, ..., n$, então a equação anterior implica, por exemplo, que a soma dos valores observados y_i é igual à soma das estimativas dos valores esperados μ_i.

Relativamente à matriz de informação de Fisher temos, pelo facto das segundas derivadas da *log-verosimilhança* ℓ_i

$$\frac{\partial^2 \ell_i}{\partial\beta_i\partial\beta_k} = -\frac{\omega_i z_{ij}}{\phi}\frac{\partial\mu_i}{\partial\beta_k}$$

não dependerem das observações y_i, que

$$-E\Big[\frac{\partial^2 \ell_i}{\partial\beta_i\partial\beta_k}\Big] = \frac{\omega_i z_{ij}}{\phi}\frac{\partial\mu_i}{\partial\beta_k},$$

ou seja a matriz de informação de Fisher $\mathcal{I}(\boldsymbol{\beta})$ coincide com $-\mathcal{H}(\boldsymbol{\beta})$, o simétrico da matriz Hessiana.

Note-se contudo, que embora as funções de ligação canónica conduzam a propriedades estatísticas desejáveis para o modelo, tais como, suficiência, facilidade de cálculo, unicidade das estimativas de máxima verosimilhança e, por vezes, interpretação simples, não há razão para, à partida, escolher a função de ligação canónica e nem sempre é com ela que se obtêm os melhores resultados. Por exemplo, no modelo gama, a utilização da função de ligação canónica obriga a impor restrições aos parâmetros. Como já se disse anteriormente, o aspecto importante a ter em consideração na escolha da ligação é a adaptabilidade e adequabilidade do modelo.

2.2 Estimação dos Parâmetros do Modelo

Os estimadores de máxima verosimilhança (EMV) de $\boldsymbol{\beta}$ são obtidos como solução das equações de verosimilhança (2.7). A solução não corresponde necessariamente a um máximo global da função $\ell(\boldsymbol{\beta})$. Contudo, em muitos modelos a função *log-verosimilhança* $\ell(\boldsymbol{\beta})$ é côncava, de modo que o máximo local e global coincidem. Para funções estritamente côncavas os estimadores de máxima verosimilhança são mesmo únicos, quando existem. O problema da existência e unicidade dos estimadores de máxima verosimilhança será referido na secção 2.2.4. Partindo do princípio que existe solução e que ela é única, subsiste ainda um problema com o cálculo das estimativas de máxima verosimilhança. É que as equações (2.7) não têm, em geral, solução analítica e, portanto, a sua resolução implica o recurso a métodos numéricos.

Uma das razões que fez com que os MLG tivessem sucesso, foi o facto da possibilidade de usar um único algoritmo (sugerido por Nelder e Wedderburn, 1972), para resolver (2.7), havendo apenas a necessidade de proceder a pequenos ajustamentos de acordo com a distribuição de probabilidade e a função de ligação em consideração. Além disso o algoritmo proposto opera através de uma sequência de problemas de mínimos quadrados ponderados para os quais existem técnicas numéricas bem testadas.

2.2.1 Método iterativo de mínimos quadrados ponderados

O esquema iterativo para a resolução numérica das equações de verosimilhança que se vai apresentar, é baseado no método de *scores*

de Fisher.

Seja $\widehat{\boldsymbol{\beta}}^{(0)}$ uma estimativa inicial para $\boldsymbol{\beta}$. O processo de *scores* de Fisher procede com o cálculo das sucessivas iteradas através da relação:

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \widehat{\boldsymbol{\beta}}^{(k)} + \Big[\mathcal{I}(\widehat{\boldsymbol{\beta}}^{(k)})\Big]^{-1} s(\widehat{\boldsymbol{\beta}}^{(k)}),$$

onde $\mathcal{I}(\cdot)^{-1}$, a inversa (que se supõe existir) da matriz de informação de Fisher dada em (2.10) e $s(\cdot)$, o vector de *scores*, são calculados para $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{(k)}$.

A diferença existente entre este algoritmo e o algoritmo de Newton-Raphson para resolver sistemas de equações não lineares, reside na utilização da matriz de informação de Fisher em vez da matriz Hessiana. A vantagem desta substituição deve-se ao facto de, em geral, ser mais fácil calcular a matriz de informação $\mathcal{I}$, para além de ser sempre uma matriz semi-definida positiva.

A expressão anterior pode ser ainda escrita na forma

$$\Big[\mathcal{I}(\widehat{\boldsymbol{\beta}}^{(k)})\Big]\widehat{\boldsymbol{\beta}}^{(k+1)} = \Big[\mathcal{I}(\widehat{\boldsymbol{\beta}}^{(k)})\Big]\widehat{\boldsymbol{\beta}}^{(k)} + s(\widehat{\boldsymbol{\beta}}^{(k)}). \tag{2.12}$$

Atendendo a (2.8) e (2.9), o lado direito da equação (2.12) é um vector com elemento genérico de ordem l dado por:

$$\sum_{j=1}^{p}\Big[\sum_{i=1}^{n}\frac{z_{ij}z_{il}}{var(Y_i)}\Big(\frac{\partial\mu_i}{\partial\eta_i}\Big)^2\Big]\beta_j^{(k)} + \sum_{i=1}^{n}\frac{(y_i-\mu_i)z_{il}}{var(Y_i)}\frac{\partial\mu_i}{\partial\eta_i}$$

e, portanto, na forma matricial tem-se

$$\mathcal{I}(\widehat{\boldsymbol{\beta}}^{(k)})\widehat{\boldsymbol{\beta}}^{(k+1)} = Z^T W^{(k)}\mathbf{u}^{(k)},$$

onde $\mathbf{u}^{(k)}$ é um vector com elemento genérico

$$\begin{aligned} u_i^{(k)} &= \sum_{j=1}^{p} z_{ij}\beta_j^{(k)} + (y_i-\mu_i^{(k)})\frac{\partial\eta_i^{(k)}}{\partial\mu_i^{(k)}} \\ &= \eta_i^{(k)} + (y_i-\mu_i^{(k)})\frac{\partial\eta_i^{(k)}}{\partial\mu_i^{(k)}} \end{aligned} \tag{2.13}$$

e a matriz $W^{(k)}$ representa a matriz W definida em (2.11) e calculada em $\widehat{\boldsymbol{\mu}}^{(k)}$.

Assim, atendendo a (2.10), tem-se a expressão final para a estimativa de $\boldsymbol{\beta}$ na $(k+1)$-ésima iteração

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \left(Z^T W^{(k)} Z\right)^{-1} Z^T W^{(k)} \mathbf{u}^{(k)}. \tag{2.14}$$

A análise desta expressão leva-nos a perceber o facto de anteriormente termos afirmado que o "algoritmo proposto opera através de uma sequência de problemas de mínimos quadrados ponderados". Com efeito, a equação (2.14) é idêntica à que se obteria para os estimadores de mínimos quadrados ponderados se se fizesse, em cada passo, a regressão linear de respostas $\mathbf{u}^{(k)}$ em Z, sendo $W^{(k)}$ uma matriz de pesos. Por isso este algoritmo é referido como "algoritmo iterativo de mínimos quadrados ponderados".

Note-se, ainda por análise da expressão (2.14), que, apesar de o elemento genérico de W conter ϕ, ele não entra no cálculo de $\widehat{\boldsymbol{\beta}}^{(k+1)}$ e portanto pode-se fazer, sem perda de generalidade, $\phi = 1$, quando se está a calcular as estimativas de $\boldsymbol{\beta}$. Assim, é irrelevante, para o cálculo de $\widehat{\boldsymbol{\beta}}$ o conhecimento ou não do parâmetro de dispersão.

Resumindo: O cálculo das estimativas de máxima verosimilhança de $\boldsymbol{\beta}$ processa-se, iterativamente, em duas etapas:

i) Dado $\widehat{\boldsymbol{\beta}}^{(k)}$ (com k a iniciar-se em 0), calcula-se $\mathbf{u}^{(k)}$ usando a expressão (2.13) e $W^{(k)}$ usando (2.11).

ii) A nova iterada $\widehat{\boldsymbol{\beta}}^{(k+1)}$ é calculada usando (2.14).

As iterações param quando é atingido um critério adequado, por exemplo, quando

$$\frac{\|\widehat{\boldsymbol{\beta}}^{(k+1)} - \widehat{\boldsymbol{\beta}}^{(k)}\|}{\|\widehat{\boldsymbol{\beta}}^{(k)}\|} \leq \epsilon,$$

para algum valor de $\epsilon > 0$ previamente definido.

Em geral a convergência atinge-se após algumas iteradas. Se o processo iterativo não parecer convergir, isto pode ser devido a uma má estimativa inicial ou, muitas vezes, à não existência de estimador de máxima verosimilhança dentro da região de valores admissíveis para o vector parâmetro $\boldsymbol{\beta}$.

Para calcular a iterada de ordem zero, $\widehat{\boldsymbol{\beta}}^{(0)}$, que dá início ao processo iterativo, pode calcular-se a estimativa de mínimos quadrados não ponderados para os dados $(g(y_i), \mathbf{z}_i), i = 1, ..., n)$, onde $g(\cdot)$ é a função de ligação. Se para algum y_i, $g(y_i)$ não estiver definido, como é o caso quando g é a função logarítmica e $y_i = 0$, pode modificar-se ligeiramente a observação y_i de modo a que $g(y_i)$ passe a estar definido.

Para que o algoritmo se processe sem problemas é preciso que a matriz $\mathcal{I}(\widehat{\boldsymbol{\beta}}^{(k)})$ tenha inversa em cada iterada. Dado que se assumiu previamente que $Z^T Z$ é de característica completa, a inversa existe desde que os elementos da matriz $W^{(k)}$ sejam todos positivos ou quase todos positivos.

2.2.2 Estimação do parâmetro de dispersão

O parâmetro de dispersão também pode ser estimado usando o método da máxima verosimilhança. Há, no entanto, um método mais simples que dá, em geral, bons resultados. Este método baseia-se na distribuição de amostragem, para grandes valores de n, da estatística de Pearson generalizada.

Suponhamos então que se aplicou o algoritmo iterativo de mínimos quadrados ponderados e se obteve uma estimativa $\widehat{\boldsymbol{\beta}}$ para $\boldsymbol{\beta}$. Devido à propriedade de invariância dos estimadores de máxima

verosimilhança, as estimativas de máxima verosimilhança para os parâmetros μ_i são dadas por

$$\widehat{\mu}_i = h(\mathbf{z}_i^T \widehat{\boldsymbol{\beta}}),$$

onde a função $h(\cdot)$ é a inversa da função de ligação.

Por outro lado, como se tem que

$$var(Y_i) = b''(\theta_i)\frac{\phi}{\omega_i} = \frac{V(\mu_i)\phi}{\omega_i}, \quad i = 1, ..., n,$$

então

$$E\Big[\frac{\omega_i(Y_i - \mu_i)^2}{V(\mu_i)}\Big] = \phi \qquad i = 1, ..., n.$$

Pela lei fraca dos grandes números, se

$$\frac{1}{n^2}\sum_{i=1}^{n}\frac{\omega_i^2 E(Y_i - \mu_i)^4}{[V(\mu_i]^2} \longrightarrow 0,$$

quando $n \to \infty$, então

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\omega_i(Y_i - \mu_i)^2}{V(\mu_i)} \xrightarrow{P} \phi.$$

Segue-se então que se $V(\cdot)$ é uma função contínua e $\widehat{\mu}_i \xrightarrow{P} \mu_i$ para todo o i, então

$$\frac{1}{n-p}\sum_{i=1}^{n}\frac{\omega_i(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)} \xrightarrow{P} \phi.$$

Assim podemos estimar ϕ por:

$$\widehat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{\omega_i(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}, \tag{2.15}$$

sendo $\widehat{\phi}$ um estimador consistente de ϕ. Como, por outro lado se tem que, para grandes valores de n,

$$\frac{1}{\phi}\sum_{i=1}^{n}\frac{\omega_i(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)} = \sum_{i=1}^{n}\frac{\omega_i(Y_i - \widehat{\mu}_i)^2}{var(Y_i)}$$

tem uma distribuição aproximada de um χ^2 com $n-p$ graus de liberdade, também se conclui que $\widehat{\phi}$ é assintoticamente centrado.

A Estatística do lado direito da equação (2.15) é conhecida como *estatística de Pearson generalizada*, a qual também é útil para julgar da qualidade de ajustamento de um modelo.

Esta estimativa de ϕ é mais simples e produz, em geral, maior estabilidade numérica que a de máxima verosimilhança. Note-se que, no caso do modelo de regressão linear normal , $\omega_i = 1$ e $V(\widehat{\mu}_i) = 1$ e portanto $\widehat{\phi}$ coincide com a estimativa habitual de σ^2.

2.2.3 Propriedades assintóticas dos estimadores de máxima verosimilhança

Na secção anterior vimos como, formulado um modelo linear generalizado, podemos proceder à estimação por máxima verosimilhança do vector parâmetro $\boldsymbol{\beta}$ dos coeficientes de regressão. Para fazer inferências sobre estes parâmetros, nomeadamente obter intervalos de confiança e fazer testes de hipóteses, há necessidade de conhecer a distribuição de amostragem de $\widehat{\boldsymbol{\beta}}$. Em geral, não é possível nos MLG, obter as distribuições de amostragem exactas para os estimadores de máxima verosimilhança dos β's. Iremos apelar então, para resultados conhecidos da teoria assintótica, que se verificam quando os modelos em estudo satisfazem certas condições de regularidade ; essas condições são, com efeito, satisfeitas para os MLG. Não iremos entrar em detalhes de natureza teórica, deixando ao cuidado do leitor interessado a leitura, por exemplo, de Fahrmeir and Kaufmann (1985), onde são estabelecidas, com o rigor adequado, condições gerais que garantem a consistência e a normalidade assintótica do estimador $\widehat{\boldsymbol{\beta}}$.

Como vimos, o estimador de máxima verosimilhança, $\widehat{\boldsymbol{\beta}}$, de $\boldsymbol{\beta}$ obtém-se como solução de

$$s(\widehat{\boldsymbol{\beta}}) = 0,$$

onde $s(\boldsymbol{\beta})$ é o vector *score*. Também é sabido que, em condições de regularidade , este vector aleatório é tal que

$$E(S(\boldsymbol{\beta})) = \mathbf{0}, \qquad cov(S(\boldsymbol{\beta})) = E(S(\boldsymbol{\beta})S(\boldsymbol{\beta})^T) = \mathcal{I}(\boldsymbol{\beta}),$$

onde $\mathcal{I}(\boldsymbol{\beta})$ é a matriz de informação de Fisher já definida. Por outro lado, pelo teorema do limite central, temos a garantia de que, pelo menos assintoticamente, $S(\boldsymbol{\beta})$ tem uma distribuição normal multivariada, *i.e.*,

$$S(\boldsymbol{\beta}) \overset{a}{\sim} N_p(\mathbf{0}, \mathcal{I}(\boldsymbol{\beta}))$$

e que, portanto, para grandes amostras e supondo que o modelo com o vector parâmetro $\boldsymbol{\beta}$ especificado é verdadeiro, a estatística $S(\boldsymbol{\beta})^T\mathcal{I}^{-1}(\boldsymbol{\beta})S(\boldsymbol{\beta})$ tem uma distribuição assintótica de um qui-quadrado, *i.e.*,

$$S(\boldsymbol{\beta})^T\mathcal{I}^{-1}(\boldsymbol{\beta})S(\boldsymbol{\beta}) \overset{a}{\sim} \chi^2_p.$$

Se desenvolvermos $S(\boldsymbol{\beta})$ em série de Taylor em torno de $\widehat{\boldsymbol{\beta}}$ e retivermos apenas os termos de $1^{\underline{a}}$ ordem, obtemos a relação:

$$S(\boldsymbol{\beta}) \approx S(\widehat{\boldsymbol{\beta}}) + \frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \tag{2.16}$$

Atendendo a que $S(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$ e substituindo a matriz de informação observada $-\mathcal{H}(\widehat{\boldsymbol{\beta}}) = \frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}$ pela matriz de informação de Fisher, isto é, fazendo $-H(\widehat{\boldsymbol{\beta}}) = \mathcal{I}(\boldsymbol{\beta})$ em (2.16) (o que admitimos ser aproximadamente válido para grandes amostras), obtemos

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \approx \mathcal{I}^{-1}(\boldsymbol{\beta})S(\boldsymbol{\beta}) \tag{2.17}$$

A expressão (2.17) e os resultados anteriores relativos ao vector *score* são fundamentais para a dedução das propriedades assintóticas dos estimadores de máxima verosimilhança de $\boldsymbol{\beta}$. Com efeito, de (2.17) pode concluir-se que,

- $E(\widehat{\boldsymbol{\beta}}) \approx \boldsymbol{\beta}$, ou melhor, $\widehat{\boldsymbol{\beta}}$ é um estimador de $\boldsymbol{\beta}$ assintoticamente centrado,
- $cov(\widehat{\boldsymbol{\beta}}) \approx E[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T] = \mathcal{I}^{-1}(\boldsymbol{\beta})$, isto é, a matriz de covariância de $\widehat{\boldsymbol{\beta}}$ é aproximadamente igual ao inverso da matriz de informação de Fisher,
- a distribuição assintótica de $\widehat{\boldsymbol{\beta}}$ é normal p-variada, com vector médio $\boldsymbol{\beta}$ e matriz de covariância $\mathcal{I}^{-1}(\boldsymbol{\beta})$, isto é,

$$\widehat{\boldsymbol{\beta}} \overset{a}{\sim} N_p(\boldsymbol{\beta}, \mathcal{I}^{-1}(\boldsymbol{\beta})),$$

- A estatística de Wald $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathcal{I}(\boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ tem uma distribuição assintótica de um χ^2 com p graus de liberdade.

Note-se que os resultados apresentados sobre as propriedades distribucionais de $\widehat{\boldsymbol{\beta}}$ são exactos para o modelo normal.

Os resultados apresentados são úteis para fazer inferências sobre $\boldsymbol{\beta}$. Com efeito, a distribuição assintótica normal multivariada serve de base para a construção de testes de hipóteses e de intervalos de confiança. Por exemplo, para obter intervalos de confiança para os parâmetros componentes do vector $\boldsymbol{\beta}$, podemos usar, pelas propriedades da distribuição normal multivariada (veja-se, e.g., Johnson and Wichern, 1998), o facto de $\hat{\beta}_j$ ter ainda uma distribuição assintótica $N(\beta_j, \sigma_{jj})$, onde σ_{jj} é o j-ésimo elemento da diagonal da matriz $\mathcal{I}^{-1}(\boldsymbol{\beta})$. O conhecimento da matriz de covariância permite-nos ainda calcular correlações entre os diferentes $\hat{\beta}_j$'s. Por outro

lado, a estatística de Wald é, como veremos, uma das estatísticas usadas para fazer testes de hipóteses sobre o vector parâmetro $\boldsymbol{\beta}$. Claro que, como $\boldsymbol{\beta}$ é desconhecido e a matriz de informação de Fisher depende de $\boldsymbol{\beta}$, ela é de facto desconhecida. No entanto, na prática, costuma substituir-se no cáculo de $\mathcal{I}^{-1}(\boldsymbol{\beta})$, o vector $\boldsymbol{\beta}$ pela sua estimativa $\widehat{\boldsymbol{\beta}}$.

A existência de um parâmetro de dispersão ϕ desconhecido, pode afectar a estrutura assintótica, embora não afecte, como se viu anteriormente, o estimador $\widehat{\boldsymbol{\beta}}$. Na prática, porém, quando existe um ϕ desconhecido, ele é substituído por uma sua estimativa consistente $\widehat{\phi}$.

A distribuição normal p-dimensional será uma boa aproximação para a distribuição de $\widehat{\boldsymbol{\beta}}$ se o logaritmo da verosimilhança for uma função "razoavelmente" quadrática. Este é, em geral, o caso para grandes amostras, mas em pequenas amostras a *log-verosimilhança* pode-se afastar de uma função quadrática. Pode no entanto acontecer que haja uma reparametrização, digamos $h^*(\boldsymbol{\beta}) = \boldsymbol{\gamma}$, que conduza a uma *log-verosimilhança* aproximadamente quadrática. Nesse caso, pode obter-se testes de hipóteses e regiões de confiança mais precisas baseadas na distribuição assintótica de $\widehat{\boldsymbol{\gamma}}$.

2.2.4 Existência e unicidade dos EMV

Um problema importante, e que tem sido tratado por vários autores apenas para MLG particulares, é o problema da existência e unicidade dos estimadores de máxima verosimilhança, já que à partida não há garantia de que a função de verosimilhança tenha um máximo único, ou mesmo que tenha um máximo; outro aspecto que também é importante, pelo menos do ponto de vista das apli-

cações, é saber se a verosimilhança tem um máximo na fronteira do espaço admissível para o vector parâmetro, já que a existência de tal máximo pode levar a problemas de natureza computacional.

Não há, no entanto, uma teoria geral sobre o problema da existência e unicidade de estimadores de máxima verosimilhança para os modelos lineares generalizados, pois que nem todos os modelos têm propriedades comuns no que diz respeito a esta questão. Há, contudo, resultados obtidos para casos particulares; Haberman (1974) dedicou o seu estudo a modelos log-lineares e binomiais; Silvapulle (1981) apresentou condições necessárias e suficientes para a existência de estimadores de máxima verosimilhança para os modelos binomiais com funções de ligação gerais, dedicando particular atenção aos modelos logístico e *probit*; Wederburn (1976) estudou condições de existência e unicidade dos estimadores de máxima verosimilhança nos modelos normal, binomial, Poisson e gama. As conclusões desse estudo encontram-se resumidas na tabela 2.1[2]. Mais referências podem ser encontradas em Fahrmeir and Tutz (1994).

2.3 Testes de Hipóteses

A maior parte dos problemas de inferência relacionados com testes de hipóteses sobre o vector parâmetro $\boldsymbol{\beta}$, podem ser formulados em termos de hipóteses lineares da forma:

$$H_0 : C\boldsymbol{\beta} = \boldsymbol{\xi} \quad versus \quad H_1 : C\boldsymbol{\beta} \neq \boldsymbol{\xi}, \tag{2.18}$$

onde C é uma matriz $q \times p$, com $q \leq p$, de característica completa q e $\boldsymbol{\xi}$ é um vector de dimensão q previamente especificado.

Casos especiais importantes de (2.18) são:

[2]Esta tabela é reproduzida do trabalho de Wederburn (1976).

Tabela 2.1: Propriedades das estimativa de máxima verosimilhança de $\boldsymbol{\beta}$ para várias distribuições e funções de ligação; F significa existência de estimativa finita; I significa existência de estimativa no interior do espaço paramétrico; U significa unicidade.

a) Modelos para os quais $\mu \geq 0$ ou $\mu > 0$ se a função de ligação não estiver definida para $\mu = 0$

função de ligação	normal	Poisson	gama
$\mu^\lambda (\lambda < -1)$	I	F,I	F,I
$\mu^\lambda (-1 \leq \lambda < 0)$	I	F,I	F,I,U
$\ln \lambda$	I	F,I,U	F,I,U
$\mu^\lambda (0 < \lambda < 1)$	F	F,I,U	F,I
μ	F,U	F,I,U	F,I
$\mu^\lambda (\lambda > 1)$	F	F,I	F,I

b) Modelos para a distribuição binomial

função de ligação	
μ	I,U
$\sin^{-1} \sqrt{\mu}$	I,U
$\Phi^{-1}(\mu)$	F,I,U
$\ln \frac{\mu}{1-\mu}$	F,I,U
$\ln\{-\ln(1-\mu)\}$	F,I,U

c) Modelos para a distribuição normal sem restrições a μ

função de ligação	
μ	F,I,U
e^μ	F,I

- Hipótese da nulidade de uma componente do vector parâmetro, nomeadamente

$$H_0 : \beta_j = 0 \quad versus \quad H_1 : \beta_j \neq 0,$$

para algum j, sendo neste caso $q = 1$, $C = (0, ..., 0, 1, 0, ..., 0)$ e ocupando o 1 a j-ésima posição e $\xi = 0$.

- Hipótese da nulidade de um subvector do vector parâmetro,

$$H_0 : \boldsymbol{\beta}_r = \mathbf{0} \quad versus \quad H_1 : \boldsymbol{\beta}_r \neq \mathbf{0},$$

para algum subvector de r componentes de $\boldsymbol{\beta}$. Se tivermos, por exemplo $H_0 : (\beta_1, ..., \beta_r)^T = (0, ..., 0)^T$, então $q = r$ e

$$C = (\, I_r \quad O_{r\times(p-r)} \,) \qquad \boldsymbol{\xi} = \mathbf{0}_r,$$

onde I_r é a matriz identidade de dimensão r, $O_{r\times(p-r)}$ é uma matriz de zeros de dimensão $r \times (p - r)$ e $\mathbf{0}_r$ é o vector nulo de dimensão r.

Estas hipóteses correspondem a testar submodelos do modelo original, importante na selecção de covariáveis, como se irá ver no capítulo 3; a 1ª corresponde a testar um submodelo com todas as covariáveis do modelo original à excepção da covariável z_j relativa ao parâmetro de regressão β_j e a segunda corresponde a testar um modelo sem as r covariáveis relativas aos parâmetros supostos nulos sob a hipótese H_0. Uma das situações em que isso acontece surge, por exemplo, quando uma variável é policotómica tomando, digamos, $r + 1$ valores distintos. Nesse caso é, como já se disse, aconselhável construir r variáveis dicotómicas para a representar, havendo nesse caso r parâmetros β's associados a ela. Assim, para

averiguar se essa variável deve ou não ser incluída no modelo, interessa testar globalmente se os r parâmetros são significativamente diferentes de zero.

De um modo geral pode ter interesse testar certas relações estruturais entre as componentes do vector $\boldsymbol{\beta}$. Se essas relações se puderem escrever na forma $C\boldsymbol{\beta} = \boldsymbol{\xi}$, então os testes que iremos formular serão adequados para esses estudos.

Existem essencialmente três estatísticas para testar hipóteses do tipo (2.18) relativas às componentes do vector $\boldsymbol{\beta}$, as quais são deduzidas das distribuições assintóticas dos estimadores de máxima verosimilhança e de funções adequadas desses estimadores.

I A *Estatística de Wald*, baseada na normalidade assintótica do estimador de máxima verosimilhança $\widehat{\boldsymbol{\beta}}$.

II A *Estatística de Wilks* ou *Estatística de razão de verosimilhanças*, baseada na distribuição assintótica da razão do máximo das verosimilhanças sob as hipóteses H_0 e $H_0 \cup H_1$.

III A *Estatística de Rao* ou *Estatística score*, baseada nas propriedades assintóticas da função *score*.

Iremos, seguidamente, descrever cada uma destas estatísticas de teste e como podem ser utilizadas para testar as hipóteses de interesse apresentadas.

2.3.1 Teste de Wald

Suponhamos que a hipótese nula estabelece que $C\boldsymbol{\beta} = \boldsymbol{\xi}$, onde C é uma matriz $q \times p$ de característica completa q. Seja $\widehat{\boldsymbol{\beta}}$ o estimador de máxima verosimilhança de $\boldsymbol{\beta}$, o qual tem uma distribuição

assintótica $N_p(\boldsymbol{\beta}, \mathcal{I}^{-1}(\widehat{\boldsymbol{\beta}}))$[3]. Dado que o vector $C\widehat{\boldsymbol{\beta}}$ é uma transformação linear de $\widehat{\boldsymbol{\beta}}$ então, pelas propriedades da distribuição normal multivariada,

$$C\widehat{\boldsymbol{\beta}} \overset{a}{\sim} N_q\Big(C\boldsymbol{\beta}, C\mathcal{I}^{-1}(\widehat{\boldsymbol{\beta}})C^T\Big)$$

e, consequentemente, sob a hipótese, nula a estatística

$$\mathcal{W} = (C\widehat{\boldsymbol{\beta}} - \boldsymbol{\xi})^T[CI^{-1}(\widehat{\boldsymbol{\beta}}))C^T]^{-1}(C\widehat{\boldsymbol{\beta}} - \boldsymbol{\xi}), \tag{2.19}$$

tem uma distribuição assintótica de um χ^2 com q graus de liberdade.

À estatística $\mathcal{W}$ em (2.19) damos o nome de *Estatística de Wald*.

Assim, a hipótese nula é rejeitada, a um nível de significância α, se o valor observado da *estatística de Wald* for superior ao quantil de probabilidade $1-\alpha$ de um χ^2_q.

Exemplo 2.3 Suponhamos que queremos testar a hipótese

$$H_0 : \beta_j = 0 \quad versus \quad H_1 : \beta_j \neq 0,$$

para algum j.

Então, usando (2.19) e designando como anteriormente por σ_{jj} o j-ésimo elemento da diagonal de $\mathcal{I}^{-1}(\widehat{\boldsymbol{\beta}})$, a *estatística de Wald* resume-se a,

$$\mathcal{W} = (\hat{\beta}_j - \beta_j)^T[\sigma_{jj}]^{-1}(\hat{\beta}_j - \beta_j)$$

e, portanto, sob H_0,

$$\mathcal{W} = \frac{\hat{\beta}_j{}^2}{\sigma_{jj}} \overset{a}{\sim} \chi^2_1,$$

resultado já obtido anteriormente com base na distribuição assintótica normal de $\hat{\beta}_j$.

[3]Note-se que aqui já substituímos o vector $\boldsymbol{\beta}$ pela sua estimativa $\widehat{\boldsymbol{\beta}}$, admitindo que para grandes amostras $\mathcal{I}(\boldsymbol{\beta}) \approx \mathcal{I}(\widehat{\boldsymbol{\beta}})$.

A *estatística de Wald* é, em geral, a mais utilizada para testar hipóteses nulas sobre componentes individuais, embora também se use para testar hipóteses nulas do tipo $\boldsymbol{\beta}_r = \mathbf{0}$ quando o subvector $\boldsymbol{\beta}_r$ representa o vector correspondente a uma recodificação de uma variável policotómica[4].

2.3.2 Teste de razão de verosimilhanças

A *estatística de Wilks* ou *estatística de razão de verosimilhanças* é definida por

$$\begin{aligned} \Lambda &= -2\ln\frac{\max_{H_0} L(\boldsymbol{\beta})}{\max_{H_0\cup H_1} L(\boldsymbol{\beta})} \\ &= -2\{\ell(\widetilde{\boldsymbol{\beta}}) - \ell(\widehat{\boldsymbol{\beta}})\} \end{aligned} \tag{2.20}$$

onde $\widetilde{\boldsymbol{\beta}}$, o estimador de máxima verosimilhança restrito, é o valor de $\boldsymbol{\beta}$ que maximiza a verosimilhança sujeito às restrições impostas pela hipótese $C\boldsymbol{\beta} = \boldsymbol{\xi}$.

O teorema de Wilks (e.g., Cox and Hinkley, 1974) estabelece que, sob certas condições de regularidade , a estatística Λ tem, sob H_0, uma distribuição assintótica de um χ^2 sendo o número de graus de liberdade igual à diferença entre o número de parâmetros a estimar sob $H_0 \cup H_1$ (neste caso p) e o número de parâmetros a estimar sob H_0 (neste caso $p - q$).

Assim, sob H_0,

$$\Lambda = -2\{\ell(\widetilde{\boldsymbol{\beta}}) - \ell(\widehat{\boldsymbol{\beta}})\} \overset{a}{\sim} \chi^2_q.$$

De acordo com o teste de razão de verosimilhanças a hipótese nula $H_0 : C\boldsymbol{\beta} = \boldsymbol{\xi}$ é rejeitada a favor de $H_1 : C\boldsymbol{\beta} \neq \boldsymbol{\xi}$, a um nível

[4]Por exemplo, o *software* SPSS usa a estatística de Wald para estas situações.

de significância α, se o valor observado da estatística Λ for superior ao quantil de probabilidade $1-\alpha$ de um χ^2_q.

Exemplo 2.4 Suponhamos que queremos testar

$$H_0 : \boldsymbol{\beta}_r = \mathbf{0} \quad versus \quad H_1 : \boldsymbol{\beta}_r \neq \mathbf{0},$$

onde $\boldsymbol{\beta}_r$ é um subvector de $\boldsymbol{\beta}$ de dimensão r.

O uso da *estatística de razão de verosimilhanças* não envolve grandes dificuldades computacionais, já que, para calcular a estatística Λ basta usar o método iterativo de mínimos quadrados ponderados para obter,

(i) a estimativa de máxima verosimilhança, $\widehat{\boldsymbol{\beta}}_0$, do vector parâmetro $\boldsymbol{\beta}_0$ que corresponde ao subvector de $\boldsymbol{\beta}$ sem as componentes que constituem $\boldsymbol{\beta}_r$, e a respectiva *log-verosimilhança* $\ell(\widehat{\boldsymbol{\beta}}_0)$,

(ii) a estimativa de máxima verosimilhança $\widehat{\boldsymbol{\beta}}$, do vector parâmetro $\boldsymbol{\beta}$ e a respectiva *log-verosimilhança* $\ell(\widehat{\boldsymbol{\beta}})$,

ou seja ajustar dois modelos aos dados (em que um é um submodelo do outro).

Hipóteses mais gerais da forma (2.18) requerem, contudo, mais trabalho computacional.

A *estatística de razão de verosimilhanças* é mais utilizada para comparar modelos que estão encaixados, isto é, modelos em que um é submodelo do outro, tal como iremos ver no capítulo seguinte.

2.3.3 Estatística de Rao

Seja novamente $\widetilde{\boldsymbol{\beta}}$ o estimador de máxima verosimilhança de $\boldsymbol{\beta}$ sujeito à restrição imposta pela hipótese nula $C\boldsymbol{\beta} = \boldsymbol{\xi}$.

A *Estatística de Rao* ou *Estatística score* para testar (2.18) é definida por

$$\mathcal{U} = [S(\widetilde{\boldsymbol{\beta}})]^T \mathcal{I}^{-1}(\widetilde{\boldsymbol{\beta}}) S(\widetilde{\boldsymbol{\beta}}) \tag{2.21}$$

A ideia por trás da sugestão desta estatística é a seguinte: Se $\widehat{\boldsymbol{\beta}}$ é o estimador de $\boldsymbol{\beta}$ não restrito , isto é, calculado sem quaisquer restrições, então sabemos que $s(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$. Se substituirmos $\widehat{\boldsymbol{\beta}}$ pelo estimador de máxima verosimilhança sob H_0, isto é por $\widetilde{\boldsymbol{\beta}}$ $s(\widetilde{\boldsymbol{\beta}})$ será significativamente diferente de zero se H_0 não for verdadeira. A estatística $\mathcal{U}$ mede assim a "distância" entre $s(\widetilde{\boldsymbol{\beta}})$ e $\mathbf{0}$.

Tal como para os outros testes, usando a *estatística score* rejeita-se H_0 a um nível de significância α se o valor observado de $\mathcal{U}$ for superior ao quantil de probabilidade $1-\alpha$ de um χ^2_q.

A *estatística score* é útil em situações em que já se calculou um estimador restrito para $\boldsymbol{\beta}$. Tem a vantagem em relação à *estatística de razão de verosimilhanças* de não requerer o cálculo do estimador não restrito. Além disso, tal como a *estatística de Wald* pode ser utilizada em modelos com parâmetro de sobredispersão, já que para o seu cálculo só há necessidade de conhecer os momentos de $1^{\underline{a}}$ e $2^{\underline{a}}$ ordem.

Como se viu, sob H_0 as distribuições assintóticas das três estatísticas são idênticas. A qualidade da aproximação das distribuições exactas das estatísticas $\Lambda, \mathcal{W}$ e $\mathcal{U}$ para a distribuição assintótica, depende da dimensão da amostra n e da forma do logaritmo da função de verosimilhança. Fahrmeir and Tutz (1994) exemplificam a situação com um modelo de Poisson com função de ligação linear.

Se a *log-verosimilhança* for uma função quadrática em $\boldsymbol{\beta}$ então as três estatísticas coincidem. Para amostras pequenas pode haver uma diferença considerável no valor destas estatísticas. Uma

discussão detalhada pode encontrar-se em Buse (1982).

2.4 Quasi-verosimilhança

Todo o estudo de inferência desenvolvido até aqui em MLG, foi feito supondo válido o modelo

$$f(y_i|\theta_i,\phi,\omega_i) = \exp\left\{\frac{\omega_i}{\phi}(y_i\theta_i - b(\theta_i)) + c(y_i,\phi,\omega_i)\right\} \qquad (2.22)$$

o qual é um caso particular do modelo referido em (1.4) com $a_i(\phi) = \frac{\phi}{\omega_i}$ para ω_i's conhecidos. Esta hipótese é muitas vezes irrealista. Por exemplo, no modelo normal, implica que as variáveis aleatórias Y_i, $i = 1, ..., n$, têm a mesma variância, se $\omega_i = 1$, ou variâncias proporcionais, isto é, $var(Y_i) = \frac{\sigma^2}{\omega_i}$.

Um modelo mais geral seria tal que $a_i(\phi) = \phi_i$, ou seja, um modelo com n parâmetros ϕ_i perturbadores. Neste caso há demasiados parâmetros e portanto o estimador de máxima verosimilhança para $\boldsymbol{\beta}$ pode mesmo não existir. Uma solução será assumir também uma estrutura para os ϕ do tipo, *e.g.*, $\phi_i = h^*(\boldsymbol{\eta}, \mathbf{z}_i)$, com $h^*(\cdot,\cdot)$ devidamente especificada, onde $\boldsymbol{\eta}$ é um vector parâmetro desconhecido, o qual pode incluir $\boldsymbol{\beta}$ como subvector. Estimação dos parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{\eta}$ pode fazer-se, como habitualmente, por máxima verosimilhança. Não vamos prosseguir aqui esta abordagem, mas pode encontrar-se mais detalhes em Smyth (1989).

Outro problema, já referido, que surge com o modelo proposto é o da possibilidade de existência de *sobredispersão*. Uma solução possível apontada para tratar deste problema com o modelo binomial, ou com o modelo Poisson, é introduzir um parâmetro de *sobredispersão* ϕ desconhecido, isto é, admitir que para estes modelos ainda se tem $var(Y_i) = \phi V(\mu_i)$ (sabemos que, tanto no modelo

binomial como no de Poisson, isto é verdade se se fizer $\phi = 1$). O problema é que, ao proceder assim, já não podemos escrever o modelo na forma da família exponencial. O modelo para Y_i passa apenas a estar definido através do valor médio e da variância, não sendo possível o recurso à verosimilhança para fazer inferências.

No modelo linear $Y_i = \mathbf{z}_i^T \boldsymbol{\beta} + \varepsilon_i$ em que apenas se admite que os erros são não correlacionados, o vector parâmetro $\boldsymbol{\beta}$ é estimado usando, como se sabe, o método dos mínimos quadrados. Os estimadores assim obtidos coincidem com os de máxima verosimilhança quando o modelo é normal. Um modo semelhante de ultrapassar as dificuldades associadas à não especificação da distribuição nos MLG, ou até de resolver o problema anteriormente referido em que admitimos que $a_i(\phi) = \phi_i$, $i = 1, ..., n$, sem especificar ϕ, passa pelo recurso ao conceito de *quasi-verosimilhança* que iremos introduzir.

Consideremos o caso em que só especificamos os valores médios de Y_i e as suas variâncias, isto é, admitimos que

$$E(Y_i) = \mu_i \qquad var(Y_i) = \phi V(\mu_i).$$

Consideremos a variável (omitindo o indíce i para simplificação)

$$U = U(\mu, Y) = \frac{Y - \mu}{\phi V(\mu)}.$$

Esta variável é tal que

$$E(U) = 0, \qquad var(U) = \frac{1}{\phi V(\mu)}, \qquad -E\Big[\frac{\partial U}{\partial \mu}\Big] = var(U),$$

comportando-se assim como uma função *score*. Como a função *score* é a derivada da função de *log-verosimilhança* podemos esperar que o integral de U (caso exista) se comporte como uma função de *log-verosimilhança*.

Definição 2

Seja Ξ o espaço paramétrico, isto é $\mu \in \Xi$. Diz-se que a função $Q : \Xi \to I\!R$, definida por

$$Q(\mu, y) = \int_y^\mu u(t, y)dt = \int_y^\mu \frac{y - t}{\phi V(t)}dt$$

é uma função de **quasi-verosimilhança** *(correctamente seria de quasi-log-verosimilhança).* $\diamondsuit$

No caso em que temos n observações de variáveis aleatórias independentes, definimos

$$Q(\mu, y) = \sum_{i=}^{n} Q(\mu_i, y_i)$$

Esta função além de partilhar de muitas das propriedades formais que o logaritmo de uma função verosimilhança pode mesmo ser uma função de *log-verosimilhança*. Prova-se que se existir uma função de *log-verosimilhança* ℓ tal que

$$\frac{\partial \ell}{\partial \mu} = \frac{y - \mu}{\phi V(\mu)},$$

com $E(Y) = \mu$ e $var(Y) = \phi V(\mu)$, então ℓ tem a estrutura correspondente a uma função de *log-verosimilhança* da família exponencial.

Nos MLG sabemos que $\mu_i = h(\eta_i) = h(\mathbf{z}_i^T \boldsymbol{\beta})$. Assim a função de *quasi-verosimilhança* $Q(\mu, y)$ é dada por

$$Q(\mu, y) = \sum_{i=1}^{n} \frac{y_i - h(\mathbf{z}_i^T \boldsymbol{\beta})}{var(Y_i)},$$

e, se igualarmos a zero as derivadas de $Q(\mu_i, y_i)$ em ordem a β_j, para $j = 1, ...p$, obtemos o sistema de equações

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0 \quad j = 1, ..., p$$

$$= \sum_{i=1}^{n} \frac{(y_i - \mu_i) z_{ij}}{V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \qquad (2.23)$$

as quais não dependem de ϕ e coincidem com as obtidas em (2.7).

À função $s^*(\boldsymbol{\beta}) = \frac{\partial Q}{\partial \boldsymbol{\beta}}$ damos o nome de função *quasi-score* ou *função de estimação generalizada* e às equações em (2.23) damos o nome de *equações de quasi-verosimilhança*, sendo as estimativas resultantes, estimativas de *quasi-máxima verosimilhança*. Note-se que quando a função de variância é igual a 1 o método reduz-se ao método dos mínimos quadrados.

As propriedades assintóticas dos estimadores de quasi-verosimilhança $\widehat{\boldsymbol{\beta}}^*$ podem ser obtidas sob condições de regularidade semelhantes às necessárias para os estimadores de máxima verosimilhança. Em particular pode mostrar-se que

$$\widehat{\boldsymbol{\beta}}^* \overset{a}{\sim} N_p\Big(\boldsymbol{\beta}, (\mathcal{I}^*)^{-1}(\widehat{\boldsymbol{\beta}}^*) V(\widehat{\boldsymbol{\beta}}^*) (\mathcal{I}^*)^{-1}(\widehat{\boldsymbol{\beta}}^*)\Big),$$

onde

$$\mathcal{I}^*(\boldsymbol{\beta}) = E\Big(- \frac{\partial s^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\Big)$$

e

$$V(\boldsymbol{\beta}) = cov(s^*(\boldsymbol{\beta})).$$

Comparando este resultado com o obtido quando os modelos estão completamente especificados, vemos que, essencialmente, apenas a matriz de covariância de $\widehat{\boldsymbol{\beta}}^*$ tem de ser corrigida. Assim, o método de quasi-verosimilhança permite a obtenção de estimadores consistentes e assintoticamente normais para $\boldsymbol{\beta}$, com apenas uma perda de eficiência. Para que esta perda de eficiência seja pequena é necessário que a estrutura de variância proposta seja o mais próxima possível da verdadeira estrutura de variância.

Também é possível proceder a testes de hipóteses semelhantes aos da secção anterior através do uso de estatísticas de Wald e de

Rao modificadas. Por exemplo, a *Estatística de Wald modificada* para testar

$$H_0 : C\boldsymbol{\beta} = \boldsymbol{\xi} \quad versus \quad H_1 : C\boldsymbol{\beta} \neq \boldsymbol{\xi},$$

é dada por

$$\mathcal{W}_m = (C\widehat{\boldsymbol{\beta}}^* - \boldsymbol{\xi})^T [CAC^T]^{-1} (C\widehat{\boldsymbol{\beta}}^* - \boldsymbol{\xi}),$$

onde $A = (\mathcal{I}^*)^{-1}(\widehat{\boldsymbol{\beta}}^*)V(\widehat{\boldsymbol{\beta}}^*)(\mathcal{I}^*)^{-1}(\widehat{\boldsymbol{\beta}}^*)$ é a matriz de covariância corrigida. $\mathcal{W}_m$ ainda tem uma distribuição assintótica χ^2_r, onde r é a característica da matriz C. Veja-se detalhes em White (1982).

Tratamento semelhante pode ser feito para o caso em que o modelo se encontra completamente especificado, mas os parâmetros ϕ são distintos. Veja-se, para o efeito, Shao (1998, pg. 246-247).

Capítulo 3

Selecção e Validação de Modelos

No estudo feito até aqui admitiu-se que o modelo proposto, em termos da combinação - distribuição da variável resposta e função de ligação - era um modelo adequado. No entanto, quando se trabalha com muitas covariáveis, tem interesse saber qual o modelo mais parcimonioso, isto é, com o menor número de variáveis explicativas, que ofereça uma boa interpretação do problema posto e que ainda se ajuste bem aos dados. O problema da selecção do modelo corresponde à procura do "melhor modelo", no sentido de ser um modelo que atinge um bom equilíbrio entre os três factores "bom ajustamento", "parcimónia" e "interpretação". Dado que no processo de selecção há uma série de modelos em consideração, convém descrever vários que são comummente referidos durante o processo.

- **Modelo completo ou saturado**

  Consideremos o modelo linear generalizado sem a estrutura linear $\boldsymbol{\eta} = Z\boldsymbol{\beta}$, isto é com n parâmetros $\mu_1, ..., \mu_n$, linearmente

independentes, sendo a matriz do modelo a matriz identidade de $n \times n$. Este modelo atribui toda a variação dos dados à componente sistemática. Como as estimativas de máxima verosimilhança dos μ_i são as próprias observações, isto é, $\widehat{\mu}_i = y_i$, o modelo ajusta-se exactamente, reproduzindo os próprios dados. Não oferece qualquer simplificação e, como tal, não tem interesse na interpretação do problema, já que não faz sobressair características importantes transmitidas pelos dados. Além disso tem pouca hipótese de ser um modelo adequado em réplicas do estudo. A sua consideração é contudo necessária na formulação da teoria da selecção de modelos, como iremos ter oportunidade de ver.

- **Modelo nulo**

  O modelo mais simples que se pode imaginar é o modelo com um único parâmetro. Corresponde a assumir que todas as variáveis Y_i têm o mesmo valor médio μ. É um modelo, de interpretação sem dúvida simples, mas que raramente captura a estrutura inerente aos dados. A matriz do modelo é, neste caso, um vector coluna unitário. Contrariamente ao modelo anterior, este modelo atribui toda a variação nos dados à componente aleatória.

- **Modelo maximal**

  O modelo maximal é o modelo que contém o maior número de parâmetros, e portanto, o mais complexo, que estamos preparados a considerar.

- **Modelo minimal**

  Contrariamente ao modelo maximal, o modelo minimal é o modelo mais simples, com o menor número de parâmetros,

que ainda se ajusta adequadamente aos dados. Este modelo embora adaptando-se aos dados e podendo até ser adequado para réplicas do estudo, pode esconder características ainda importantes dos dados.

- **Modelo corrente**

  Em geral trabalha-se com modelos encaixados, isto é, passa-se do modelo maximal para o modelo minimal por exclusão de termos da desvio. O modelo corrente, é qualquer modelo com q parâmetros linearmente independentes situado entre o modelo maximal e o modelo minimal, e que está a ser sujeito a investigação.

3.1 Qualidade de Ajustamento

3.1.1 Função desvio

O modelo saturado é útil para julgar da qualidade de ajustamento de um determinado modelo em investigação, que passamos a designar por M, através da introdução de uma medida da distância dos valores ajustados $\widehat{\mu}$ com esse modelo e dos correspondentes valores observados y. Essa medida de discrepância entre o modelo saturado e o modelo corrente, é baseada na estatística de razão de verosimilhanças de Wilks referida na secção 2.3.2[5].

Como vimos na secção 2.1.1, o logaritmo da função de verosimilhança (*função log-verosimilhança*) de um modelo linear generaliza-

[5]Seguindo a sugestão de Cordeiro (1986) traduzimos o termo "deviance" por desvio.

do é dada por

$$\ln L(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\omega_i[y_i q(\mu_i) - b(q(\mu_i))]}{\phi} + c(y_i, \phi, \omega_i)$$

em que se substituiu θ_i por $q(\mu_i)$, para fazer salientar, na *função log-verosimilhança* , a relação funcional existente entre θ_i e μ_i.

Como para o modelo saturado - que passamos a designar por S - se tem $\widehat{\mu}_i = y_i$, o máximo da *função log-verosimilhança* para este modelo é

$$\ell_S(\widehat{\boldsymbol{\beta}}_S) = \sum_{i=1}^{n} \frac{\omega_i[y_i q(y_i) - b(q(y_i))]}{\phi} + c(y_i, \phi, \omega_i).$$

Por outro lado, se designarmos por $\widehat{\mu}_i$ a estimativa de máxima verosimilhança de μ_i, para $i = 1, .., n$, o máximo da *função log-verosimilhança* para o modelo em investigação com, digamos, m parâmetros na desvio é

$$\ell_M(\widehat{\boldsymbol{\beta}}_M) = \sum_{i=1}^{n} \frac{\omega_i[y_i q(\widehat{\mu}_i) - b(q(\widehat{\mu}_i))]}{\phi} + c(y_i, \phi, \omega_i).$$

Os índices em $\widehat{\boldsymbol{\beta}}$ e ℓ correspondem ao modelo em relação ao qual são calculados.

Se compararmos o modelo em investigação M com o modelo saturado S através da estatística de razão de verosimilhanças, obtemos

$$\begin{aligned} D^*(\mathbf{y}; \widehat{\boldsymbol{\mu}}) &= -2(\ell_M(\widehat{\boldsymbol{\beta}}_M) - \ell_S(\widehat{\boldsymbol{\beta}}_S)) \\ &= -2\sum_i \frac{\omega_i}{\phi}\Big\{\Big[y_i q(\widehat{\mu}_i) - b(q(\widehat{\mu}_i))\Big] - \Big[y_i q(y_i) - b(q(y_i))\Big]\Big\} \\ &= \frac{D(\mathbf{y}; \widehat{\boldsymbol{\mu}})}{\phi}. \end{aligned} \tag{3.1}$$

A $D^*(\mathbf{y}; \widehat{\boldsymbol{\mu}})$ definido em (3.1) damos o nome de **desvio reduzido**; ao numerador $D(\mathbf{y}; \widehat{\boldsymbol{\mu}})$ damos o nome de **desvio** para o modelo corrente. Note-se que o *desvio* é só função dos dados.

Como se pode observar de (3.1) o *desvio* pode ser decomposto

$$\begin{aligned} D(\mathbf{y};\widehat{\boldsymbol{\mu}}) &= \sum_i 2\omega_i\Big\{y_i(q(y_i)-q(\widehat{\mu}_i))-b(q(y_i))+b(q(\widehat{\mu}_i))\Big\} \\ &= \sum_i d_i \end{aligned}$$

na soma de parcelas d_i que medem a diferença dos logaritmos das verosimilhanças observada e ajustada para cada observação. A soma destas componentes é assim uma medida da discrepância total entre as duas log-verosimilhanças.

É fácil de verificar que o *desvio* é sempre maior ou igual a zero, e decresce à medida que covariáveis vão sendo adicionadas ao modelo nulo, tomando obviamente o valor zero para o modelo saturado.

Uma outra propriedade importante do *desvio* é a aditividade para modelos encaixados. Com efeito, suponhamos que temos dois modelos intermédios M_1 e M_2 estando M_2 encaixado em M_1, isto é, são modelos do mesmo tipo, mas o modelo M_2 contém menos parâmetros na desvio que o modelo M_1. Se designarmos por $D(\mathbf{y};\widehat{\boldsymbol{\mu}}_j)$ o desvio do modelo M_j, $j=1,2$, então a estatística da razão de verosimilhanças para comparar estes dois modelos resume-se a

$$-2(\ell_{M_2}(\widehat{\boldsymbol{\beta}}_2)-\ell_{M_1}(\widehat{\boldsymbol{\beta}}_1)) = \frac{D(\mathbf{y};\widehat{\boldsymbol{\mu}}_2)-D(\mathbf{y};\widehat{\boldsymbol{\mu}}_1)}{\phi}.$$

Dos resultados do capítulo anterior sabe-se que, sob a hipótese do modelo M_1 ser verdadeiro, então

$$\frac{D(\mathbf{y};\widehat{\boldsymbol{\mu}}_2)-D(\mathbf{y};\widehat{\boldsymbol{\mu}}_1)}{\phi} \overset{a}{\sim} \chi^2_{p_1-p_2},$$

onde p_j, representa a dimensão do vector $\boldsymbol{\beta}$ para o modelo M_j, $j=1,2$. A comparação de modelos encaixados, pode então ser feitas à custa da diferença dos *desvios* de cada modelo.

O *desvio* também costuma ser usado para julgar da adequabilidade de um modelo. No entanto não se conhece, em geral, a distribuição quer exacta, quer assintótica do desvio. Há certos casos especiais, *e.g.*, distribuição normal ou gaussiana inversa, para os quais resultados exactos podem ser obtidos. Também, por vezes, o desvio pode ser aproximado pela distribuição χ^2. Em regra geral, no entanto, esta aproximação é bastante má mesmo para grandes amostras. Assim, a análise do desvio não é mais do que um guia no estudo da adequabilidade de um modelo, embora muitas vezes na prática se faça comparação do valor observado do desvio, para um modelo com p parâmetros na desvio, com o valor crítico de um χ^2_{n-p}. Se esse valor observado for superior a $\chi^2_{n-p,\alpha}$, então o modelo é considerado não adequado. O aperfeiçoamento deste teste através da introdução de um factor de correcção é discutido em Cordeiro (1986).

Exemplo 3.1 Modelo Normal

Para o caso do modelo normal a *função log-verosimilhança* é

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{1}{\sigma^2}(y_i\mu_i - \frac{\mu_i^2}{2}) + c(y_i, \phi, \omega_i).$$

Assim, para o modelo saturado

$$\ell_S(\widehat{\boldsymbol{\beta}}_S) = \sum_{i=1}^{n} \frac{1}{\sigma^2}(y_i^2 - \frac{y_i^2}{2}) + c(y_i, \phi, \omega_i) = \sum_{i=}^{n} \frac{y_i^2}{2\sigma^2} + c(y_i, \phi, \omega_i)$$

e para o modelo corrente

$$\begin{aligned} \ell_M(\boldsymbol{\beta}_M) &= \sum_{i=1}^{n} \frac{1}{\sigma^2}(y_i\widehat{\mu}_i - \frac{\widehat{\mu}_i^2}{2}) + c(y_i, \phi, \omega_i) \\ &= \sum_{i=}^{n} \frac{1}{2\sigma^2}(2y_i\widehat{\mu}_i - \widehat{\mu}_i^2) + c(y_i, \phi, \omega_i). \end{aligned}$$

Consequentemente o *desvio reduzido* é

$$D^*(\mathbf{y};\widehat{\boldsymbol{\mu}}) = 2(\ell_S(\widehat{\boldsymbol{\beta}}_S) - \ell_M(\widehat{\boldsymbol{\beta}}_M)) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \widehat{\mu}_i)^2,$$

que, como se sabe da teoria do modelo linear, sob a hipótese do modelo M ser verdadeiro, tem uma distribuição exacta de um χ^2 com $n-p$ graus de liberdade, sendo p a dimensão do vector $\boldsymbol{\beta}$ para o modelo em questão.

Exemplo 3.2 Modelo Poisson

Para o caso do modelo Poisson com função de ligação canónica, temos que a *função log-verosimilhança* é dada por

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \ln \mu_i - \mu_i - \ln y_i!,$$

sendo

$$\ell_M(\widehat{\boldsymbol{\beta}}_M) = \sum_{i=1}^{n} y_i \ln \widehat{\mu}_i - \widehat{\mu}_i - \ln y_i!,$$

$$\ell_S(\widehat{\boldsymbol{\beta}}_S) = \sum_{i=1}^{n} y_i \ln y_i - y_i - \ln y_i!.$$

Consequentemente o desvio [6] para o modelo de Poisson é

$$D^*(\mathbf{y};\widehat{\boldsymbol{\mu}}) = 2[\sum_{i=1}^{n} y_i \ln \frac{y_i}{\widehat{\mu}_i} - \sum_{i=1}^{n}(y_i - \widehat{\mu}_i)].$$

Quando a matriz de especificação do modelo tem uma primeira coluna unitária tem-se que $\sum_{i=1}^{n}(y_i - \widehat{\mu}_i) = 0$, como aliás se referiu na secção 2.1.2. Assim, nestas condições, o desvio coincide com a estatística habitual G^2 usada para julgar da adequabilidade dos modelos log-lineares em tabelas de contingência (veja-se *e.g.*, Christensen, 1997).

[6] Note-se que como $P(Y_i = 0) = e^{\mu_i}$, o termo $y_i \ln y_i$ não aparece na expressão do desvio para os casos em que $y_i = 0$

Outros exemplos da função *desvio* para modelos lineares generalizados encontram-se na tabela 3.1.

Tabela 3.1: Expressões da função *desvio* para alguns modelos.

normal	$\sum_i (y_i - \widehat{\mu}_i)^2$
Poisson	$2[\sum_i y_i \ln \frac{y_i}{\widehat{\mu}_i} - \sum_i (y_i - \widehat{\mu}_i)]$
binomial	$2[\sum_i m_i y_i \ln \frac{y_i}{\widehat{\mu}_i} + \sum_i m_i (1 - y_i) \ln \frac{1-y_i)}{1-\widehat{\mu}_i}]$
gama	$2 \sum_i \left\{ - \ln \frac{y_i}{\widehat{\mu}_i} + \frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i} \right\}$
gaussiana inversa	$\sum_i \frac{(y_i - \widehat{\mu}_i)^2}{y_i \widehat{\mu}_i^2}$

3.1.2 Estatística de Pearson generalizada

Outra medida da adequabilidade de modelos é a *estatística de Pearson generalizada* já definida na secção 2.2.2,

$$X^2 = \sum_i \frac{\omega_i (y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}. \tag{3.2}$$

Para a distribuição normal, a estatística X^2 coincide, tal como o *desvio*, com a soma dos quadrados dos resíduos. Para os modelos Poisson e Binomial coincide com a estatística original de Pearson. Novamente é costume usar X^2 para testar a adequabilidade de um modelo comparando o valor observado com o quantil de probabilidade $1 - \alpha$ de uma distribuição de χ^2 com $n - p$ graus de liberdade. Contudo, tal como acontece com o *desvio*, a aproximação pelo χ^2 da distribuição de X^2 pode ser, em certos modelos, má mesmo para grandes amostras, havendo a necessidade de agrupar os dados

o mais possível, garantindo ao mesmo tempo que o número de observações em cada grupo, digamos n_i não seja pequeno. Assim, as estatísticas a usar para testar a adequabilidade do modelo em consideração devem ser do tipo

$$D(\mathbf{y};\widehat{\boldsymbol{\mu}}) = \sum_{i=1}^{g} 2\omega_i\Big\{y_i(q(y_i) - q(\widehat{\mu}_i)) - b(q(y_i)) + b(q(\widehat{\mu}_i))\Big\},$$

e

$$X^2 = \sum_{i=1}^{g} \frac{\omega_i(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)},$$

onde g é o número de grupos, e o número de observações em cada grupo é suficientemente grande em todos os grupos. Neste caso, a suposição de que ambas as estatísticas têm uma distribuição aproximada de um $\phi\chi^2$ com $g - p$ graus de liberdade, já é menos problemática.[7]

A propriedade da aditividade da função *desvio* faz com que esta seja preferida, em relação à estatística de Pearson, como uma medida da discrepância, embora esta última tenha a vantagem de ter uma interpretação mais directa.

3.2 Selecção de Modelos

Como já se disse, em problemas práticos que requerem uma análise estatística via modelos lineares generalizados, há geralmente um número elevado de covariáveis que podem ser potencialmente importantes para explicar a variabilidade inerente aos dados. Também tem, frequentemente, interesse investigar a influência de possíveis interacções entre as covariáveis. Isto implica obviamente a existência de um número elevado de modelos a considerar de

[7]Consulte-se, *e.g.*, Fahrmeir and Tutz (1994, pg. 48).

modo a escolher um modelo possível para explicar o fenómeno em estudo. Se pensarmos num modelo maximal como aquele que entra em linha de conta, na sua desvio, com todas as possíveis covariáveis e interacções de interesse entre elas, um submodelo deste é qualquer modelo que é obtido dele por exclusão de algum ou alguns dos termos da desvio. A existência de um número elevado de modelos a considerar, quer se parta do modelo maximal, quer se parta do modelo minimal, traz problemas de ordem combinatória - o número de combinações possíveis torna-se rapidamente não manejável - e de ordem estatística - como decidir sobre o equilíbrio entre o efeito da inclusão ou exclusão de um termo na discrepância entre y e $\widehat{\mu}$ e a complexidade de um modelo maior? Há pois necessidade de estabelecer uma estratégia para a selecção do melhor, ou dos melhores modelos, já que raramente se pode falar na existência de um único "melhor modelo".

Ter um submodelo M_1 de um modelo M, com vector parâmetro $\boldsymbol{\beta}$ de dimensão p, corresponde a ter um modelo com vector de parâmetros $\boldsymbol{\beta}_1$ que é um subvector de $\boldsymbol{\beta}$. Sem perda de generalidade, podemos assumir a partição de $\boldsymbol{\beta}$ em $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^T$. Assim, a adequabilidade de um submodelo pode ser testada formalmente como

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0}, \qquad versus \qquad H_1 : \boldsymbol{\beta}_2 \neq \mathbf{0}.$$

Esta hipótese pode ser testada usando a metodologia explicada na secção 2.3.

Designemos, como anteriormente, a função *score*, a matriz de informação de Fisher e a sua inversa para o modelo especificado em H_1, respectivamente, por $s(\boldsymbol{\beta})$, $\mathcal{I}(\boldsymbol{\beta})$ e $A(\boldsymbol{\beta})$. Então, em conformidade com a partição relativa ao vector $\boldsymbol{\beta}$, podemos particioná-las

do seguinte modo

$$s = \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \qquad \mathcal{I} = \begin{pmatrix} I_{11} & I_{12} \\ I_{12}^T & I_{22} \end{pmatrix} \qquad A = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{pmatrix}.$$

Seja ainda $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2)^T$ e $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\beta}}_1, \mathbf{0})^T$ os estimadores de máxima verosimilhança de $\boldsymbol{\beta}$, sob o modelo em H_1 e H_0, respectivamente. Para o caso em que existe um parâmetro ϕ desconhecido, sejam $\widehat{\phi}$ e $\widetilde{\phi}$ estimadores consistentes de ϕ sob H_1 e H_0, respectivamente. Então, de acordo com a secção 2.3, podemos considerar as seguintes estatísticas de teste para testar H_0 contra H_1

- A estatística de razão de verosimilhanças

$$\Lambda = -2\{\ell(\widetilde{\boldsymbol{\beta}}_1, \mathbf{0}, \widetilde{\phi}) - \ell(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\phi})\},$$

- A estatística de Wald

$$\mathcal{W} = \widehat{\boldsymbol{\beta}}_2^T \hat{A}_{22}^{-1} \widehat{\boldsymbol{\beta}}_2,$$

- A estatística de Rao

$$\mathcal{U} = \tilde{s}_2^T \tilde{A}_{22} \tilde{s}_2,$$

  onde $\tilde{A}, \tilde{s}$ e $\hat{A}, \hat{s}$, significa que os cálculos relativos à matriz A e à função *score* são feitos em $(\widetilde{\boldsymbol{\beta}}, \widetilde{\phi})$ e $(\widehat{\boldsymbol{\beta}}, \widehat{\phi})$, respectivamente.

Ainda se pode considerar uma estatística de razão de verosimilhanças modificada dada por

$$\Lambda_m = -2\{\ell(\widetilde{\boldsymbol{\beta}}_1^*, \mathbf{0}, \widetilde{\phi}) - \ell(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\phi})\},$$

onde $\widetilde{\boldsymbol{\beta}}_1^*$ é uma aproximação de $1^{\underline{a}}$ ordem a $\widetilde{\boldsymbol{\beta}}_1$ dada por

$$\widetilde{\boldsymbol{\beta}}_1^* = \widehat{\boldsymbol{\beta}}_1 - \hat{A}_{12}^T \hat{A}_{22}^{-1} \widehat{\boldsymbol{\beta}}_2.$$

Sob a hipótese H_0 todas as estatísticas de teste têm uma distribuição assintótica de um χ^2_r, onde r é a dimensão do vector $\boldsymbol{\beta}_1$, desde que, obviamente, se verifiquem as condições de regularidade necessárias.

Segundo Fahrmeir and Tutz (1994) é preferível, em geral, usar a estatística de razão de verosimilhanças se o número de covariáveis for pequeno e as amostras tiverem uma dimensão moderada. Quando as amostras são de dimensão elevada, as estatísticas de teste tendem a dar resultados semelhantes e é razoável usar quer $\mathcal{W}, \mathcal{U}$ ou Λ_m, já que são mais fáceis e rápidas de calcular.

A estatística a usar pode depender da metodologia de selecção que se está a seguir. Por exemplo, a estatística de Wald, por usar a estimativa não restrita de máxima verosimilhança, é útil na comparação de modelos quando se começa por formar o modelo maximal e se consideram modelos alternativos pela exclusão de covariáveis. (selecção *backward*). A estatística de Rao, pelo contrário, é útil na escolha de modelos, quando se parte do modelo nulo, *i.e.*, o modelo sem covariáveis, ou de um modelo minimal e se consideram modelos alternativos pela inclusão de covariáveis (selecção *forward*).

A análise do *desvio* é uma generalização, para os MLG, da análise de variância usada na análise de modelos lineares normais. A diferença entre os *desvios reduzidos* de dois modelos encaixados coincide com a estatística de razão de verosimilhanças, quando a hipótese H_0 diz respeito ao modelo menor e a hipótese H_1 ao modelo maior.

Outro critério de selecção possível é o critério de informação de Akaike (1974), o qual é baseado na *função log-verosimilhança* , com a introdução de um factor de correcção como modo de penalização da complexidade do modelo. A estatística correspondente para o

modelo em H_0 é,

$$AIC = -2\ell(\widetilde{\boldsymbol{\beta}}_1, \mathbf{0}, \widetilde{\phi}) + 2r,$$

onde $r = dim(\boldsymbol{\beta}_1)$. Um valor baixo para AIC é considerado como representativo de um melhor ajustamento e na selecção de modelos devemos ter como objectivo a minimização de AIC. Note-se a seguinte relação existente entre AIC e o *desvio reduzido* relativo ao modelo especificado por H_0 (estamos a supor que o parâmetro ϕ ou é conhecido, ou é substituído por uma estimativa consistente)

$$\begin{aligned} AIC_r &= -2\ell(\widetilde{\boldsymbol{\beta}}_1, \mathbf{0}) + 2\ell(\widehat{\boldsymbol{\beta}}_S) - 2\ell(\widehat{\boldsymbol{\beta}}_S) + 2r \\ &= D_r^* + 2r - 2\ell(\widehat{\boldsymbol{\beta}}_S) \end{aligned}$$

onde o indice r serve para especificar o modelo em consideração e S, como habitualmente, refere-se ao modelo saturado.

Cordeiro (1986) sugere ainda a seguinte modificação do critério de Akaike para seleccionar modelos,

$$C_r^* = D_r^* + 2r - n = AIC_r + 2\ell(\widehat{\boldsymbol{\beta}}_S) - n.$$

Um gráfico de C_r^* contra r fornece uma boa indicação para comparação de modelos. Se o modelo for verdadeiro é de esperar que C_r^* seja próximo de r.

Se tivermos dois modelos encaixados M_1 e M_2 com, digamos r_1 e r_2 parâmetros respectivamente, onde $r_1 > r_2$, vem

$$AIC_{r_1} - AIC_{r_2} = C_{r_1}^* - C_{r_2}^* = D_{r_1}^* - D_{r_2}^* + 2(r_1 - r_2)$$

e, supondo que o modelo M_2 é verdadeiro, tem-se (Cordeiro, 1986) $E(AIC_{r_1} - AIC_{r_2}) = r_1 - r_2 + 0(n^{-1})$. Na comparação de modelos sucessivamente mais ricos, o declive esperado do segmento de recta

que une AIC_{r_1} e AIC_{r_2} deve estar próximo de 1 supondo o modelo menor M_2 verdadeiro. Pares de modelos que exibem declive maior do que 1 são indicação de que o modelo maior não é significativamente melhor que o modelo menor.

3.3 Análise de Resíduos

A análise de resíduos é útil, não só para uma avaliação local da qualidade de ajustamento de um modelo no que diz respeito à escolha da distribuição, da função de ligação e de termos do preditor linear, como também para ajudar a identificar observações mal ajustadas, *i.e.*, que não são bem explicadas pelo modelo.

Um resíduo R_i deve exprimir a discrepância entre o valor observado y_i e o valor $\widehat{\mu}_i$ ajustado pelo modelo. No modelo linear normal em que o vector das respostas $\mathbf{Y}$ se pode escrever na forma

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = Z\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I)$$

tem-se $\widehat{\boldsymbol{\beta}} = (Z^T Z)^{-1} Z^T \mathbf{y}$ e o vector dos resíduos é naturalmente dado por $\mathbf{R} = \mathbf{y} - \widehat{\mathbf{y}}$ onde $\widehat{\mathbf{y}} = \widehat{\boldsymbol{\mu}} = Z\widehat{\boldsymbol{\beta}}$ é o vector dos valores ajustados. No caso dos modelos lineares generalizados não existe necessariamente uma componente ε_i para o qual o resíduo R_i seja uma estimativa; faz portanto sentido, como iremos ver, considerar outras definições de resíduos.

Outra quantidade de interesse na análise dos resíduos, no caso do modelo linear normal, é a matriz de projecção $H = Z(Z^T Z)^{-1} Z^T$ a qual é tal que $\widehat{\mathbf{y}} = H\mathbf{y}$ (e que por essa razão se designa, em inglês, por *hat matrix*). Esta matriz é simétrica e idempotente. Os seus elementos h_{ij} são uma medida da influência exercida por y_j em $\widehat{y}_i$. A influência exercida por y_i em $\widehat{y}_i$ é reflectida pelo elemento da

diagonal principal h_{ii}. Como se tem que $\sum h_{ii} = p$ e $0 \leq h_{ii} \leq 1$, um ponto é considerado influente se $h_{ii} > \frac{2p}{n}$ (Hoaglin and Welsch, 1978).

3.3.1 Matriz de projecção generalizada

Como vimos em (2.14), o processo iterativo conduz, no modelo linear generalizado a

$$\widehat{\boldsymbol{\beta}} = \left(Z^T W Z\right)^{-1} Z^T W \mathbf{u}.$$

Esta equação é idêntica à que se obteria para os estimadores de mínimos quadrados ponderados para o problema de regressão

$$\mathbf{u} = Z\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

ou, alternativamente, a solução de minimos quadrados para o modelo linear

$$\mathbf{u}_0 = Z_0\boldsymbol{\beta} + \widetilde{\varepsilon},$$

onde $\mathbf{u}_0 = W^{\frac{1}{2}}\mathbf{u}$, $\quad Z_0 = W^{\frac{1}{2}}Z$ e, portanto, a matriz de projecção correspondente é

$$\begin{aligned} H = Z_0(Z_0^T Z_0)^{-1} Z_0^T &= W^{\frac{1}{2}} Z (Z^T W^{\frac{1}{2}} W^{\frac{1}{2}} Z)^{-1} Z^T W^{\frac{1}{2}} \\ &= W^{\frac{1}{2}} Z I^{-1}(\boldsymbol{\beta}) Z^T W^{\frac{1}{2}}. \end{aligned} \tag{3.3}$$

O lado direito de (3.3) advém do facto de $\mathcal{I}(\boldsymbol{\beta}) = Z^T W Z$, tal como se viu em (2.10). Esta matriz H é também simétrica e idempotente e pode ser vista como uma matriz de projecção para a qual ainda se tem traço(H)=característica(H). Os elementos da diagonal principal desta matriz são ainda tais que $0 \leq h_{ii} \leq 1$ e valores elevados de h_{ii} correspondem a pontos extremos. Contudo, em contraste com o modelo linear normal, esta matriz não depende apenas da matriz Z,

mas também das estimativas dos parâmetros do modelo, através de W. Como tal, pontos extremos não correspondem necessariamente, apenas, a valores elevados de h_{ii} (McCullagh and Nelder, 1989, pg. 405).

3.3.2 Definições de resíduos

Como se disse pretende-se, ao definir resíduo relativamente à i-ésima observação, uma quantidade R_i que exprima a discrepância entre o valor observado y_i e o valor $\widehat{\mu}_i$ ajustado pelo modelo. É conveniente, para uma análise adequada dos resíduos, que eles sejam padronizados e reduzidos, isto é, que tenham variância constante unitária e, preferencialmente, que sejam aproximadamente normalmente distribuídos. Cálculos um pouco morosos, mas relativamente simples permitem estabelecer que, assintoticamente, se tem (Williams, 1987)

$$\begin{aligned} var(\widehat{\eta}_i) &= \varpi_i h_{ii}, \\ var(\widehat{\mu}_i) &= var(Y_i) h_{ii}, \\ var(Y_i - \widehat{\mu}_i) &= var(Y_i)(1 - h_{ii}), \end{aligned}$$

onde ϖ_i é o i-ésimo elemento da matriz W definido em (2.11) e h_{ii} é o i-ésimo elemento da diagonal principal da matriz de projecção generalizada definida em (3.3).

Têm sido propostas várias definições de resíduos generalizados.

À semelhança da definição de resíduo para o modelo linear normal podemos definir o **resíduo de Pearson** por:

$$R_i^P = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{var(Y_i)}}}$$

$$= \frac{(y_i - \widehat{\mu}_i)w_i}{\sqrt{\widehat{\phi}V(\widehat{\mu}_i)}}. \tag{3.4}$$

O resíduo R_i^P, assim definido, corresponde à contribuição de cada observação para o cálculo da estatística de Pearson generalizada. Atendendo a que se tem $var(Y_i - \widehat{\mu}_i) \approx var(Y_i)(1 - h_{ii})$, o correspondente resíduo, convenientemente padronizado é

$$R_i^{\star P} = \frac{(y_i - \widehat{\mu}_i)w_i}{\sqrt{\widehat{\phi}V(\widehat{\mu}_i)(1 - h_{ii})}}.$$

A desvantagem do resíduo de Pearson é que a sua distribuição é, geralmente, bastante assimétrica para modelos não normais.

Tal como foi sugerido por Anscombe (1953), de modo a conseguir resíduos com uma distribuição o mais próxima possível da normal, pode considerar-se uma transformação adequada $A(y_i)$ da observação y_i e definir o resíduo como

$$R_i^A = \frac{A(y_i) - E[A(Y_i)]}{\sqrt{var[A(Y_i)]}}. \tag{3.5}$$

Fazendo aproximações de primeira ordem tem-se que $E[A(Y_i)] \approx A(\mu_i)$ e $var[A(Y_i)] \approx [A'(\mu_i)]^2 var(Y_i)$. Substituindo em (3.4) e considerando as estimativas correspondentes, obtém-se os chamados **resíduos de Anscombe**

$$R_i^A = \frac{A(y_i) - A(\widehat{\mu}_i)]}{\sqrt{\widehat{var(Y_i)}}A'(\widehat{\mu}_i)}. \tag{3.6}$$

Barndorff-Nielsen (1978) mostra que a transformação a considerar nos modelos lineares generalizados é da forma

$$A(x) = \int \frac{1}{V^{1/3}(x)} dx,$$

onde $V(x)$ é a função de variância.

Outro tipo de resíduo é baseado na função *desvio*. Podemos usar a contribuição da i-ésima observação para a função desvio definida em (3.1),

$$D(\mathbf{y};\widehat{\boldsymbol{\mu}}) = \sum_i 2\omega_i\Big\{y_i(q(y_i) - q(\widehat{\mu}_i)) - b(q(y_i)) + b(q(\widehat{\mu}_i))\Big\} = \sum_i d_i,$$

para dar uma nova definição de resíduo.

Assim o **desvio residual** correspondente à i-ésima observação é definido por

$$R_i^D \;=\; \delta_i \sqrt{d_i}, \tag{3.7}$$

onde $\delta_i = sinal(y_i - \widehat{\mu}_i)$. O desvio residual padronizado é também obtido dividindo o desvio residual R_i^D por $\sqrt{\widehat{\phi}(1-h_{ii})}$, ou seja

$$R^{\star D}_i = \frac{R_i^D}{\sqrt{\widehat{\phi}(1-h_{ii})}}. \tag{3.8}$$

Exemplo 3.3 Resíduos no modelo normal

Para o modelo normal é fácil de verificar que os três tipos de resíduos coincidem. Com efeito, atendendo a que para este modelo $V(x) = 1$, tem-se que $A(x) = x$ e, portanto o resíduo de Pearson (puro, *i.e.*, não padronizado), $R_i^P = y_i - \widehat{\mu}_i$, coincide com o resíduo de Anscombe. Por outro lado, dado que $d_i = (y_i - \widehat{\mu}_i)^2$, tem-se que o desvio residual é também dado por $y_i - \widehat{\mu}_i$

Exemplo 3.4 Resíduos no modelo Poisson

No modelo Poisson tem-se, como se sabe, $V(x) = x$. Deste modo $\int V^{-1/3}(x)dx = \frac{3}{2}x^{2/3}$ e portanto os resíduos de Pearson e de Anscombe puros são, respectivamente

$$R_i^P = \frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i^{1/2}} \qquad R_i^A = \frac{3(y_i^{2/3} - \widehat{\mu}_i^{2/3})}{2\widehat{\mu}_i^{1/6}}.$$

Consultando a tabela 3.1 facilmente se obtém para o desvio residual

$$R_i^D = \delta_i 2^{1/2}(y_i \ln \frac{y_i}{\widehat{\mu}_i} - y_i + \widehat{\mu}_i)^{1/2},$$

onde $\delta_i = sinal(y_i - \widehat{\mu}_i)$.

Exemplo 3.5 Resíduos no modelo binomial

No modelo binomial tem-se que $V(x) = x(1-x)$ e $\omega_i = m_i$. Deste modo $A(x) = \int x^{-1/3}(1-x)^{-1/3}dx$ e o resíduo de Anscombe é dado por:

$$R_i^A = \frac{m_i^{1/2}[A(y_i) - A(\widehat{\mu}_i)]}{[\widehat{\mu}_i(1-\widehat{\mu}_i)]^{1/6}}.$$

Cox and Snell (1968) calculam este resíduo através da função beta incompleta.

Facilmente se obtêm as seguintes expressões para os resíduos de Pearson e desvio residual

$$\begin{aligned} R_i^P &= \frac{m_i^{1/2}(y_i - \widehat{\mu}_i)}{[\widehat{\mu}_i(1-\widehat{\mu}_i)]^{1/2}} \\ R_i^D &= \delta_i \Big[2m_i \Big(\ln \frac{y_i}{\widehat{\mu}_i} + (1-y_i) \ln \Big(\frac{1-y_i}{1-\widehat{\mu}_i} \Big) \Big) \Big]^{1/2}, \end{aligned}$$

onde $\delta_i = sinal(y_i - \widehat{\mu}_i)$.

Cordeiro (1986) faz um estudo comparativo entre os resíduos de Anscombe e os desvios residuais para os modelos Poisson, gama e gaussiano inverso.

Pierce and Schafer (1986) introduzem uma generalização dos resíduos de Anscombe e sugerem outro tipo de resíduos destinados a estabilizar a variância.

Na tabela 3.2 apresentamos um quadro resumo dos três tipos de resíduos para os modelos que temos vindo a considerar.

Tabela 3.2: Expressões dos resíduos para alguns modelos.

	R_i^P	R_i^A	R_i^D
normal	$y_i - \widehat{\mu}_i$	$y_i - \widehat{\mu}_i$	$y_i - \widehat{\mu}_i$
Poisson	$\frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i^{1/2}}$	$\frac{3(y_i^{2/3} - \widehat{\mu}_i^{2/3})}{2\widehat{\mu}_i^{1/6}}$	$\delta_i 2^{1/2}(y_i \ln \frac{y_i}{\widehat{\mu}_i} - y_i + \widehat{\mu}_i)^{1/2}$
binomial	$\frac{m_i^{1/2}(y_i - \widehat{\mu}_i)}{[\widehat{\mu}_i(1-\widehat{\mu}_i)]^{1/2}}$	$\frac{m_i^{1/2}[A(y_i) - A(\widehat{\mu}_i)]}{[\widehat{\mu}_i(1-\widehat{\mu}_i)]^{1/6}}$	$\delta_i[2m_i(\ln \frac{y_i}{\widehat{\mu}_i} + (1-y_i)\ln \frac{1-y_i}{1-\widehat{\mu}_i})]^{1/2}$
gama	$\frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i}$	$\frac{3(y_i^{1/3} - \widehat{\mu}_i^{1/3})}{\widehat{\mu}_i^{1/3}}$	$\delta_i[2(\ln \frac{\widehat{\mu}_i}{y_i} + \frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i})]^{1/2}$
gaussiana inversa	$\frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i^{3/2}}$	$\widehat{\mu}_i^{-1/2} \ln \frac{y_i}{\widehat{\mu}_i}$	$\frac{y_i - \widehat{\mu}_i}{y_i^{1/2}\widehat{\mu}_i}$

$\delta_i = sinal(y_i - \widehat{\mu}_i)$ e $A(x) = \int [x(1-x)]^{-1/3} dx$.

3.3.3 Análise informal dos resíduos

Na adaptação de um modelo podemos encontrar anomalias, tanto na componente aleatória do modelo, como na componente sistemática, as quais podem ser detectadas através de uma análise informal dos resíduos, usando representações gráficas adequadas. Essas representações gráficas variam consoante a natureza das anomalias que se pretende detectar. As ideias aqui expostas são essencialmente baseadas no capítulo 12 de McCullagh and Nelder (1989).

- Uma primeira representação gráfica pode ser dos resíduos contra $\widehat{\eta}$ ou alguma transformação adequada do valor predito $\widehat{\mu}$. Tanto McCullagh and Nelder (1989) como Cordeiro (1986), sugerem usar os desvios residuais $R^{\star D}$ em vez dos resíduos de Pearson, já que estes, embora apresentem propriedades de $2^{\underline{a}}$ ordem razoáveis, podem ter distribuições bem diferentes da normal; as transformações de $\widehat{\mu}$ sugeridas por McCullagh and Nelder (1989) são

  i) $\widehat{\mu}$ para o modelo normal;

  ii) $2\sqrt{\widehat{\mu}}$ para o modelo de Poisson;

  iii) $2\,\mathrm{sin}^{-1}\sqrt{\widehat{\mu}}$ para o modelo binomial;

  iv) $2\ln\widehat{\mu}$ para o modelo gama;

  v) $-2\widehat{\mu}^{-1/2}$ para o modelo gaussiano inverso.

  No caso de não haver anomalias, os resíduos devem estar distribuídos em torno de zero com uma amplitude constante para diferentes valores de $\widehat{\mu}$. Anomalias tais como - (i) escolha errada da função de ligação; (ii) escolha errada da escala de uma ou mais covariáveis; (iii) omissão de um termo quadrático

numa covariável - podem ser detectadas através de uma curvatura no gráfico. Técnicas de alisamento podem ser úteis na avaliação da existência ou não de curvatura.

- Os resíduos também podem ser representados graficamente contra uma variável explicativa presente no preditor linear; novamente a existência de qualquer tendência no gráfico pode indicar uma escolha errada da função de ligação, ou uma escolha errada da escala da covariável em questão.

- Para detectar uma falsa distribuição populacional para a resposta Y, costuma usar-se uma representação gráfica dos resíduos ordenados contra pontos percentuais da distribuição de probabilidade de referência $F(\cdot)$; esses pontos podem ser definidos por

$$F^{-1}\Big(\frac{i-\alpha}{n-2\alpha+1}\Big) \qquad 0 \leq \alpha \leq 0.5.$$

- A existência de observações dependentes ou exibindo correlação serial, pode ser detectada através da representação gráfica dos resíduos $R_i^{\star D}$ contra i.

- Avaliação da função de variância

  Para avaliar da adequabilidade da função de variância escolhida, pode fazer-se uma representação gráfica dos resíduos absolutos contra os valores preditos (ou transformações adequadas desses valores preditos, como já foi referido); uma função de variância mal escolhida dá origem a uma tendência no gráfico. Uma tendência positiva indica que a função de variância cresce muito lentamente com a média; por exemplo, uma função de variância do tipo $V(\mu) \propto \mu$ deve ser substituída por $V(\mu) \propto \mu^{\lambda}, \lambda > 1$, com λ a escolher. Por outro

lado uma tendência negativa indica a situação contrária, isto é, a variância cresce demasiadamente rápido em relação à média e, portanto, uma função de variância do tipo, *e.g.*, $V(\mu) \propto \mu$, deve ser substituída por uma função de variância do tipo $V(\mu) \propto \mu^{\lambda}, \lambda < 1$.

É possível fazer uma avaliação formal da função de variância e estimar, *e.g.*, a quantidade λ adequada. Para tal veja-se, *e.g.*, a secção 4.3 de Fahrmeir and Tutz (1994), ou a secção 12.6.2 de McCullagh and Nelder (1989).

- Avaliação da função de ligação

  Para uma avaliação informal da adequabilidade da função de ligação, pode fazer-se uma representação gráfica da variável dependente ajustada definida em (2.13)

  $$\mathbf{u} = \widehat{\boldsymbol{\eta}} + \widehat{D}(\mathbf{y} - \widehat{\boldsymbol{\mu}}),$$

  contra $\widehat{\boldsymbol{\eta}}$, onde $\widehat{D}$ significa que a matriz diagonal D com elemento genérico $\frac{\partial \eta_i}{\partial \mu_i}$ é calculada para os valores estimados. Se os pontos se distribuírem, aproximadamente, sobre uma linha recta, então a função de ligação é adequada; se, por outro lado, se observar uma curvatura para cima, isso significa que é necessário usar uma função de ligação com potência superior; uma curvatura para baixo é sinónimo da situação contrária.

  Existem métodos formais para a avaliação da adequabilidade da função de ligação. Hinkley (1985) sugere considerar $\widehat{\eta}^2$ como uma nova covariável a adicionar ao preditor linear e verificar se há um declínio da função desvio. Veja-se ainda a secção 12.6.3 de McCullagh and Nelder (1989).

- Averiguação da adequabilidade da escala em que as covariáveis estão representadas

Uma má escolha da escala em que uma ou mais covariáveis estão representadas pode afectar a adequabilidade do modelo de modo a parecer, erradamente, que outro tipo de anomalias estão presentes, tal como, por exemplo, uma má escolha da função de ligação; é pois importante saber distinguir em que situação nos encontramos.

O objectivo, neste caso, é o de averiguar se um termo no preditor linear, do tipo, *e.g.*, βx, deve ser substituído por um termo do tipo $\beta h(x, \theta)$, onde $h(\cdot, \theta)$ é uma transformação apropriada da covariável em questão (do tipo, por exemplo, Box-Cox);

Uma representação gráfica adequada é a dos *resíduos parciais* $R_{parciais,i}$ contra os valores observados da covariável x_i, sendo o vector de *resíduos parciais* definidos por

$$\mathbf{R}_{parcial} = \mathbf{u} - \widehat{\boldsymbol{\eta}} + \widehat{\gamma}\mathbf{x},$$

onde $\mathbf{u} - \widehat{\boldsymbol{\eta}}$ é o vector dos resíduos medidos na escala linear, $\mathbf{u}$ é a variável dependente ajustada já definida e $\widehat{\gamma}$ é a estimativa do parâmetro para a variável explicativa em consideração.

Se a escala de x é satisfatória o gráfico deve ser aproximadamente linear.

- Avaliação da omissão de uma covariável

  Para averiguar se uma covariável que se omitiu, digamos z^*, deve ou não ser incluída no modelo, deve fazer-se uma representação gráfica dos *resíduos aumentados* contra z^*. Esses são definidos do modo que passamos a descrever.[8]

[8] Baseado na secção 7.6 de Cordeiro (1986).

Suponhamos que a componente sistemática correcta deve conter uma covariável adicional, isto é,

$$g(\boldsymbol{\mu}) = Z\boldsymbol{\beta} + h(\mathbf{z}^*; \gamma),$$

onde $h(\cdot; \gamma)$ pode representar

(i) um termo adicional em uma ou mais covariáveis originais, *e.g.*, um termo quadrático ou uma interacção;

(ii) uma contribuição linear ou não linear de alguma covariável omitida, *e.g.*, $h(z^*; \gamma) = z^*\gamma$ ou $h(z^*; \gamma) = \frac{\gamma}{z^*}$.

O objectivo é definir resíduos $\widetilde{R}$ para o modelo ajustado $\boldsymbol{\eta} = Z\boldsymbol{\beta}$ tal que

$$E(\widetilde{R}) = h(z^*; \gamma).$$

Se isto acontecer, um gráfico de $\widetilde{R}_i$ contra z_i^* exibirá, desprezando a variação aleatória, a função $h(z^*; \gamma)$.

De um modo semelhante ao que se fez para definir os resíduos parciais, os *resíduos aumentados* são obtidos através do acréscimo $[Z(Z^T\widehat{W}Z)^{-1}Z^T\widehat{W}]h(\mathbf{z}^*; \widehat{\gamma})$ aos resíduos medidos na escala linear $\mathbf{R} = \mathbf{u} - \widehat{\boldsymbol{\eta}} = [I - Z(Z^T\widehat{W}Z)^{-1}Z^T\widehat{W}]\mathbf{u}$.

Assim o vector dos *resíduos aumentados* é dado por

$$\widetilde{\mathbf{R}} = [I - Z(Z^T\widehat{W}Z)^{-1}Z^T\widehat{W}]\mathbf{u} + Z(Z^T\widehat{W}Z)^{-1}Z^T\widehat{W}]h(\mathbf{z}^*; \widehat{\gamma}).$$

A análise gráfica dos resíduos aumentados pode ser bastante útil na selecção de covariáveis, quando se tem muitas covariáveis a considerar. A formação da componente sistemática pode ser feita, passo a passo, com a introdução de uma única covariável de cada vez pelo método descrito.

3.4 Observações Discordantes

Na secção anterior estudámos como averiguar, usando informalmente os resíduos, a existência de desvios sistemáticos do modelo. Nesta secção iremos estudar como se pode averiguar da existência de desvios isolados do modelo, isto é, da existência de uma ou mais observações mal ajustadas pelo modelo, não seguindo o padrão das restantes observações; iremos designar, genericamente, essas observações por *observações discordantes*.

Na análise destes desvios isolados há, essencialmente, três noções importantes a considerar:

- *repercussão* ("*leverage*") - A repercussão mede o efeito que a observação tem nos valores preditos, sendo um indicativo de quão influente uma observação é.

- *influência* - Uma observação é influente se, uma sua ligeira modificação, ou exclusão do modelo, produz alterações significativas nas estimativas dos parâmetros do modelo. A sua presença pode, por isso, originar um impacto indevido nas conclusões a retirar do modelo. Observações influentes não têm, necessariamente, resíduos elevados.

- *consistência* - Uma observação com um resíduo elevado é em geral uma observação inconsistente. Esta inconsistência pode ser devida a um valor extremo da variável resposta ou (e) de uma ou mais covariáveis. Uma observação consistente deve seguir a tendência sugerida pelas restantes observações. Pode haver, no entanto, observações consistentes com repercussões elevadas.

Uma observação inconsistente (*outlier*) não é necessariamente uma observação influente.

3.4.1 Medida de repercussão

A definição geral da repercussão da j-ésima observação no valor predito da i-ésima resposta é a amplitude da derivada do i-ésimo valor predito $\widehat{\mu}_i$ em relação ao valor observado da j-ésima resposta, y_j. No caso dos MLG, esta medida é dada pelo ij-ésimo elemento da matriz de projecção generalizada

$$H = W^{1/2} Z (Z^T W Z)^{-1} Z^T W^{1/2},$$

definida em (3.3). Pode mostrar-se que

$$V^{-1/2}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \approx H V^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}),$$

onde $V = diag(V(\mu_i))$. Deste modo H mede a influência que $\mathbf{Y}$ tem em $\boldsymbol{\mu}$. Assim, uma medida do efeito de repercussão da i-ésima observação na determinação de $\widehat{\mu}_i$ é dada por h_{ii}, isto é, pelo i-ésimo elemento da diagonal principal de H. Dado que

$$tra(H) = \sum_{i=1}^{n} h_{ii} = p,$$

digamos que, em média, cada valor h_{ii} deve estar próximo de p/n. Pode pois considerar-se que um ponto tem repercussão elevada se $h_{ii} > \frac{2p}{n}$, ou, equivalentemente, se

$$h_{ii}^{\star} = \frac{n h_{ii}}{p} > 2.$$

Esta matriz, contrariamente ao que acontece com a matriz de projecção para o modelo linear normal, depende não só das covariáveis

através de Z, como também do modelo adaptado através de W. Assim, uma observação extrema, isto é, com um valor elevado para uma ou mais das covariáveis, não tem necessariamente uma repercussão elevada se o seu peso (elemento correspondente de W) for pequeno.

Gráficos de h_{ii} contra $\widehat{\mu}_i$, ou contra i, são geralmente úteis na identificação de pontos com repercussão elevada.

3.4.2 Medida de influência

Um indicador da influência da i-ésima observação $(y_i, \mathbf{z}_i)$ no vector estimado $\widehat{\boldsymbol{\beta}}$, pode ser calculado pela diferença $\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}$, onde $\widehat{\boldsymbol{\beta}}_{(i)}$ e $\widehat{\boldsymbol{\beta}}$ representam, respectivamente, as estimativas de máxima verosimilhança do vector parâmetro $\boldsymbol{\beta}$ obtidas da amostra sem a observação $(y_i, \mathbf{z}_i)$ e da amostra com todas as observações. Se $\widehat{\boldsymbol{\beta}}_{(i)}$ for substancialmente diferente $\widehat{\boldsymbol{\beta}}$, então a observação $(y_i, \mathbf{z}_i)$ pode ser considerada influente.

No caso dos modelos lineares normais a medida de influência utilizada é a sugerida por Cook (1977) $D_i = (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})^T (Z^T Z)(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})/ps^2$.

Dado que, agora $cov(\widehat{\boldsymbol{\beta}}) = (Z^T W Z)^{-1}$, tal como se viu em 2.2.3, é natural considerar como generalização da medida de influência de Cook

$$D_i = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})^T (Z^T W Z)(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})}{p\widehat{\phi}}. \tag{3.9}$$

Contudo, no caso dos MLG, a estimação de $\widehat{\boldsymbol{\beta}}_{(i)}$ necessita do recurso a métodos iterativos. O processo é, pois, computacionalmente caro para poder ser feito para todas as observações; alternativamente pode obter-se uma aproximação a $\widehat{\boldsymbol{\beta}}_{(i)}$ fazendo apenas o 1º passo do

processo iterativo, usando $\widehat{\boldsymbol{\beta}}$ como valor inicial; recorrendo a (2.14), a estimativa a um passo é dada por

$$\widehat{\boldsymbol{\beta}}_{(i),1} = \mathcal{I}_{(i)}^{-1}(\widehat{\boldsymbol{\beta}}) Z_{(i)}^T W_{(i)}(\widehat{\boldsymbol{\beta}}) \mathbf{u}(\widehat{\boldsymbol{\beta}}), \tag{3.10}$$

onde o índice (i) refere que os cálculos são feitos sem a i-ésima observação. Para o cálculo aproximado de D_i usa-se esta aproximação de $\widehat{\boldsymbol{\beta}}_{(i)}$ em (3.9).

Uma fórmula mais simples para $\widehat{\boldsymbol{\beta}}_{(i),1}$ é dada por Williams (1987), nomeadamente

$$\widehat{\boldsymbol{\beta}}_{(i),1} = \widehat{\boldsymbol{\beta}} - \varpi_i^{1/2}(1-h_{ii})^{-1/2} R^{\star P}_i (Z^T W Z)^{-1} \mathbf{z}_i, \tag{3.11}$$

onde ϖ_i, elemento de W está definido em (2.11).

Aplicações do modelo linear generalizado mostram que a utilização de $\widehat{\boldsymbol{\beta}}_{(i),1}$ em (3.9) em vez do verdadeiro valor $\widehat{\boldsymbol{\beta}}_{(i)}$ subestima o valor de D_i. Contudo, segundo Williams (1987), o ponto importante é que a aproximação considerada geralmente identifica os casos mais influentes. Após serem identificados estes casos, pode então obter-se exactamente a influência da sua omissão.

3.4.3 Medida de consistência

A possibilidade de uma determinada observação $(y_i, \mathbf{z}_i)$ ser inconsistente pode também ser averiguada adaptando o modelo sem essa observação e calculando os resíduos da observação eliminada em relação ao correspondente valor predito $\widehat{\mu}_{(i)} = h(\mathbf{z}_i^T \widehat{\boldsymbol{\beta}}_{(i)})$. Os resíduos assim obtidos são chamados *resíduos de eliminação* ("deletion residuals"). Por exemplo, a correspondente expressão para os resíduos de eliminação de Pearson é

$$R^{\star P}_{(i)} = \frac{(y_i - \widehat{\mu}_{(i)}) w_i}{[\widehat{\phi} V(\widehat{\mu}_{(i)})(1+h_{(ii)})]^{1/2}},$$

onde $h_{(ii)} = \mathbf{z}_i^T (Z_{(i)}^T W_{(i)} Z_{(i)})^{-1} \mathbf{z}_i$ e o sinal + no denominador de $R^{\star P}_{(i)}$ aparece devido ao facto de y_i e $\widehat{\mu}_{(i)}$ serem agora independentes.

Novamente, o cálculo exacto destes resíduos de eliminação para todas as observações é computacionalmente dispendioso e é necessário encontrar outra alternativa. Em geral a solução encontrada é o cálculo desses resíduos usando as estimativas obtidas após o 1º passo do processo iterativo, tal como se referiu na secção anterior.

Williams (1987) sugere a utilização de outro tipo de resíduos que ele denomina por *resíduo de verosimilhança*. Para o efeito, seja G_i a redução operada no desvio reduzido quando se elimina do modelo a i-ésima observação, *i.e.*

$$\begin{aligned} G_i &= D^*(\mathbf{y}; \widehat{\boldsymbol{\mu}}) - D^*(\mathbf{y}_{(i)}, \widehat{\boldsymbol{\mu}}_{(i)}) \\ &= \phi^{-1}[d_i + \sum_{j \neq i} d_j - \sum_{j \neq i} d_{(i),j}], \end{aligned}$$

onde, $\mathbf{y}_{(i)}$ designa o vector $\mathbf{y}$ sem o elemento y_i e d_j foi definido na secção 3.1.1.

Assim, a contribuição de y_i para G_i é exactamente $\phi^{-1} d_i$. Por outro lado, Williams (1987) mostra que, usando a aproximação de $\widehat{\boldsymbol{\beta}}_{(i)}$ dada em (3.11), e fazendo um desenvolvimento em série de Taylor da função desvio, o decréscimo de $\sum_{j \neq i} d_j$ quando $\widehat{\boldsymbol{\mu}}$ é substituído por $\widehat{\boldsymbol{\mu}}_{(i)}$ é aproximadamente $\phi\, h_{ii} (R^{\star P}_i)^2$. Deste modo G_i pode ser aproximado por R^2_{Gi} definido por

$$R^2_{Gi} = \phi^{-1} d_i + h_{ii} (R^{\star P}_i)^2 = (1 - h_{ii})(R^{\star D}_i)^2 + h_{ii} (R^{\star P}_i)^2 \qquad (3.12)$$

O *resíduo de verosimilhança* é então dado por

$$R^*_{Gi} = \delta_i \sqrt{(1 - h_{ii})(R^{\star D}_i)^2 + h_{ii} (R^{\star P}_i)^2},$$

onde $\delta_i = sinal(y_i - \widehat{\mu}_i)$.

Observações com valores elevados de R^*_{Gi} podem ser consideradas observações inconsistentes. O valor R^*_{Gi} é intermédio entre ${R^\star}^D_i$ e ${R^\star}^P_i$, estando em geral mais próximo de ${R^\star}^D_i$ porque o valor esperado de h_{ii} dado por p/n é em geral pequeno.

Williams (1987) sugere representar graficamente R^*_{Gi} contra i, h_{ii} ou $\widehat{\eta}_i$, para estudar as observações quanto à sua consistência. Sugere ainda usar $\max R^2_{Gi}$ como estatística para testar se a observação correspondente é um *outlier*.

Capítulo 4

Aplicações I: Modelos Discretos

A aplicação dos modelos lineares generalizados tem-se verificado em diferentes áreas científicas, sobretudo nas ciências biomédicas, agronomia e ciências sociais. A partir deste capítulo ilustraremos os MLG em diversas situações caracterizadas pela natureza dos dados ou pelo objectivo da analise estatística. Os MLG com variável resposta discreta e contínua são aqui chamados de **modelos discretos** e de **modelos contínuos**, respectivamente. Neste capítulo analisaremos alguns modelos discretos, enquanto exemplos de modelos contínuos serão estudados no capítulo seguinte.

A ilustração dos modelos lineares generalizados discretos para dados binários ou na forma de contagens faz-se com os modelos de regressão logística, *probit*, complementar log-log e log-lineares, incluindo exemplos com dados agrupados. A secção 4.1 apresenta modelos de regressão logística num estudo retrospectivo de um processo infeccioso pulmonar com pacientes diagnosticados com tipo malig-

no ou benigno. Na secção 4.2 encontra-se um exemplo de dados na forma de proporção com a adopção de três MLG: logístico, *probit* e complementar log-log. Na última secção analisa-se um conjunto de dados em forma de contagens através de modelos log-lineares que desempenham um papel importante na análise de dados categorizados.

4.1 Modelos de Regressão Logística

Na subsecção 1.4.2 apresentámos potenciais modelos lineares generalizados para analisar dados binários, sendo o modelo logístico o mais popular desses modelos, provavelmente, devido à simplicidade da sua implementação computacional. A adopção do modelo estrutural (1.11) para a probabilidade de sucesso caracteriza o modelo de regressão logística.

Exemplo 4.1 Processo Infeccioso Pulmonar

No sector de Anatomia e Patologia do Hospital Heliópolis (São Paulo/Brasil) realizou-se um estudo retrospectivo com 175 pacientes entre 1970 e 1982, cujos dados se encontram em Paula et al. (1984). O objectivo principal desse estudo era avaliar a associação entre algumas variáveis histológicas e o tipo, maligno ou benigno, do Processo Infeccioso Pulmonar (PIP).

Nesse estudo de caso-controle, os casos foram todos os pacientes diagnosticados, no período e hospital há pouco mencionados, como portadores do PIP de origem maligna (71 pacientes). Os controles foram formados por uma amostra de 104 pacientes de uma população de 270, os quais foram também diagnosticados na mesma época e local e tiveram confirmado o PIP de origem benigna.

A observação de cada um dos pacientes fez-se através de variáveis histológicas nos fragmentos de tecidos retirados da região pulmonar. Dessas variáveis somente as intensidades de histiócitos-linfócitos (HL) e de fibrose-frouxa (FF) foram consideradas importantes na discriminação dos dois tipos de PIP. Porém, o conjunto de covariáveis será formado por dois factores potenciais de confundimento, sexo e idade. A descrição da codificação destas variáveis encontra-se na tabela 4.1.

Tabela 4.1: Variáveis do processo infeccioso pulmonar.

variáveis	codificação
tipo de PIP (Y)	1=maligno
	0=benigno
idade (x_1)	em anos
sexo (x_2)	1=masculino
	0=feminino
intensidade de histiócitos-linfócitos	1=alta(3,4)
HL (x_3)	0=baixa(1,2)
intensidade de fibrose frouxa	1=alta(3,4)
FF (x_4)	0=baixa(1,2)

Fonte: Paula et al. (1984).

Note-se que os dados sobre a variável resposta binária Y e o vector de covariáveis $\mathbf{x} = (x_1, \ldots, x_4)^T$ não foram obtidos prospectivamente. Isto é, os dados sobre os casos e controles resultam de uma amostragem directa de um modelo para $P(\mathbf{x} \mid Y)$, $Y = 0, 1$, contrariamente aos dados prospectivos que estão associados ao modelo $\pi(\mathbf{x}) = P(Y \mid \mathbf{x})$. Entretanto, a sua análise pode ser processada de

modo análogo àquele previsto para os dados de um estudo prospectivo, visto que o uso de um modelo prospectivo logístico revela-se conveniente pelo facto da metodologia de máxima verosimilhança para os dados retrospectivos envolver ainda um modelo numa forma logística. Para maiores detalhes, veja Silva (1992, sec. 2.8).

De acordo com a característica deste estudo retrospectivo, o modelo de regressão logístico (1.11) será ajustado a estes dados, *i.e.*, a probabilidade de tipo maligno do PIP para o i-ésimo paciente (π_i) está relacionada com o seu vector de covariáveis $\mathbf{z}_i = (1, x_{i1}, \ldots, x_{i4})^T$ através de

$$\pi_i = \frac{\exp(\mathbf{z}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{z}_i^T \boldsymbol{\beta})}, \tag{4.1}$$

onde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_4)^T$ e $i = 1, \ldots, 175$. Os parâmetros de regressão β_j, $j \neq 0$, são estimados directamente deste modelo, enquanto β_0 pode ser estimado posteriormente, se as probabilidades de selecção amostral dos tipos de PIP, ϕ_1 e ϕ_0, forem conhecidas.

4.1.1 Selecção do modelo logístico

Para a selecção de covariáveis que formem o "melhor" modelo logístico usaremos um método de selecção *stepwise* baseados em *p-values* relativos aos testes de razão de verosimilhanças de Wilks entre modelos com inclusão ou exclusão de covariáveis, ou mesmo de suas interacções. Neste caso, o grau de importância de uma covariável é medido pelo *p-value* do teste da razão de verosimilhanças entre os modelos que a incluem e a excluem. Quanto menor for este valor tanto mais importante será considerada a covariável. Como a covariável mais importante por este critério não é necessariamente significativa do ponto de vista estatístico, há que impor um limite

superior P_E para estes *p-values*, a fim de atrair candidatos importantes em princípio à entrada.

Dado que a presença de várias covariáveis num modelo pode tornar uma ou outra dispensáveis, faremos a verificação da importância da presença de cada covariável confrontando o seu respectivo *p-value* com um limite inferior P_S, superior a P_E. As covariáveis com um *p-value* associado superior a P_S serão assim candidatas à remoção. Na primeira etapa selecciona-se os efeitos principais das covariáveis e nas etapas seguintes as interacções de $1^{\underline{a}}$ ordem, $2^{\underline{a}}$ ordem, e assim sucessivamente, associadas às covariáveis presentes no modelo obtido na primeira etapa.

A primeira etapa começa com o ajustamento do modelo só com a ordenada na origem (modelo nulo) e é constituída pelos seguintes passos (Hosmer and Lemeshow, 1989, cap. 3):

1. Construímos testes da razão de verosimilhanças entre o modelo inicial e os modelos logísticos simples formados com cada uma das covariáveis do estudo. O mínimo dos *p-values* associados a cada teste será comparado com o *p-value* de entrada P_E. Se P_E for maior incluímos a covariável referente àquele valor mínimo e passamos ao passo seguinte; caso contrário, paramos a selecção e seleccionamos o último modelo;

2. Partindo do modelo incluindo a covariável seleccionada no passo anterior, introduzimos individualmente as demais covariáveis. Cada um destes modelos com duas covariáveis é testado contra o modelo inicial deste passo. Novamente o mínimo dos *p-values*, se for menor do que P_E, implica a inclusão no modelo da sua respectiva covariável, e a passagem ao passo seguinte. Caso contrário, paramos com a selecção;

3. Comparamos o ajuste do modelo logístico contendo as covariáveis seleccionadas nos passos anteriores com os modelos que dele resultam por exclusão individual de cada uma das covariáveis. Se o máximo dos *p-values* destes testes da razão de verosimilhanças for menor do que P_S, a covariável associada a este nível permanece no modelo. Caso contrário, ela é removida. Em qualquer circunstância, o algoritmo segue para o passo seguinte.

4. O modelo resultante do passo anterior será ajustado, e antes de tornar-se o modelo inicial da etapa 2 (selecção de interacções de primeira ordem das covariáveis incluídas), repetiremos os passos anteriores quantas vezes forem necessárias até termos a indicação de parada nestes passos ou todas as covariáveis inclusas no modelo.

Uma vez seleccionadas as covariáveis "importantes", i.e., os seus efeitos principais na etapa 1, os passos anteriores são repetidos com o objectivo de seleccionar as interacções que envolvem aquelas covariáveis.

Note-se que este procedimento exige o cálculo das estimativas de máxima verosimilhança em cada passo, o que encarece o trabalho computacional, particularmente em grandes amostras. Apesar desta desvantagem este procedimento será utilizado na selecção das covariáveis do exemplo 4.1.

Na primeira etapa da selecção dos efeitos principais do modelo (4.1), o valor observado da estatística do teste da razão de verosimilhanças (2.20) que compara o modelo nulo (modelo inicial) com o modelo com inclusão da covariável idade é $\Lambda = 236.34 - 190.92 = 45.42$. O uso da distribuição qui-quadrado com 1 grau de liberdade

produz o *p-value* de 0.000. Este valor faz parte da tabela 4.2, bem como os outros *p-values* que formam os cinco passos da etapa 1.

Baseando-se nos *p-values* da tabela 4.2 pode-se encontrar quais as covariáveis a incluir ou excluir em cada passo de decisão da etapa 1 do método de selecção. O passo 1 inclui a covariável idade, pois o seu *p-value*, que é o mínimo neste passo, é inferior a $P_E = 0.20$ (valor padrão para inclusão de variáveis). O passo seguinte nesta etapa inclui a variável HL, e agora com duas variáveis incluídas no modelo serão testadas as exclusões individuais destas covariáveis. Os *p-values* associados a esses testes encontram-se na linha de referência do passo 3 e abaixo da curva em forma de escada da tabela 4.2. O máximo desses valores estará identificado por um asterisco e, sendo inferior a $P_S = 0.25$ (valor padrão de exclusão), a variável associada a este *p-value* não é retirada do modelo. Seguindo esta lógica, encontramos os *p-values* mínimos em cada passo de decisão como o primeiro elemento acima da curva em "escada". Sendo todos inferiores a P_E concluímos pela entrada no modelo de todas covariáveis. Relativamente às exclusões observamos que os *p-values* com asterisco são inferiores a P_S, e assim nenhuma das covariáveis sai do modelo. Em resumo, o modelo resultante da etapa 1 com este procedimento de selecção é o modelo com todos os efeitos principais do conjunto das covariáveis.

De forma análoga processar-se-á a etapa 2, cujos *p-values* para tomada de decisão em cada passo encontram-se na tabela 4.3. Concluímos então que só três interacções de primeira ordem serão incluídas no modelo, e nenhuma delas foi excluída posteriormente. Essas interacções são idade.HL, HL.FF e sexo.FF.

Na etapa 3 nenhuma interacção de segunda ordem foi incluída, já que o mínimo dos *p-values* dos testes de inclusão não foi inferior

Tabela 4.2: *P-values* da etapa 1 do método de selecção.

passo de decisão	idade	HL	sexo	FF
1	0.000	0.000	0.288	0.001
2	0.000	0.000	0.100	0.002
3	0.000	0.000*	0.043	0.082
4	0.000	0.000	0.053*	0.123
5	0.000	0.000	0.063	0.123*

Tabela 4.3: *P-values* da etapa 2 do método de selecção.

passo de decisão	id.HL	HL.FF	sex.FF	id.FF	id.sex	HL.sex
1	0.015	0.020	0.082	0.065	0.655	0.084
2	0.015	0.038	0.082	0.243	0.222	0.128
3	0.028	0.038*	0.017	0.254	0.275	0.207
4	0.033*	0.008	0.017	0.230	0.417	0.806

a P_E. Assim, o modelo resultante da selecção *stepwise* acima possui todos os efeitos principais do conjunto de covariáveis e as interacções de primeira ordem idade.HL, sexo.FF e HL.FF.

4.1.2 Avaliação e interpretação do modelo seleccionado

De acordo com a subsecção 4.1.1, o modelo de regressão logístico seleccionado no procedimento adoptado para a selecção de cova-

riáveis ou interacções de covariáveis tem a seguinte forma estrutural

$$\ln\Big[\frac{\pi_i}{1-\pi_i}\Big] = \beta_0 + \sum_{j=1}^{4} x_{ij}\beta_j + x_{i1}x_{i3}\beta_5 + x_{i2}x_{i4}\beta_6 + x_{i3}x_{i4}\beta_7. \quad (4.2)$$

As estimativas dos parâmetros de regressão e dos respectivos desvios padrões assintóticos do modelo logístico (4.2) encontram-se na tabela 4.4.

Tabela 4.4: Estimativas dos parâmetros e desvios padrões associados ao modelo logístico 4.2.

efeito	parâmetro	estimativa	desvio padrão
constante	β_0	-0.033	1.0280
idade	β_1	0,039	0.0173
sexo	β_2	-1.387	0.5826
HL	β_3	-5.430	1.6724
FF	β_4	-5.197	1.6896
idade.HL	β_5	0.060	0.0292
sexo.FF	β_6	3.188	1.4720
HL.FF	β_7	2.801	1.1120

O teste de ajustamento do modelo logístico (4.2) produz para a função desvio o valor 145.45 correspondente a um *p-value* 0.8844 (calculado de uma distribuição qui-quadrado com 167 graus de liberdade) e, portanto, conclui-se que há adequação do modelo. Outra avaliação do modelo em causa faz-se através do estudo dos resíduos, *e.g.*, calculando os desvios residuais (3.7) para os 175 indivíduos. Neste caso foram encontrados três pacientes com valores considerados aberrantes: $R^{\star D}_{9} = -2.33$, $R^{\star D}_{92} = -2.09$ e $R^{\star D}_{117} = 2.23$. Uma representação gráfica desses resíduos encontra-se na figura 4.1.

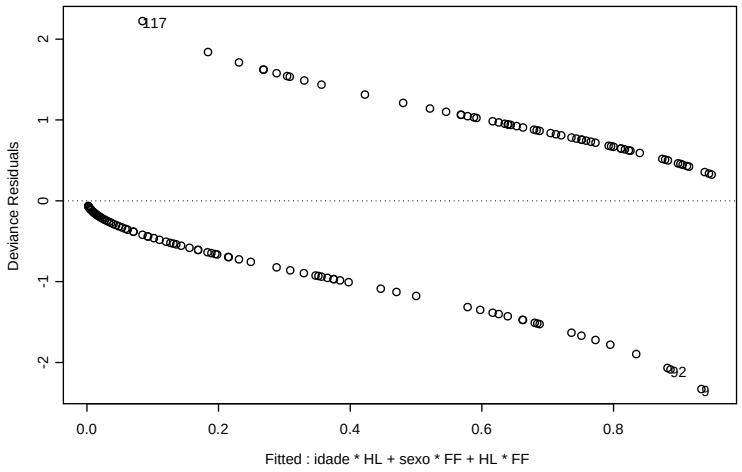

Figura 4.1: Gráfico dos desvios residuais × valores ajustados.

Como o interesse principal é estudar associação entre o tipo de PIP e as variáveis histológicas no conjunto de covariáveis, formaremos as razões de chances para os níveis dessas variáveis. Por exemplo, a razão de chances[9] de um paciente com nível alto de intensidade de histiócitos-linfócitos (HL), em relação ao nível baixo de intensidade HL, estar com PIP do tipo maligno, denotada aqui por ψ_{HL}. Supondo que os pacientes sejam do mesmo sexo (x_2) e tenham a mesma idade (x_1) e o mesmo nível de intensidade de fibrose frouxa (x_4), a razão de chances em causa pode ser estimada, de acordo com o modelo (4.2), por

$$\widehat{\psi}_{HL} \;=\; \exp\{-5.43 + 0.062\, x_1 + 2.801\, x_4\}\ . \qquad (4.3)$$

Da expressão (4.3) podemos concluir que a chance de um PIP

[9]Seguindo a sugestão de Paulino e Singer (1997) traduzimos o termo "*odds ratio*" por razão de chances.

maligno é menor para os pacientes com alta intensidade HL que para os pacientes com baixa intensidade HL, isto no nível de baixa intensidade de fibrose frouxa (FF) e no intervalo de variação amostral da idade (15-87 anos). Já na categoria alta de intensidade FF, $\widehat{\psi}_{HL}$ torna-se maior do que a unidade após a idade de 42 anos (aproximadamente), pelo que a afirmação anterior é válida se substituirmos menor por maior. Em ambos os níveis de intensidade de fibrose frouxa a razão de chances referida cresce com o aumento da idade.

Para ilustramos a aplicação da expressão (4.3), suponhamos que dois pacientes de 60 anos e do mesmo sexo tenham sido submetidos a exames no hospital referido a fim de ser diagnosticado o tipo de PIP. Após os exames, admitamos que se constatou para ambos o nível baixo de intensidade FF, enquanto apenas um apresentou alta intensidade HL. Deste modo, a chance estimada do paciente, cujo exame não detectou alta intensidade HL, estar com PIP maligno, em relação ao outro, é $\widehat{\psi}_{HL}^{-1} = [\exp(-1.71)]^{-1} = 5.5$.

Analogamente, seja ψ_{FF} a razão de chances de um paciente com alta intensidade FF estar com PIP do tipo maligno relativamente ao nível baixo desta intensidade. Supondo que os pacientes são semelhantes nas demais covariáveis e recordando que x_2 e x_3 são, respectivamente, sexo e intensidade HL, o parâmetro em causa pode ser estimado por

$$\widehat{\psi}_{FF} \;=\; \exp\{-5.197 + 3.188\, x_2 + 2.801\, x_3\}\ . \qquad (4.4)$$

Da estimativa (4.4) podemos deduzir que a chance de um PIP maligno é menor para os pacientes com alta intensidade FF que para os pacientes com baixa intensidade de fibrose frouxa, isto entre as mulheres independentemente do nível de intensidade HL e para

os homens com baixa intensidade HL. Para as mulheres com alta intensidade HL ocorre o contrário nesta chance. Em ambos os níveis de intensidade HL a razão de chances ψ_{FF} é maior para os homens do que para as mulheres.

Se houver interesse em prever $\pi(\mathbf{x})$, probabilidade de um paciente da população com uma determinada configuração estar com PIP do tipo maligno, deveremos antes estimar β_0 verdadeiramente, *i.e.*,

$$\widehat{\beta}_0^* = \widehat{\beta}_0 - \ln\Big(\frac{71/71}{104/270}\Big) = -0.033 - (0.954) = -0.987 \ .$$

Deste modo, ficamos aptos a estimar $\pi(\mathbf{x})$ para qualquer valor de $\mathbf{x}$, como se ilustra na tabela 4.5.

Tabela 4.5: Estimativas de $\pi(\mathbf{x})$ em algumas situações.

idade	sexo	intensidade HL	intensidade FF	$\pi(\mathbf{x})$
51	masculino	alto	alto	0.267
51	masculino	baixo	baixo	0.639
51	feminino	baixo	baixo	0.876
20	feminino	baixo	baixo	0.678
45	masculino	alto	baixo	0.083
50	feminino	alto	baixo	0.374
60	masculino	baixo	alto	0.252

4.2 Modelos de Dose-resposta

Os modelos lineares generalizados são frequentemente usados em toxicologia, onde se pretende frequentemente descrever o efeito de

um medicamento tóxico na morte dos indivíduos em estudo. Este caso envolve uma covariável contínua e uma variável resposta binária e a relação entre elas é frequentemente denominada de modelo de dose-resposta.

Nos modelos de dose-resposta a probabilidade de sucesso π está restrito ao intervalo $(0,1)$ para valores do preditor linear $\eta = \mathbf{z}^T\boldsymbol{\beta} = \beta_0 + \beta_1 x$ em $(-\infty, +\infty)$, sendo razoável modelarmos π como uma função de distribuição acumulada $F(\cdot)$, *i.e.*, $\pi = g^{-1}(\eta) = F(\eta)$, onde $g(\pi) = \eta$ é uma função de ligação. De acordo com a subsecção 1.4.2, se a função de distribuição for a logística, normal reduzida ou Gumbel, o modelo de dose-resposta será um modelo logístico, *probit* ou complementar log-log, respectivamente.

Os modelos de dose-resposta visam não só a predição da probabilidade de sucesso para uma dosagem específica ($\pi(x)$) mas também a determinação da dosagem necessária para se atingir uma probabilidade de sucesso P. Essa dosagem é chamada de **dose letal**. A notação escolhida para uma dose letal de $100P\%$ de sucesso é DL_{100P}, logo

$$P \;=\; F(\beta_0 + \beta_1 DL_{100P}), \quad 0 < P < 1. \tag{4.5}$$

A dose letal mais comum em toxicologia é a dose mediana (DL_{50}), embora em certos casos haja interesse em estimar as doses extremas, *e.g.*, DL_1 ou DL_{99}.

Para grandes amostras, pode-se construir um intervalo de confiança para a dose letal referida em (4.5) usando uma aproximação para a variância assintótica do seu estimador de máxima verosimilhança. Por exemplo, sob o modelo logístico, o estimador de máxima verosimilhança de DL_{100P} é, pela propriedade de invariância,

$$\widehat{DL}_{100P} \;=\; \frac{1}{\widehat{\beta}_1}\Big[\ln\Big(\frac{P}{1-P}\Big) - \widehat{\beta}_0\Big] \;\equiv\; \mathcal{D}(\widehat{\boldsymbol{\beta}}) \;, \tag{4.6}$$

onde $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)^T$ é o estimador de máxima verosimilhança do parâmetro de regressão do modelo. Logo, um intervalo de $100(1-\alpha)\%$ de confiança para DL_{100P}, em grandes amostras, é

$$\widehat{DL}_{100P} \stackrel{+}{-} z_{1-\alpha/2}\sqrt{\widehat{var}_A[\mathcal{D}(\widehat{\boldsymbol{\beta}})]}\ , \tag{4.7}$$

onde $\widehat{var}_A[\mathcal{D}(\widehat{\boldsymbol{\beta}})]$ é a variância de $\mathcal{D}(\widehat{\boldsymbol{\beta}})$, aproximada por

$$Var_A[\mathcal{D}(\widehat{\boldsymbol{\beta}})] = \mathcal{B}(\boldsymbol{\beta})^T[\mathcal{I}(\boldsymbol{\beta})]^{-1}\mathcal{B}(\boldsymbol{\beta})$$

e avaliada em $\widehat{\boldsymbol{\beta}}$, $\mathcal{B}(\boldsymbol{\beta}) \equiv \partial\mathcal{D}(\widehat{\boldsymbol{\beta}})/\partial\boldsymbol{\beta} = (\beta_1^{-1}, \{\beta_0 - \ln(P/(1-P))\}/\beta_2^2)^T$, $\mathcal{I}(\boldsymbol{\beta})$ é a matriz de informação de Fischer e $z_{1-\alpha/2}$ é o percentil $100(1-\alpha/2)\%$ da normal reduzida (Silva, 1992, cap. 2).

Exemplo 4.2 Mortalidade de besouros

Em Bliss (1935) encontra-se um estudo sobre o comportamento de besouros adultos à exposição ao gás carbono (CS_2) durante cinco horas. Foram observados 481 besouros divididos em 8 grupos, onde cada um deles recebeu uma dosagem distinta do gás. Posteriormente, anotou-se o total de besouros mortos em cada grupo de dosagem. A variável resposta é a proporção de besouros mortos na dosagem x e a covariável transformada é $x = \log_{10} CS_2(mg/litro)$. Um objectivo do estudo é estimar a curva de dose-resposta quanto à mortalidade de besouros a partir de diferentes dosagens.

Ajustando-se os modelos logístico, *probit* e complementar log-log aos dados do exemplo 4.2, as equações de regressão estimadas são, respectivamente,

$$\begin{aligned} \ln[\widehat{\pi}(x)/(1-\widehat{\pi}(x))] &= -60.459 + 34.121\,x, \\ \Phi^{-1}(\widehat{\pi}(x)) &= -34.803 + 19.652\,x, \\ \ln(-\ln(1-\widehat{\pi}(x))) &= -39.533 + 22.017\,x. \end{aligned}$$

Os valores ajustados por estes modelos apresentam-se na tabela 4.6. Observamos assim uma grande concordância entre os valores ajustados pelos modelos logístico e *probit*, contrariamente aos valores análogos obtidos com o modelo complementar log-log. Esta conclusão pode ser verificada também na figura 4.2.

Tabela 4.6: Mortalidade de besouros (Bliss, 1935).

dosagem	proporções de besouros mortos			
x	observada	logístico	*probit*	clog-log
1.6907	0.1017	0.0590	0.0573	0.0947
1.7242	0.2167	0.1642	0.1789	0.1877
1.7552	0.2903	0.3614	0.3782	0.3373
1.7842	0.5000	0.6035	0.6024	0.5412
1.8113	0.8254	0.7933	0.7859	0.7571
1.8369	0.8983	0.9019	0.9024	0.9168
1.8610	0.9839	0.9544	0.9615	0.9854
1.8839	1.0000	0.9786	0.9867	0.9991

Os testes de ajustamento dos modelos de dose-resposta logístico, *probit* e complementar log-log produziram para a estatística da função desvio os valores 11.23, 10.12 e 3.45, respectivamente, com os correspondentes *p-values* 0.0815, 0.1197 e 0.7506 comparados com uma qui-quadrado de 6 graus de liberdade. Os modelos logístico e *probit* não constituem, assim, um instrumento capaz de uma descrição satisfatória dos dados, o mesmo não acontece com o modelo complementar log-log, visto que o seu *p-value* indica um bom ajustamento do modelo.

Para os modelos de dose-resposta ajustados acima, a dosagem de

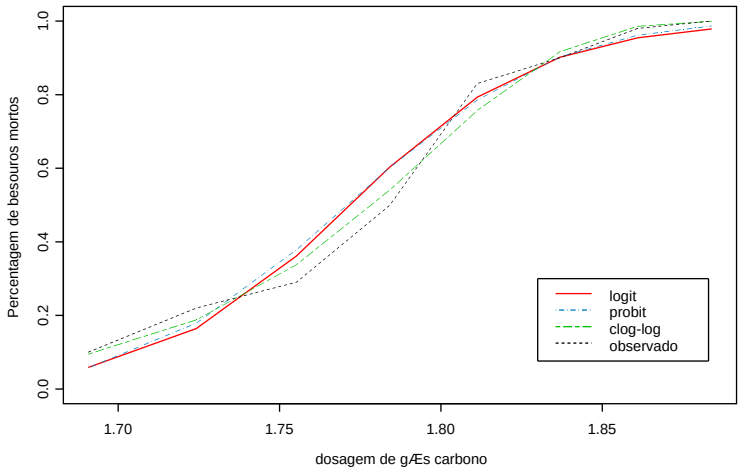

Figura 4.2: Gráfico dos valores ajustados nos 3 modelos.

gás carbono que mata $100P\%$ de besouros é estimada por

$$\widehat{DL}_{100P} = \frac{1}{34.121}\Big[\ln\Big(\frac{P}{1-P}\Big) + 60.459\Big]$$

para o modelo logístico,

$$\widehat{DL}_{100P} = \frac{1}{19.652}\Big[\Phi^{-1}(P) + 34.803\Big]$$

para o modelo *probit* e

$$\widehat{DL}_{100P} = \frac{1}{22.017}\Big[\ln(-\ln(1-P)) + 39.533\Big]$$

para o modelo complementar log-log. Estimativas de doses letais nos três modelos lineares generalizados em causa encontram-se na tabela 4.7. Novamente os modelos logístico e *probit* têm estimativas parecidas para essas doses letais, contrariamente ao modelo complementar log-log que apresenta estimativas quer inferiores nas

caudas da curva de dose-resposta quer superiores junto ao meio dessa curva. Obviamente que as doses letais estimadas com o modelo de dose-resposta complementar log-log são mais fiáveis, visto que este modelo foi o único a ter um bom ajustamento.

Tabela 4.7: Estimativas de doses letais.

dose letal	modelo		
	logístico	*probit*	clog-log
DL_{50}	1.771	1.771	1.779
DL_{99}	1.907	1.889	1.865
DL_1	1.637	1.652	1.587

4.3 Modelos Log-lineares

Os modelos log-lineares ou modelos de regressão de Poisson desempenham um papel importante na análise de dados categorizados equiparável ao dos modelos de regressão normal na análise de dados contínuos. Estes modelos são usados para analisar tabelas de contingência multidimensionais, mesmo quando o modelo probabílistico não é um produto de distribuições de Poisson (subsecção 1.4.3). Para maiores detalhes, veja Paulino e Singer (1997).

Exemplo 4.3 Infecções urinárias

Os dados da tabela 4.8 reportam-se a um estudo sobre infecções urinárias realizado com 468 pacientes (Koch et al., 1985, pg. 120). Os pacientes foram classificados quanto ao tipo de diagnóstico (complicado ou não), tratamento (A,B,C) e a cura da infecção (sim,não). Entre as questões de interesse, têm-se:

1. verificação de existência de associação entre tratamento e tipo de diagnóstico relativamente à cura da infecção;

2. comparação dos tratamentos para cada tipo de diagnóstico, visto que frequentemente as infecções com diagnóstico complicado são mais difíceis de curar.

Tabela 4.8: Dados do estudo de infecção urinária.

diagnóstico	tratamento	curados	não curados
complicado	A	78	20
	B	101	11
	C	68	46
não complicado	A	40	5
	B	54	5
	C	34	6

Fonte: Koch et al. (1985).

As variáveis tipo de diagnóstico, tratamento e cura da infecção do exemplo 4.3 são aqui denotadas por X_1, X_2 e X_3, respectivamente. Os níveis destas variáveis são $X_1 = 1$ (complicado), $X_1 = 2$ (não complicado), $X_2 = 1$ (A), $X_2 = 2$ (B), $X_2 = 3$ (C), $X_3 = 1$ (curado) e $X_3 = 2$ (não curado). A frequência e o valor médio da cela (i, j, k) da tabela de contingência em causa são, respectivamente, n_{ijk} e μ_{ijk}, $i = 1, 2$, $j = 1, 2, 3$ e $i, k = 1, 2$.

Admitindo que todas as variáveis aleatórias associadas às frequências n_{ijk} são independentes, o modelo probabilístico para a tabela de contingência em causa é o produto de distribuições de Poisson,

cuja f.m.p. é

$$f(\mathbf{n}|\boldsymbol{\mu}) = \prod_{i=1}^{2}\prod_{j=1}^{3}\prod_{k=1}^{2}\frac{\mathrm{e}^{-\mu_{ijk}}\mu_{ijk}^{n_{ijk}}}{n_{ijk}!}, \tag{4.8}$$

onde $\mathbf{n} = (n_{111}, \ldots, n_{232})^T$, $\boldsymbol{\mu} = (\mu_{111}, \ldots, \mu_{232})^T$, $n_{ijk} \in \mathbb{N}_0$ e $\mu_{ijk} \in \mathbb{R}^+$, $i = 1, 2$, $j = 1, 2, 3$ e $k = 1, 2$.

Sob o modelo probabilístico (4.8), podemos ajustar alguns modelos log-lineares hierárquicos, *e.g.*, o modelo de inexistência de interacção de segunda ordem $(X_1.X_2, X_1.X_3, X_2.X_3)$ dado por

$$\ln \mu_{ijk} = u + u_i^{X_1} + u_j^{X_2} + u_k^{X_3} + u_{ij}^{X_1X_2} + u_{ik}^{X_1X_3} + u_{jk}^{X_2X_3}, \tag{4.9}$$

onde os parâmetros u_i's e u_{ij}'s são definidos consoante a parametrização do modelo em causa. Se a parametrização for em termos de cela de referência (111), os parâmetros do modelo (4.9) tem a seguinte definição:

$$\begin{array}{rcll} u & = & \ln \mu_{111}, & \\ u_i^{X_1} & = & \ln(\mu_{i11}/\mu_{111}), & \\ u_j^{X_2} & = & \ln(\mu_{1j1}/\mu_{111}), & \\ u_k^{X_3} & = & \ln(\mu_{11k}/\mu_{111}), & \\ u_{ij}^{X_1X_2} & = & \ln(\mu_{ij1}\mu_{111}/(\mu_{i11}\mu_{1j1})) & \equiv \ln \Delta_{ij}^{X_3}, \\ u_{ik}^{X_1X_3} & = & \ln(\mu_{i1k}\mu_{111}/(\mu_{i11}\mu_{11k})) & \equiv \ln \Delta_{ik}^{X_2}, \\ u_{jk}^{X_2X_3} & = & \ln(\mu_{1jk}\mu_{111}/(\mu_{1j1}\mu_{11k})) & \equiv \ln \Delta_{jk}^{X_1}, \end{array} \tag{4.10}$$

para $i = 1, 2$, $j = 1, 2, 3$ e $k = 1, 2$. Note-se que os parâmetros em (4.10) satisfazem as seguintes restrições de identificabilidade:

$$\begin{array}{c} u_1^{X_1} = u_1^{X_2} = u_1^{X_3} = 0 \\ u_{11}^{X_1X_2} = u_{11}^{X_1X_3} = u_{11}^{X_2X_3} = 0 \end{array}$$

para $i = 1, 2$, $j = 1, 2, 3$ e $k = 1, 2$.

Os testes de ajustamento de modelos log-lineares hierárquicos para os dados do exemplo 4.3 indicam inadequação desses modelos,

exceptuando o modelo (4.9) e um modelo resultante deste quando $u_{ij}^{X_1X_2} = 0$ (denotado por M^*). Os valores observados da função desvio desses testes encontram-se na tabela 4.9, bem como os respectivos graus de liberdade e *p-values* calculados a partir da distribuição do qui-quadrado. Neste quadro, o modelo M^* com *p-value*= 0.4974 é o "melhor" modelo log-linear hierárquico no ajustamento dos dados em causa.

Tabela 4.9: Testes de ajustamento de modelos hierárquicos.

modelo log-linear	estatística	g. liberdade	*p-value*
(X_1)	229.83	10	0.0000
(X_2)	298.34	9	0.0000
(X_3)	118.80	10	0.0000
(X_1, X_2, X_3)	45.218	7	0.0000
$(X_1.X_2, X_3)$	42.373	5	0.0000
$(X_1.X_3, X_2)$	34.311	6	0.0000
$(X_2.X_3, X_1)$	14.279	5	0.0139
$(X_1.X_2, X_1.X_3)$	31.467	4	0.0000
$(X_1.X_2, X_2.X_3)$	11.435	3	0.0096
$(X_1.X_3, X_2.X_3)$	3.373	4	0.4974
$(X_1.X_2, X_1.X_3, X_2.X_3)$	2.555	2	0.2787

Na tabela 4.10 encontram-se as estimativas dos parâmetros do modelo M^* e dos seus erros padrões. Sob o modelo M^*, a razão de chances $\Delta_{ij}^{X_3} = 1$, $\forall\, i, j$, em (4.10), *i.e.*, há independência condicional entre X_1 e X_2 para cada nível de X_3. Logo, a primeira questão de interesse é respondida com a ausência de associação entre o tipo de diagnóstico e de tratamento nos pacientes curados ou não curados. Quanto à segunda questão de interesse podemos concluir, com

base no modelo M^*, que a comparação entre tratamentos não depende do tipo de diagnóstico. Observe-se que as razões de chances não unitárias definidas em (4.10), sob o modelo M^*, são estimadas por

$$\widehat{\Delta}_{22}^{X_2} = 0.401, \quad \widehat{\Delta}_{22}^{X_1} = 0.487 \quad \text{e} \quad \widehat{\Delta}_{32}^{X_1} = 2.406.$$

A interpretação destes valores permite informar sobre a eficácia dos tratamentos aplicados para a cura dos pacientes. Por exemplo, a chance de cura de um paciente tratado com C é 2.406 vezes maior do que à de um paciente tratado com A, contrariamente à chance de cura com o tratamento B que é inferior (0.487 vezes) à do tratamento A.

Tabela 4.10: Estimativas do modelo log-linear M^*.

parâmetro	estimativa	erro padrão
u	4.353	0.09928
$u_2^{X_1}$	-0.6574	0.1089
$u_2^{X_2}$	0.2727	0.1222
$u_3^{X_2}$	-0.1457	0.1352
$u_2^{X_3}$	-1.323	0.2282
$u_{22}^{X_1X_3}$	-0.9139	0.2952
$u_{22}^{X_2X_3}$	-0.7190	0.3425
$u_{32}^{X_2X_3}$	0.8781	0.2784

Capítulo 5

Aplicações II: Modelos Contínuos

Neste capítulo daremos continuidade a ilustração dos modelos lineares generalizados no que se refere aos **modelos contínuos**, *i.e.*, modelos com variável resposta contínua. O modelo de regressão linear normal é o modelo mais popular entre os MLG no ajustamento desse tipo de dados. Entretanto, algumas vezes o modelo normal é adoptado sem respeitar as características principais da situação em estudo. Nesse caso, outros modelos de regressão podem ser ajustados aos dados, *e.g.*, o modelo de regressão gama frequentemente usado quando a variável resposta assume somente valores positivos. Ne secção 5.1 analisamos um conjunto de dados relativo a um estudo de doenças vasculares em Portugal, onde o modelo gama mostra-se mais adequado que o modelo normal.

Outra situação de interesse em modelos contínuos ocorre quando a variável resposta não é observada para todas as elementos do conjunto de dados (censura). Por exemplo, o tempo de vida de um

indivíduo não é observado se ele ainda estiver vivo depois do fim do período em estudo. A análise estatística de tempos de vida é conhecida por análise de sobrevivência, onde alguns dos seus modelos fazem parte dos modelos lineares generalizados. Na secção 5.2 apresentamos um modelo de sobrevivência paramétrico (exponencial) para um estudo de tempos de remissão de pacientes com leucemia.

5.1 Modelos de Regressão Gama

O modelo de regressão gama é usado na análise de dados contínuos com suporte positivo para a distribuição da variável resposta. Eles também são adoptados quando a variância crescente com a média ou mais frequentemente quando o coeficiente de variação dos dados for aproximadamente constante. Nesta secção faremos a análise de alguns modelos contínuos no exemplo a seguir, sobretudo o modelo gama como alternativa aos demais MLG.

Exemplo 5.1 Doenças Vasculares

Em Teles (1995) foi analisado um conjunto de dados relativo a um estudo de doenças vasculares que constituem a primeira causa de morte em Portugal. Usualmente os factores de riscos nos enfartes ou acidente vascular cerebral (AVC) são a hipercolesterolémia, a hipertensão arterial e o tabagismo. Porém, acredita-se que a ocorrência precoce de doenças vasculares pode ser influenciada pela hiperhomocisteinemia, traduzida pelos valores elevados das variáveis homocisteinemia basal (HSP) e homocisteinemia após sobrecarga com metionima oral (HSP). O objectivo deste estudo é encontrar quais as variáveis que influenciam significativamente os

valores das variáveis HBP e HSP.

Nesse estudo, foram observados 145 indivíduos, dos quais 121 são homens, que sofreram enfarte ou AVC com idades entre 26 a 56 anos. Para cada indivíduo foram registados os valores de 30 variáveis clínico-laboratoriais. A variável fumar deu origem a duas variáveis mudas identificando o fumador activo e o ex-fumador, contrariamente às outras sete variáveis qualitativas. Os dados deste estudo encontram-se em Teles (1995) mas a descrição da codificação dessas variáveis está na tabela 5.1, exceptuando HBP e HSP.

A selecção das variáveis explicativas que influenciam a variável resposta HSP pode ser feita assumindo o modelo normal sem problemas de inadequação do modelo. Teles (1995) efectua uma selecção de covariáveis para o "melhor" modelo usando o método *stepwise* (*backward*). As covariáveis escolhidas como as mais importantes na explicação de HSP são **fumar**, **gluc**, **ureia**, **AST**, **GGT** e **HBP**. Por questões de simplicidade, a análise deste estudo é feita aqui somente com a variável resposta HBP e as 29 covariáveis observadas, incluindo as duas variáveis mudas da covariável fumar e excluindo a variável HSP.

Numa análise exploratória da variável HBP podemos notar uma assimetria no seu histograma (figura 5.1), podendo-se antever uma inadequação do modelo normal facilmente verificada com um teste de ajustamento do modelo. Neste caso, dois potenciais candidatos para modelar os dados em causa são os modelos log-normal e gama. O primeiro pode ser obtido simplesmente com a transformação logaritímica de HBP, *i.e.*,

$$\ln HBP_i = \mathbf{z}_i^T \boldsymbol{\beta} + \epsilon_i, \tag{5.1}$$

onde $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, 145$.

Tabela 5.1: Codificação das variáveis nas doenças vasculares.

variáveis	codificação
sexo	0 (feminino) e 1 (masculino)
fumar	0 (não fumador), 1 (fumador) e 2 (ex-fumador)
dislip	0 (sem dislipidemia) e 1 (com dislipidemia)
diab	0 (sem diabetes) e 1 (com diabetes)
obes	0 (não obeso) e 1 (obeso)
inativ	0 (vida inactiva) e 1 (vida activa)
histfa	0 (sem história familiar) e 1 (com história familiar)
hipert	0 (sem hipertensão) e 1 (com hipertensão)
idade	(em anos)
padia	pressão arterial diastólica
pasis	pressão arterial sistólica
hem	hemoglobina
htc	hematócrito
leuc	número de leucócitos
gluc	glucose
ureia	ureia
creat	creatinina
AST	aspartato aminotransferase
GGT	gama-glutamil-transpeptidase
falc	fosfatase alcalina
bil	bilirrubina total
Na	sódio
K	potássio
col	colesterol total
LDL	colesterol das LDL
HDL	colesterol das HDL
Tg	triglicéridos
VGM	volume globular médio

Para o modelo gama basta considerar o seguinte modelo multiplicativo

$$HBP_i = \exp(\mathbf{z}_i^T \boldsymbol{\beta})\, \epsilon_i, \tag{5.2}$$

onde $\epsilon_i \sim Ga(\nu, \nu/\mu_i)$, $i = 1, \ldots, 145$.

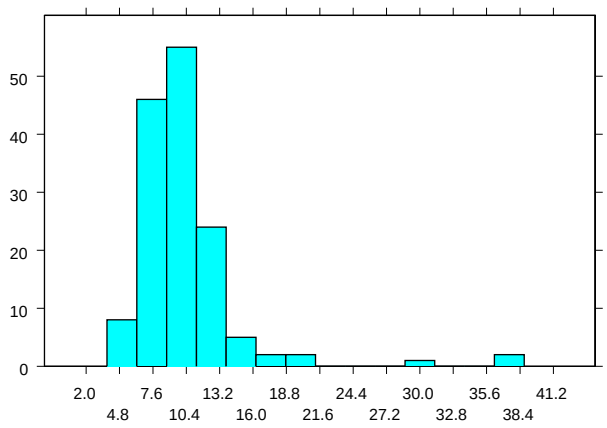

Figura 5.1: Histograma da covariável HBP.

A selecção de covariáveis para o modelo (5.1) pode ser feita em muitos pacotes estatísticos, visto que esses possuem geralmente módulos com métodos de selecção de covariáveis para modelos normais, ou seja, modelos com a variável $\ln HBP$. Usando o *software* S-plus com medidas AIC e método de selecção *backward*, o melhor de efeitos principais seleccionado incluiu as covariáveis **sexo**, **creat**, **falc**, **pasis** e **padia**. Neste modelo, as funções desvios nula (modelo nulo) e reduzida (modelo completo) são estimadas, respectivamente, por 12.087 e 8.214 com os seus graus de liberdade iguais a 126 e 121. Estes valores permitem encontrar para o modelo em causa uma

medida de ajustamento, conhecida na análise de modelos lineares como coeficiente de determinação ajustado, *i.e.*,

$$\rho^2 = 1 - \frac{126}{121}\frac{8.214}{12.087} = 0.2923, \tag{5.3}$$

o que indica pouca explicação do modelo log-normal na variação total dos dados e, portanto, partiremos para o ajustamento do modelo gama.

O método de selecção do "melhor" modelo de regressão gama para os dados de doenças vasculares é do tipo *stepwise* (*backward*), onde o modelo inicial é o modelo com os efeitos principais de todas as 29 variáveis explicativas ou covariáveis. No primeiro passo ajustaremos o modelo inicial anotando o valor observado da sua função desvio e os respectivos graus de liberdade. Este valor da função desvio dividido pelos respectivos graus de liberdade é usado para estimar o parâmetro escala ϕ (McCullagh and Nelder ,1989). Outra estimativa de ϕ pode ser calculada com base da estatística de Pearson generalizada (secção 2.2.2). Em cada um dos passos restantes uma variável explicativa será eliminada e o modelo sem esta variável é ajustado analogamente ao modelo do passo 1.

Note-se que o teste de eliminação da variável de um passo é feita com base na diferença entre as funções desvio do modelo sem a variável em causa (M_2) e do modelo ajustado no passo imediatamente anterior (M_1). Isso fará com que o modelo (M_2) seja encaixado no modelo (M_1) e, portanto, sabe-se que esta diferença divida por ϕ segue, para grandes amostras, uma distribuição qui-quadrado com $p_1 - p_2$ graus de liberdade, onde p_i é o grau de liberdade do modelo M_i, $i = 1, 2$. Além disso, como esse teste é baseado na validade de M_1, o ϕ a estimar neste teste é igual ao ϕ estimado no ajustamento do modelo M_1. Para mais detalhes, reveja a subsecção 3.1.1.

Os resultados do método de selecção de covariáveis acima encontram-se na tabela 5.2. A escolha da covariável a eliminar em cada passo faz-se com base no maior *p-value* dos testes de Wald para a anulação de cada parâmetro de regressão estimado no modelo do passo imediatamente anterior. A última covariável eliminada foi a LDL com *p-value* na ordem dos 5%. Esse valor é calculado a partir dos valores observados da função desvio no modelo do passo 22 (M_2) e no modelo do passo 21 (M_1), *i.e.*, $(D_{22} - D_{21})/\hat{\phi}_{21} = (7.722 - 7.478)/(7.478/117) = 3.818$, cuja estatística tem assintoticamente uma distribuição χ^2_1.

As covariáveis seleccionadas no processo de selecção acima foram **sexo**, **creat**, **falc**, **GGT**, **padia**, **histfa** e **fumar** (com duas variáveis mudas). Para cada uma delas foi testada a hipótese de nulidade do seu parâmetro de regressão. Em nenhum dos testes foi encontra um *p-value* superior a 3% e, portanto, as covariáveis em causa são as principais covariáveis que influenciam a variável HBP. Os valores estimados para os coeficientes de regressão do modelo seleccionado acima (modelo M^*) e os respectivos erros padrões encontram-se na tabela 5.3. O valor t é o quociente entre a estimativa do parâmetro e o seu erro padrão.

No ajustamento do modelo M^* foram calculadas as funções desvios nula (14.818) e reduzida (7.722) com graus de liberdade iguais a 126 e 118, respectivamente. Logo, o coeficiente de determinação ajustado para o modelo M^* é dado por

$$\rho^2 = 1 - \frac{126}{118}\frac{7.722}{14.818} = 0.4435, \tag{5.4}$$

indicando assim que o modelo em causa explica 44.35% da variação total de HBP. Como o valor de ρ^2 em (5.4) é maior do que em (5.3), o modelo gama M^* ajusta-se melhor aos dados de doenças vasculares do que o modelo log-normal.

Tabela 5.2: Resultados do método de selecção *backward*.

passo	covariável	modelo completo		teste de $\beta_j = 0$		
	eliminada	desvio	g.l.	desvio	g.l.	*p-value*
1	-	6.343	88	-	-	-
2	VGM	6.663	94	4.438	6	0.618
3	hipert	6.663	95	0.000	1	1.000
4	htc	6.663	96	0.000	1	1.000
5	idade	6.664	97	0.014	1	0.906
6	ureia	6.679	98	0.218	1	0.641
7	HDL	6.696	99	0.249	1	0.618
8	bil	6.707	100	0.163	1	0.686
9	Tg	6.725	101	0.268	1	0.605
10	dislip	6.735	102	0.150	1	0.698
11	AST	6.846	106	1.682	4	0.794
12	inativ	6.866	107	0.310	1	0.578
13	obes	6.887	108	0.327	1	0.567
14	Na	6.907	109	0.313	1	0.576
15	hem	6.937	110	0.473	1	0.492
16	K	6.988	112	0.808	2	0.668
17	col	7.056	113	1.090	1	0.296
18	gluc	7.170	114	1.827	1	0.176
19	diab	7.256	115	1.367	1	0.242
20	pasis	7.380	116	1.965	1	0.161
21	leuc	7.478	117	1.541	1	0.214
22	LDL	7.722	118	3.818	1	0.051

Para identificar observações que não são bem explicadas pelo modelo M^*, podemos efectuar uma análise de resíduos, *e.g.*, calculando os desvios residuais. O gráfico dos desvios residuais versus os

valores ajustados (figura 5.2) indica que três observações aberrantes, potenciais candidatas a *outliers*: $R^{\star D}_{46} = 0.95$, $R^{\star D}_{47} = 1.06$ e $R^{\star D}_{91} = -0.59$. Estas observações também são tidas como anómalas no gráfico do ajustamento da normal reduzida aos resíduos de Pearson, onde os seus valores se distanciam da recta traçada implicando a falta de ajustamento.

Tabela 5.3: Resultados do modelo gama seleccionado (M^*)

covariável	estimativa	erro padrão	valor t	*p-value*
ordenada na origem	2.260	0.2512	8.998	0.0000
sexo	0.292	0.0748	3.901	0.0001
histfa	-0.100	0.0504	-1.984	0.0473
creat	0.101	0.0252	3.999	0.0001
falc	0.003	0.0009	3.465	0.0005
GGT	-0.002	0.0007	-2.297	0.0216
Padia	-0.005	0.0025	-2.193	0.0283
fumar1 (fumador)	0.239	0.0737	3.245	0.0012
fumar2 (ex-fumador)	-0.006	0.0588	-0.110	0.9112

Com base no modelo gama M^*, cujas estimativas encontram-se na tabela 5.3, podemos concluir que o risco de homocisteinemia basal (HSP) é maior nos indivíduos do sexo masculino, nos fumadores e nos doentes com valores elevados para a cretinina e fosfatase alcalina. Note-se que a condição de ex-fumador parece ser um factor preventivo da hiperhomocisteinemia basal, mas isso pode ser devido ao facto dos doentes vasculares deixarem de fumar após o acidente vascular agudo provocando uma diminuição dos valores da homocisteinemia basal.

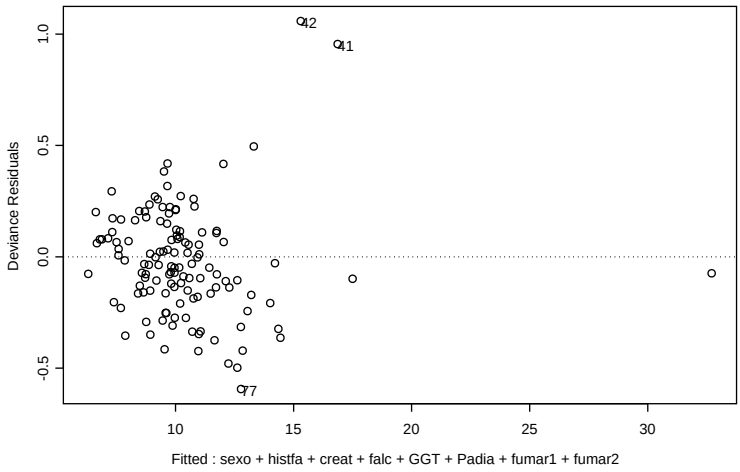

Figura 5.2: Gráfico dos desvios residuais × valores ajustados.

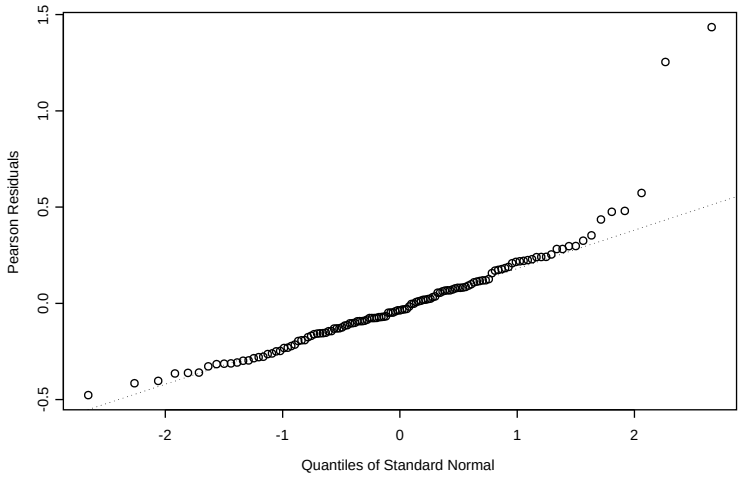

Figura 5.3: Gráfico dos resíduos de Pearson × quantis da $N(0,1)$.

5.2 Modelos de Sobrevivência

O conjunto de dados em análise de sobrevivência é formado por unidades (indivíduos) que são observadas até à ocorrência de algum evento de interesse, por exemplo, a falha das unidades (morte). Frequentemente esse evento não chega a ocorrer para algumas dessas unidades (censura), o que torna a análise desses dados peculiar. Daí o surgimento da análise de sobrevivência como metodologia estatística apropriada para o estudo de unidades sujeitas a censura.

O tempo de vida ou tempo de sobrevivência é considerado uma variável não negativa (Y) geralmente contínua. Para Y contínuo, sejam $f(y)$ e $F(y)$ $P(Y \leq y)$ são as funções de densidade de probabilidade e de distribuição de Y. Outras funções de interesse em análise de sobrevivência são a função de sobrevivência e a função risco. A função de sobrevivência é a probabilidade de uma unidade sobreviver pelo menos até ao instante y, *i.e.*,

$$S(y) \equiv P(Y \geq y), \tag{5.5}$$

enquanto a função risco (*hazard*) é a taxa de ocorrência do evento de interesse no instante y, definida por

$$\lambda(y) = \lim_{dy \to 0^+} \frac{P(y \leq Y < y + dy | Y \geq y)}{dy}. \tag{5.6}$$

As relações entre as funções acima são dadas por

$$\begin{aligned} \lambda(y) &= f(y)/S(y) \\ S(y) &= \exp[-\Lambda(y)], \end{aligned}$$

onde $\Lambda(y) \equiv \int_0^y \lambda(u)du$ é a função risco cumulativa ou integrada.

Assumindo que a função risco é constante, *i.e.*, $\lambda(y) = \delta$, $\delta > 0$, obtemos a distribuição exponencial para os tempos de vida, cuja

f.d.p. é $f(y) = \delta \exp[-\delta y]$. Quando existem covariáveis no conjunto de dados, a função risco em causa é redefinida com a substituição de δ por $\exp(\mathbf{z}^T\boldsymbol{\beta})$, o que define o modelo de sobrevivência exponencial como

$$\lambda(y) = \exp(\mathbf{z}^T\boldsymbol{\beta}). \tag{5.7}$$

Exemplo 5.2 Doentes com Leucemia

Em onze hospitais dos E.U.A., Freireich et al. (1963) analisaram 42 crianças com leucemia aguda após terem entrado num período de remissão da doença. Os pacientes foram classificados em dois tipos de remissão da doença: completa ou parcial, consoante todos ou alguns sinais da doença tivessem desaparecidos da medula óssea, respectivamente. Posteriormente, eles foram aleatoriamente distribuídos por 2 tipos de terapia: tratamento com a droga *6-Mercaptopurine* (6-MP) e placebo. Os pacientes ficaram em observação até ao retorno da doença (recaída) ou até ao fim do estudo (indicador de censura). Os tempos de sobrevivência são aqui as durações do período de remissão da doença nos pacientes. Os valores observados desses tempos consoante o tipo de terapia encontram-se na tabela 5.4.

Para cada paciente do conjunto de dados do exemplo 5.2 associaremos um par aleatório (Y_i, γ_i), onde Y_i é o tempo de vida (remissão da doença) do paciente i e γ_i é a sua função indicadora da ausência de censura, $i = 1, \ldots, 42$. Assim, supondo que os 42 pares aleatórios são independentes e oriundos de uma população com f.d.p. $f(\cdot)$ e f.s. $S(\cdot)$ indexadas pelo vector paramétrico $\boldsymbol{\beta}$ e com mecanismo de censura não informativo para $\boldsymbol{\beta}$, a função de verosimilhança de $\boldsymbol{\beta}$ dado o conjunto de dados é expressa por

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(t_i|\boldsymbol{\beta})^{\gamma_i} S(t_i|\boldsymbol{\beta})^{1-\gamma_i}. \tag{5.8}$$

Tabela 5.4: Dados de pacientes com leucemia

Durações da remissão da doença em semanas (* censura)										
Placebo										
1	12	2	12	4	23	4	22	8	8	11
11	2	15	5	1	3	17	8	5	8	
Tratamento										
10	23	16	25*	19*	35*	9*	7	22	17*	13
34*	11*	6	6	6*	32*	6	32*	20*	10*	

Freireich et al. (1963).

A função de verosimilhança (5.8) é a base da inferência sobre o parâmetro $\boldsymbol{\beta}$ no modelo exponencial (5.7). Não sendo difícil verificar que $\lambda(y)/\Lambda(y) = y^{-1}$ e, portanto, o modelo exponencial pode ser ajustado através do ajustamento do modelo Poisson com

$$\ln(\mu_i) = \ln(y_i) + \mathbf{z}^T\boldsymbol{\beta}, \tag{5.9}$$

onde $\mathbf{z} = (1, x)^T$, $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, x é a covariável do tipo de terapia dos pacientes assumindo os valores 0 para o placebo e 1 para o tratamento e o termo $\ln(y_i)$ é considerado um *offset*. Para maiores detalhes sobre a implementação do modelo (5.9) no GLIM, consulte Aitkin and Clayton (1980).

Com base no ajustamento do modelo (5.9), usando o modelo de Poisson com variável resposta γ_i, função de ligação logarítmica e $\ln(y_i)$ como *offset*, o parâmetro de regressão do modelo exponencial é estimada por

$$\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)^T = (-2.159, -1.526)^T,$$

cujos erros padrões das suas componentes estimadas são, respectivamente, 0.2179 e 0.3958. O teste de ajustamento deste modelo é feito pela função desvio com valor observado igual a 38.017. Este valor é comparado com uma distribuição χ^2_{40} proporciona um *p-value*= 0.5598 e, portando, o modelo exponencial ajusta-se bem aos dados de doentes de leucemia. Para testar a hipótese de nulidade do β_1 pela estatística de Wald, obtém-se o valor observado igual a 14.865 com *p-value* igual a 0.0001. Isso indica um efeito do tipo de terapia na duração dos tempos de remissão da doença.

Apêndice A

Programas do S-plus

Os resultados para o ajustamento de modelos lineares generalizados nos exemplos dos capítulos 4 e 5 foram obtidos com base nos *software* S-plus ou GLIM (*Generalised Linear Interactive Modelling*). Neste apêndice ilustramos alguns dos programas *input/output* do S-plus usados nos exemplos 4.1 (secção A1) e 5.1 (secção A2). Em Fahrmeir and Tutz (1994, apend. B) pode-se encontrar uma descrição resumida dos vários *software* que ajustam modelos lineares generalizados.

A.1 Exemplo 4.1

Todos os modelos lineares generalizados referidos no exemplo 4.1 (processo infeccioso pulmonar) foram ajustados no S-plus. Como o modelo (4.2) foi seleccionado como o "melhor" modelo de regressão logística para os dados em causa, o seu programa *output* encontra-se a seguir, incluindo o respectivo programa *input* nas suas primeiras linhas.

```
*** Generalized Linear Model ***

Call: glm(formula = PIP ~ idade * HL + sexo * FF +
   HL * FF, family = binomial(link= logit),
   data = pulmao, na.action = na.omit,
   control = list(epsilon = 0.001, maxit = 50,
   trace = F))

Deviance Residuals:
       Min         1Q     Median         3Q      Max
 -2.328786 -0.6075183 -0.1451404 0.7491216 2.222905

Coefficients:
                  Value Std. Error     t value
(Intercept) -0.03374410 1.02440563 -0.03294018
      idade  0.03944518 0.01730781  2.27903965
         HL -5.39156719 1.59267110 -3.38523578
       sexo -1.38570230 0.57485555 -2.41052260
         FF -5.16868629 1.62465215 -3.18141104
   idade:HL  0.05894743 0.02802903  2.10308533
    sexo:FF  3.15960450 1.39644616  2.26260388
      HL:FF  2.80006789 1.10322994  2.53806373

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 236.3412 on 174 degrees of freedom

Residual Deviance: 145.4526 on 167 degrees of freedom

Number of Fisher Scoring Iterations: 4
```

A.2 Exemplo 5.1

Os modelos de regressão gama e log-normal foram os principais MLG no método de selecção das 28 covariáveis do exemplo 5.1 (doenças vasculares). Para o modelo log-normal usámos o módulo de selecção *stepwise* do S-plus, com a devida transformação da variável resposta, enquanto para o modelo gama foi preciso ajustar vários MLG. Os programas *output/input* dos "melhores" modelos de regressão log-normal e gama, seleccionados pelo método *stepwise* (*backward*), encontram-se a seguir.

Modelo log-normal:

```
*** Generalized Linear Model ***

Call: glm(formula = log(HBP) ~ sexo + creat + falc +
   Pasis + Padia, family = gaussian, data = cardio,
   na.action = na.omit, control = list(epsilon = 0.001,
   maxit = 50, trace = F))

Deviance Residuals:
        Min          1Q      Median          3Q      Max
 -0.4979828 -0.1738579 -0.03313648 0.1365756 1.207923

Coefficients:
                   Value   Std. Error   t value
(Intercept)  2.167615163 0.2409437791  8.996352
       sexo  0.266821626 0.0666779858  4.001645
      creat  0.107322410 0.0233249571  4.601184
       falc  0.002510331 0.0008307429  3.021790
```

```
      Pasis  0.002955957 0.0023398125  1.263331
      Padia -0.009668315 0.0039059352 -2.475288

(Dispersion Parameter for Gaussian family taken to be
  0.0678844 )

    Null Deviance: 12.08704 on 126 degrees of freedom

Residual Deviance: 8.214014 on 121 degrees of freedom

Number of Fisher Scoring Iterations: 1
```

Modelo gama:

```
*** Generalized Linear Model ***

Call: glm(formula = HBP ~ sexo + histfa + creat +
    falc + GGT + Padia + fumar1 + fumar2, family =
    Gamma(link = log), data = cardio, na.action =
    na.exclude, control = list(epsilon= 0.0001,
    maxit = 50, trace = F))

Deviance Residuals:
        Min         1Q      Median        3Q     Max
 -0.5949305 -0.1636099 -0.03546889 0.1107428 1.05547

Coefficients:
                   Value   Std. Error    t value
(Intercept)  2.260377991 0.2512044953  8.9981590
       sexo  0.291984755 0.0748555323  3.9006436
```

```
     histfa -0.100006194 0.0504076823 -1.9839475
      creat  0.100905628 0.0252352178  3.9986034
       falc  0.003185380 0.0009192149  3.4653269
        GGT -0.001585101 0.0006899399 -2.2974477
      Padia -0.005513712 0.0025141684 -2.1930558
     fumar1  0.239228567 0.0737255067  3.2448548
     fumar2 -0.006446950 0.0587982894 -0.1096452

(Dispersion Parameter for Gamma family taken to be
                                          0.0771972 )

    Null Deviance: 14.8181 on 126 degrees of freedom

Residual Deviance: 7.722002 on 118 degrees of freedom
18 observations deleted due to missing values

Number of Fisher Scoring Iterations: 4
```

Apêndice B

Programas do GLIM

O GLIM (*Generalised Linear Interactive Modelling*) foi o *software* usado para ajustar os modelos lineares generalizados dos exemplos 4.2, 4.3 e 5.2, cujos programas *input/output* encontram-se na secção B1, B2 e B3, respectivamente.

B.1 Exemplo 4.2

Para os modelos de dose-resposta do exemplo 4.2 (mortalidade de besouros) adoptámos modelos lineares generalizados (binomial) com funções de ligação *logit*, *probit* e complementar log-log. A comparação destes 3 novos modelos pode ser obtida do seguinte programa *input/output*.

```
[o] GLIM 3.77 update 1 (copyright)1985
[o]                  Royal Statistical Society, London
[i] ? $input 5$
[i] $c   Ajuste do modelo logistico, probit e extremit
```

```
[i] $units 8
[i] $data dose y n
[i] $read
[i] 1.6907   6  59
[i] 1.7242  13  60
[i] 1.7552  18  62
[i] 1.7842  28  56
[i] 1.8113  52  63
[i] 1.8369  53  59
[i] 1.861   61  62
[i] 1.8839  60  60
[i] $yva y $err b n
[i] $c Modelo Logistico
[i] $fit dose $dis e r $
[o] scaled deviance = 11.232 at cycle  4
[o]            d.f. =  6
[o]          estimate        s.e.     parameter
[o]     1      -60.72       5.180     1
[o]     2       34.27       2.912     DOSE
[o]     scale parameter taken as  1.000
[o]   unit  observed    out of    fitted   residual
[o]      1         6        59     3.457      1.409
[o]      2        13        60     9.842      1.101
[o]      3        18        62    22.451     -1.176
[o]      4        28        56    33.898     -1.612
[o]      5        52        63    50.096      0.594
[o]      6        53        59    53.291     -0.128
[o]      7        61        62    59.222      1.091
[o]      8        60        60    58.743      1.133
[i] $c Modelo Probit
```

```
[i] $link p $fit dose $dis e r $
[w] -- model changed
[o] scaled deviance = 10.120 at cycle  4
[o]            d.f. =  6
[o]           estimate        s.e.     parameter
[o]     1      -34.93       2.647     1
[o]     2       19.73       1.487     DOSE
[o]     scale parameter taken as  1.000
[o]   unit  observed    out of    fitted   residual
[o]      1         6        59     3.358      1.485
[o]      2        13        60    10.722      0.768
[o]      3        18        62    23.482     -1.435
[o]      4        28        56    33.815     -1.589
[o]      5        52        63    49.615      0.735
[o]      6        53        59    53.319     -0.141
[o]      7        61        62    59.664      0.891
[o]      8        60        60    59.228      0.884
[i] $c Modelo Extremit
[i] $link c $fit dose $dis e v r $
[w] -- model changed
[o] scaled deviance = 3.4464 at cycle  4
[o]            d.f. = 6
[o]           estimate        s.e.     parameter
[o]     1      -39.57       3.238     1
[o]     2       22.04       1.798     DOSE
[o]     scale parameter taken as  1.000
[o]
[o]   unit  observed    out of    fitted   residual
[o]      1         6        59     5.589      0.183
[o]      2        13        60    11.281      0.568
```

```
 [o]      3        18        62     20.954      -0.793
 [o]      4        28        56     30.369      -0.636
 [o]      5        52        63     47.776       1.243
 [o]      6        53        59     54.143      -0.541
 [o]      7        61        62     61.113      -0.121
 [o]      8        60        60     59.947       0.230
 [i] $stop
```

B.2 Exemplo 4.3

O ajustamento de todos os modelos log-lineares (regressão de Poisson) ajustados aos dados do exemplo 4.3 (infecções urinárias) também foram realizados no GLIM, porém numa versão mais antiga desse *software*. Segue-se o programa *input/output* associado a esses modelos.

```
GLIM 3.77 update 0 (copyright)1985
                   Royal Statistical Society, London
? $c Ajuste do modelo log-linear (regressao Poisson)
$COM? $units 12
? $data n
$DAT? $read
$REA? 78 20
$REA? 101 11
$REA? 68 46
$REA? 40 5
$REA? 54 5
$REA? 34 6
? $calculate diag=%GL(2,6):trat=%GL(3,2):cura=%GL(2,1)
$CAL? $factor diag 2 trat 3 cura 2
```

```
$FAC? $yva n $err p
? $c Modelo log-linear (A)
$COM? $fit diag $
scaled deviance = 229.83 at cycle  4
           d.f. =  10
? $c Modelo log-linear (B)
$COM? $fit trat $
scaled deviance = 298.34 at cycle  4
           d.f. =   9
? $c Modelo log-linear (C)
$COM? $fit cura $
scaled deviance = 118.80 at cycle  4
           d.f. =  10
? $c Modelo log-linear (A,B,C)
$COM? $fit diag + trat + cura $
scaled deviance = 45.218 at cycle  4
           d.f. =  7
? $c Modelo log-linear (AB,C)
$COM? $fit diag + trat + cura + diag*trat $
scaled deviance = 42.373 at cycle  4
           d.f. =  5
? $c Modelo log-linear (AC,B)
$COM? $fit diag + trat + cura + diag*cura $
scaled deviance = 34.311 at cycle  4
           d.f. =  6
? $c Modelo log-linear (BC,A)
$COM? $fit diag + trat + cura + trat*cura $
scaled deviance = 14.279 at cycle  4
           d.f. =  5
? $c Modelo log-linear (AB,AC)
```

```
$COM? $fit diag + trat + cura + diag*trat + diag*cura $
scaled deviance = 31.467 at cycle  4
           d.f. =  4
? $c Modelo log-linear (AB,BC)
$COM? $fit diag + trat + cura + diag*trat + trat*cura $
scaled deviance = 11.435 at cycle  3
           d.f. =  3
? $c Modelo log-linear (AC,BC)
$COM? $fit diag + trat + cura + diag*cura + trat*cura $
scaled deviance = 3.3728 at cycle  3
           d.f. = 4
? $c Modelo log-linear (AB,AC,BC)
$COM? $fit diag + trat + cura + diag*trat + diag*cura
                                      + trat*cura $
scaled deviance = 2.5555 at cycle  3
           d.f. = 2
? $c Melhor modelo log-linear (AC,BC)
$COM? $fit diag + trat + cura + diag*cura + trat*cura
$FIT? $dis e v r $
scaled deviance = 3.3728 at cycle  3
           d.f. = 4
         estimate        s.e.     parameter
    1       4.353     0.09928     1
    2     -0.6574      0.1089     DIAG(2)
    3      0.2727      0.1222     TRAT(2)
    4     -0.1457      0.1352     TRAT(3)
    5      -1.323      0.2282     CURA(2)
    6     -0.9139      0.2952     DIAG(2).CURA(2)
    7     -0.7190      0.3425     TRAT(2).CURA(2)
    8      0.8781      0.2784     TRAT(3).CURA(2)
```

```
    scale parameter taken as  1.000
  unit  observed    fitted   residual
     1        78    77.723      0.031
     2        20    20.699     -0.154
     3       101   102.093     -0.108
     4        11    13.247     -0.617
     5        68    67.184      0.100
     6        46    43.054      0.449
     7        40    40.277     -0.044
     8         5     4.301      0.337
     9        54    52.907      0.150
    10         5     2.753      1.354
    11        34    34.816     -0.138
    12         6     8.946     -0.985
? $stop
```

B.3 Exemplo 5.2

Por fim, os resultados inferências do exemplo 5.2 (doentes com leucemia) foram obtidos também no GLIM, cujo programa *input/output* encontra-se a seguir.

```
GLIM 3.77 update 0 (copyright)1985
                   Royal Statistical Society, London

? $c Ajuste do modelo de sobrevivência exponencial
$COM? $units 42 $
? $data tempo censura grupo $
? $read
$REA?  1  1  0 10  1 1 22  1  0  7  1 1  3  1  0 32  0 1
```

```
$REA? 12  1  0 23  1 1  8  1  0 22  1 1 17  1  0  6  1 1
$REA?  2  1  0 16  1 1 11  1  0 34  0 1  8  1  0 32  0 1
$REA? 12  1  0 25  0 1  2  1  0 11  0 1  5  1  0 20  0 1
$REA?  4  1  0 19  0 1 15  1  0  6  1 1  8  1  0 17  0 1
$REA? 23  1  0 35  0 1  5  1  0  6  1 1 11  1  0 13  1 1
$REA?  4  1  0  9  0 1  1  1  0  6  0 1  8  1  0 10  0 1
? $yvariate censura  $err p  $link l $
? $calc lnt=%log(tempo)  $offset lnt $
? $fit grupo $display e $
scaled deviance = 38.017 at cycle  4
           d.f. = 40

         estimate        s.e.     parameter
    1      -2.159      0.2179     1
    2      -1.526      0.3958     GRUP
    scale parameter taken as  1.000
? $stop
```

Bibliografia

[1] Aitkin, M. (1980). The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Applied Statistics*, **29**, 156-163.

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Auto. Cntl.*, **AC-19**, 716-723.

[3] Anscombe, F.J. (1953). Contribution to the discussion of H. Hotelling's paper. *Journal of the Royal Statistical Society*, **B 15**, 229-230.

[4] Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*. John Wiley & Sons, New York.

[5] Berkson, J. (1944). Application of the logistic function to bioassay. *Journal of the American Statistical Association*, **39**, 357-365.

[6] Birch, M.W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society*, **B52**, 220-233.

[7] Bliss, C.I. (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology*, **22**, 134-167.

[8] Breslow, N.E. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics*, **33**, 38-44.

[9] Buse, A. (1982). The likelihood ratio, Wald and Lagrange multiplier test: an expository note. *The American Statistician*, **36**, 153-157.

[10] Christensen, R. (1997). *Log-Linear Models and Logistic Regression*. 2nd edition, Springer, New York.

[11] Collet, D. (1991). *Modelling Binary Data*. Chapman and Hall, London.

[12] Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics*, **19**, 15-18.

[13] Cordeiro, G.M. (1986). *Modelos Lineares Generalizados*. VII Simpósio Nacional de Probabilidades e Estatística, Campinas, São Paulo.

[14] Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

[15] Cox, D.R. and Snell, E.J. (1968). A general definition of residuals. *Journal of the Royal Statistical Association*, **B 30**, 248-275.

[16] Dyke, G.V. and Patterson, H.D. (1952). Analysis of factorial arrangements when the data are proportions. *Biometrics* **8**, 1-12.

[17] Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, **13**, 342-368.

[18] Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, New York.

[19] Feigl, P. and Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics* **21**, 826-838.

[20] Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society*, **222**, 309-368.

[21] Freireich, E. J., Gehan, E., Frei III, E., Schroeder, L.R., Wolman, I.J., Anbari, R., Burgert, E.O., Mills, S.D., Pinkel, D., Selawry, O.S., Moon, J.H., Gendel, B.R., Spurr, C.L., Storrs, R., Haurani, F., Hoogstraten B. and Lee, S. (1963). The effect of 6-Mercaptopurine on the duration of steroid-induced remissions in acute leukemia: a model for evaluation of other potentially useful therapy. *Blood*, **21**, 699-716.

[22] Glasser, M. (1967). Exponential survival with covariance. *Journal of the American Statistical Association*, **62**, 561-568.

[23] Haberman, S.J. (1974). *The Analysis of Frequency Data*. University of Chicago Press, Chicago.

[24] Hinkley, D.V. (1985). Transformation diagnostics for linear models. *Biometrika*, **72**, 487-496.

[25] Hoaglin, D.C. and Welsch, R. (1978). The hat matrix in regression and ANOVA. *American Statistician*, **32**, 17-22.

[26] Hosmer, D.W. and Lemeshow, S. (1989). *Applied Logistic Regression*. John Wiley, New York.

[27] Johnson, R.A. and Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis*. 4th edition, Prentice Hall, New Jersey.

[28] Koch, G.G., Imrey, P.B., Singer, J.M. Atkinson, S.S. and Stokes, M.E. (1985). *Analysis of Categorical Data*. Les Presses de l'Université de Montréal, Montréal.

[29] Lindsey, J.K. (1997). *Applying Generalized Linear Models*. Springer, New York.

[30] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. 2nd edition, Chapman and Hall, London.

[31] Nelder, J.A. (1966). Inverse polynomials, a useful group of multi-factor response functions. *Biometrics*, **22**, 128-141.

[32] Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, **A 135**, 370-384.

[33] Paula, G.A., Fontes, L.R. e Imanaga, A.T. (1984). Associação entre o tipo de processo infeccioso pulmonar e algumas variáveis histológicas. Relatório de Análise Estatística nº 8417. Instituto de Matemática e Estatística da Universidade de São Paulo.

[34] Paulino, C.D. and Singer, J.M. (1997). *Análise de Dados Categorizados*. Versão preliminar, Lisboa/São Paulo.

[35] Pierce, D.A. and Schafer, D.W. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association*, **81**, 977-986.

[36] Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests*. Danmarks Paedogogiske Institut, Copenhagen.

[37] Sen, P.K. and Singer, J.M. (1993). *Large Sample Methods in Statistics. An Introduction with Applications*. Chapman and Hall, New York.

[38] Shao, J. (1998). *Mathematical Statistics*. Springer, New York.

[39] Silvapulle, M.J. (1981). On the existence of maximum likelihood estimates for the binomial response models. *Journal of the Royal Statisticsl Society*, **B 43**, 310-313.

[40] Silva, G.L. (1992). *Modelos Logísticos para Dados Binários*. Tese de Mestrado em Estatística. Instituto de Matemática e Estatística da Universidade de São Paulo.

[41] Smyth, G.K.(1989). Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society*, **B 51**, 47-60.

[42] Teles, J.M.V. (1995). *Modelos Lineares Generalizados - Uma Aplicação à Medicina*. Tese de Mestrado em Estatística e Investigação Operacional. Faculdade de Ciência da Universidade de Lisboa.

[43] Wedderburn, R.W.M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, **63**, 27-32.

[44] Williams, D.A. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics*, **36**, 181-191.

[45] Write, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1-25.

[46] Zippin, C. and Armitage, P. (1966). Use of concomitant variables and incomplete survival information in the estimation of an exponential survival parameter. *Biometrics*, **22**, 665-672.

Índice Remissivo